Module for Post-Graduate Diploma in Education Programme

EPS751P:
INTRODUCTION TO EDUCATIONAL STATISTICS

PROF. Y. K. A. ETSEY
PROF. D. D. AGYEI

REPUBLIC OF GHANA

INSTITUTE OF EDUCATION, UCC

IoE/MoF/TUC/GHANA CARES TRAINING AND RETRAINING PROGRAMME FOR PRIVATE SCHOOL TEACHERS

Ministry of Finance

Trade Union Congress

University of Cape Coast

DECEMBER, 2022

TABLE OF CONTENTS

SYMBOLS AND THEIR MEANINGS

OVERVIEW

UNIT OBJECTIVES

SESSION OBJECTIVES

DO AN ACTIVITY

NOTE AN IMPORTANT POINT

TIME TO THINK AND ANSWER QUESTION(S)

REFER TO

READ OR LOOK AT

SUMMARY

SELF- ASSESSMENT TEST

ASSIGNMENT

UNIT 1: NATURE OF STATISTICS

Unit Outline

Session 1: Definitions of Statistics
Session 2: Reasons for studying statistics in education
Session 3: Variables, data and data sources
Session 4: Descriptive and inferential statistics
Session 5: Nominal and ordinal scales of measurement
Session 6: Interval and ratio scales of measurement

Dear reader, we are about to begin our study of the course in Educational Statistics. This course introduces you to the use of statistics in aiding teaching and learning. It explains the methods and procedures used to organize test scores to obtain meaningful information for decision making. This unit will attempt to make the nature of statistics clear to you. It may be exciting to begin with our own profession as teachers. As a classroom teacher, you will be expected to conduct quizzes and tests as well as give assignments and projects. These quizzes, tests, assignments and projects are to be marked and assigned scores. These scores give information about the level of performance exhibited by the learner. It is important therefore to obtain as much information as possible from the scores to help you to improve upon the teacher-learner interaction in the classroom. As a teacher, you are expected to take important decisions concerning your learners. The scores you obtain from the quizzes, tests, assignments and projects will provide information for you to enable you take the right decisions. This unit will help you to gain a better understanding of the nature of data collected and how it can be used. You will learn the meaning of statistics, acquire an understanding of basic concepts in Statistics and learn how to use appropriate statistics to describe performance.

Unit Objectives

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

1. Define the term, Statistics as used in education;
2. State the reasons why classroom teachers need to study Statistics;
3. Give examples of statistical variables;
4. Identify the sources and types of data used in Statistics;
5. Explain the differences between descriptive and inferential Statistics;
6. Differentiate among nominal, ordinal, interval and ratio scales of measurement.

SESSION 1: DEFINITIONS OF STATISTICS

You are welcome to the first session of the first Unit of your module in Educational Statistics. I hope that you will have fun with the course as you learn how to improve upon teaching and learning by the use of Statistics.

A lot of students fear the subject, Statistics because they do not like Mathematics. Many of them think that Statistics is the same as Mathematics since both of them deal with numbers and figures. If you are in this group, I advise that you remove this fear as you begin to read through this session. We use Statistics daily in our lives and we shall apply these to solving problems in the classroom. I hope that you will enjoy this course. This first session introduces you to the basic definitions of statistics. Understanding the basic definitions will provide a strong foundation for enjoying the course.

Objectives

By the end of the session, you should be able to

(a) define the term, Statistics;
(b) explain the term, educational statistics;
(c) identify and explain the characteristics of educational statistics as a subject.

Now read on...

1.1 Statistics as a body of numbers

You have probably heard the word, statistics, so many times. In one sense, statistics can be defined as the body of numbers or data collected in any field of endeavour. There are several examples.

1. Industrial statistics. This involves the collection of numbers pertaining to industries. These will include the number of employees in the industry, the number of products, and the value of products.
2. Vital statistics. These are measurements of the bust, waist and hip of a person (but mostly women). In a recent Miss Ghana competition, the vital statistics of one contestant were 36-24-36. It also refers to government records on number of births, deaths, marriages and divorces in a country.
3. Health statistics. These are numbers concerning various aspects of health. This includes the number of people who were in hospital in a particular month, the number of people who caught the H1N1 flu in a year, the number of people who died of malaria in a year or month, the number of children who were immunized in a month, life expectancy at birth etc.
4. Population statistics. These are the total number of people in a country, the number of males and females in a country, the number of adults (21 years and above), the number of youth in school, etc.

Now that you have read about statistics as a collection of numbers in any field, write down five indicators of what you think constitutes agricultural statistics. Bring your indicators to face to face for discussion.

Let us now move on to the second definition of statistics.

1.2 Statistics as computed values

The word, statistics, can also be defined as a collection of two or more values computed from a set of data. These values provide summary information on a set of data. One value is a 'statistic' (singular), with two or more values, statistics (plural).

Consider the ages (in years) of 10 adults: 48, 40, 32, 30, 25, 36, 24, 30, 30, 35. Their total age is 330 years. We can find the difference between the oldest and the youngest. This is 24 years (48-24). This difference is called the range and it is a **statistic**. We can also add all the ages and divide by 10, i.e. 330÷10 = 33. This represents the mean age of the 10 adults. This is also a **statistic**. In addition, we can find the ages that occur most frequently. In this example, it is 30 years (obtained by 3 people). This value that occurs most frequently is called the mode. It is also a **statistic**. It is also possible to obtain the proportion of the adults who are below the age of 35. This proportion is 0.4 (40%), i.e. 4 out of 10. If we obtain all four values from the same set of data, we are obtaining **statistics** (plural).

In this course, we shall compute several other statistics.

1.3 Statistics as a subject of study

You have studied subjects like Geography, History, Science, Mathematics, Education, English and Social Studies. Statistics is also considered a subject of study as any of these subjects studied in school.

As a subject, it is defined as the study of methods and procedures used in collecting, organizing, analyzing, and interpreting a body of numbers for information and decision making. The methods and procedures include such topics as Measures of Location, Measures of Variation, Correlation and Regression, Analysis of variance, Normal distribution and Hypothesis testing.

We have now looked at the three definitions of statistics. From your knowledge of these definitions, how would you define educational statistics? Pause for a minute and reflect.

Close the module now. Take your jotter and write down a definition of educational statistics. When you are done, open your module.

Now read on to find out whether the main points of your definition can be found in the definition of educational statistics.

1.4 What is Educational Statistics?

From the definitions above, Educational Statistics can be defined in two ways. Firstly, it is the body of numbers or data in the field of education. This includes the number of pupils in a school, number of teachers in a school, number of textbooks in a school; number of teachers in a region, number of pupils in a district, number of students in a University, pupil-teacher ratio in a region, primary school completion rate, high school completion rate, adult literacy rate, repetition rate etc.

Secondly it can be defined as the study of the methods and procedures used in collecting, organizing, analyzing and interpreting a body of numbers related to education for information and decision making.

The main concern of this course is the second definition. We shall study methods and procedures such as graphical representation of data, measures of central tendency, measures of variation and relative position, normal distribution and linear correlation. These methods and procedures will use examples from the field of education in the collection, analysis and interpretation of educational data.

In this session, you have learnt about the various definitions of the word, statistics. You have also learnt about what educational statistics means and the focus of this course. Keep these ideas in mind as you move on to the next session.

Self-Assessment Questions

Exercise 1.1

For each item in 1 - 3, select the best or correct option.

1. The number of mosquito nets produced in a factory can be classified as _____ statistics?
    A. health
    B. industrial
    C. population
    D. school

2. Which one of the following information is part of the vital statistics for Miss Alice Mensah?
    A. Age in 2009
    B. Country of birth
    C. Hip size
    D. Region of birth

3. The Ghana Premier Football League statistics for 2009 would include_____.
    A. age of the players' parents.
    B. home town of the players.
    C. schools players attended.
    D. total number of goals scored.

For each item in 4 - 6, indicate whether the item is False or True by circling your choice.

4. The number of students in Zion High School Form 1A is part of agricultural statistics.
    A. False
    B. True

5. The study of methods and procedures used in collecting, organizing, analyzing and interpreting data related to population for information and decision making is educational statistics.
    A. False
    B. True

6. A teacher computed the mean age of the pupils and the proportion of girls in her Primary 6 class. These values can be termed statistics.
    A. False
    B. True

SESSION 2: REASONS FOR STUDYING STATISTICS IN EDUCATION

You are welcome to the second session of Unit 1 for the Educational Statistics course. I trust that you did well in the self-assessment items. Remember that in Session 1, you first learnt the meaning of the term, Statistics and applied it to understand the second term, Educational Statistics.

In this session, you will learn why you need to study Educational Statistics as a teacher. Many teachers wonder why they should be bothered with such a difficult subject. They forget that as teachers, we deal with a lot of numbers, especially test scores. These scores contain a lot of information we need to tap as teachers. This Session will explain to you further why as teachers you need a course in Educational Statistics.

Objectives

By the end of the session, you should be able to:

(a) use appropriate statistics to describe performance;
(b) interpret information from test scores;
(c) evaluate course grades;
(d) read and understand professional journals in education.
(e) describe how statistics is used in research.

Now read on...

2.1 Using appropriate statistics to describe performance

One reason why teachers need to study Educational Statistics is that it helps them to use the appropriate statistics in describing the performance of their classes to others. You will notice that teachers are often asked about the performance of their classes as a whole or performance in single subjects. The question often asked is, 'How did your class perform in the examination?' or 'How did your students perform in Mathematics?' The obvious answers are; the performance was bad or very bad. Others say the performance was good or very good. Well, these descriptions do not give us much information because the words, bad, very bad, good and very good are relative.

The most appropriate way to describe the performance of a class is to compare it with a known criterion or standard usually in the form of an average and say the performance of the class is above average, average or below average. If for example, the pass mark for a subject is 50 and a class obtains an average (mean) of 65, the teacher can say that performance is above average rather than saying performance was good. From now on, when you are asked about the performance of your class resist the temptation of saying, bad or good and rather say above average or below average. We shall learn more about the averages later in Unit 4.

2.2 Understanding information from test scores

A second reason why teachers must study Educational Statistics is that it puts them in a position to better understand the information they receive from test scores on students. As a class teacher, you are required to give quizzes, assignments, projects, class tests and end-of-term examinations which must be marked and scores given. These scores have a lot of information in them which are useful for decision making on individual pupils/students and the class as a whole. Sadly enough, most teachers do not look for the information to help them to improve the teaching and learning in their classrooms.

Pause for a minute and reflect. What are the purposes of assessing pupils/students?

The test scores provide information to teachers for planning and organizing instruction, making selection and placement, classification and certification decisions as well as motivating and guiding students. These purposes can only be achieved if teachers are able to study test scores and not just leave them. This course will teach you the various ways by which you can understand information from test scores. For example, after a class test, you can compute the mean and the median (these will be studied in Unit 4). If the mean is greater than the median, that tells you that, in general the performance of the class is low. As a teacher, you may need to give remedial classes or re-teach some topics.

2.3 Evaluating course grades

Another reason why teachers need to study Educational Statistics is that it helps them to evaluate course grades and the differences in ability represented by different grades. In the Primary schools, the same teacher teaches different subjects to the same pupils. However, the pupils do not perform at the same level in each subject. Ali may obtain 80% in English Language but 50% in Mathematics. The teacher is tempted to say that the performance is better in English Language than Mathematics. This information, however, **may not** be accurate.

Suppose in English Language, Ali's position is 18^{th} out of 40 in the class and in Mathematics, the position is 3^{rd} out of 40; then Ali in fact is better in Mathematics than in English Language, considering the performance of the class in both subjects. This is a better information for the classroom teacher. For students' personal report cards, grades alone do not provide enough information on a student's level of performance in a subject. This information should be combined with the ranking in the subject. It is important that teachers provide the ranking in each subject in addition to the raw scores. I hope that after this course you will be providing both raw scores and subject rankings to your pupils if you have not been doing that.

2.4 Reading and understanding professional journals

Educational Statistics also helps the teacher in the critical reading and understanding of professional journals in education. Education is dynamic and changes are always occurring. Do you know that what we now call teaching/learning materials (TLMs) used to be known as teaching apparatus? Do you also know that teaching has changed from teacher-centred to student-centred? Some of you may realize that when you were in the teacher training college, now changed to colleges of education, there was no continuous assessment.

What other changes in education can you think of since you became a teacher? Pause for a minute and reflect.

These changes in education are disseminated through many channels. One of the main channels is professional journals in education. These journals publish articles on new ideas and discoveries and best practices in the field of education. To be abreast with the times, teachers need to read the professional journals in education. However, to be able to have a full understanding of the articles, a basic knowledge of statistics is needed. Journals such as the Journal of Educational Management, Journal of Educational Research, Journal of Educational Development and Practice, Journal of Research and Development in Education often use statistics in their analysis of results.

2.5 Carrying out research

Another reason why teachers must study Educational Statistics is that it provides the basic statistical tools for the collection, analysis and interpretation of research data. Most teachers carry out Educational research either on their own or as part of a team on a project. Others may also embark on further studies which may require them to write a research project, dissertation or thesis to partially fulfil the requirement for the award of the degree they seek for. In all these, it is expected that students apply their knowledge acquired in Educational Statistics in conducting the research via scientific processes including data collection and analysis and interpretation of data for decision making.

Why is it important for teachers to carry out research? Pause for a minute and reflect.

Close your module. Write down at <u>least two</u> reasons why teachers should carry out research in education in your jotter. When you are done, open your module.

Now read on to find out whether your reasons are included below.
The reasons why teachers must carry out research include:

1. It provides an increase in knowledge about educational issues.
2. It helps to improve the practice of education.
3. It provides research results that inform policy issues.

As part of the research process, data needs to be collected and analysed. The analysis of research data requires statistical tools such as measures of central tendency, variability and correlation. You have just seen that research is important for every teacher therefore you need to acquire the basic statistical tools in this course to be able to carry out meaningful research.

In this session, you have learnt about the reasons why teachers need to study Educational Statistics as a course. The reasons are:

1. It helps teachers to use the appropriate statistics in describing the performance of their classes to others. (2). It puts teachers in a position to better understand the information they receive from test scores on students. (3). It helps teachers to evaluate course grades and the differences

in ability represented by different grades. (4). It helps the teacher in the critical reading and understanding of professional journals in education. (5). It is useful for research purposes.

Self-Assessment Questions

Exercise 1.2

For each item in 1 - 5, select and circle the best or correct option.

1. Statistics is important for classroom teachers because it…
    A. enables them to write appropriate objectives.
    B. helps them to construct good test items.
    C. helps them to evaluate students' grades.
    D. is useful for promotion and certification.

2. Educational research relies on Educational Statistics mainly for…
    A. directions for literature review.
    B. ideas for problem identification.
    C. tools for data analysis.
    D. tools for writing hypothesis.

3. Teachers can understand information from professional journals in education better if they…
    A. buy the journals regularly.
    B. carry out their own research.
    C. formulate educational policies.
    D. have basic knowledge in statistics.

4. Classroom teachers need a course in Educational Statistics because it will help them to…
    A. acquire knowledge for passing promotion interviews.
    B. describe class performance more appropriately.
    C. determine the best teaching method for a class.
    D. maintain discipline in the classroom.

5. Educational Statistics helps classroom teachers to…
    A. describe the steps in selecting a educational objectives.
    B. identify procedures and strategies for motivating students.
    C. provide avenues for student participation in classroom activities.
    D. understand differences in ability represented by different grades.

SESSION 3: VARIABLES, DATA AND DATA SOURCES

You are welcome to the third session of Unit 1 for the Educational Statistics course. I trust that you did well in the self-assessment items in Session 2 where you studied the reasons why classroom teachers need to take a course in Educational Statistics. I believe you are now convinced that this course is very important for you as a teacher. In this session, we will look at some important concepts that will be used in this course. You will learn about what data is, the types of data and sources of data. Before we start our lesson on what data are, it is important you understand the concept of a variable. Thus, in this third session, you will also learn about variables and their various classifications. Understanding these variables will lay the foundation for future computations, analysis and interpretations of statistical data to be able to make the right decisions.

Objectives

By the end of the session, you should be able to

(a) define the term, variable;
(b) distinguish between categorical and numerical variables;
(c) distinguish between ordered and unordered variables;
(d) distinguish between discrete and continuous variables;
(e) identify the sources and types of data used in Statistics.

Now read on...

3.1 What is a variable?

You know that for every human being or object, there are attributes, traits or characteristics. For human beings, there is height, weight, colour of skin, natural colour of hair, age, and speed in running a 100-metre race. What other traits can you think of? Likewise, for objects like tables, books and stones, such characteristics can be found. You can determine the height of a table, the colour of a book, and the weight of a stone. Some of these attributes can take on different values in an individual or group of individuals or in an object or group of objects while other attributes cannot take on different values.

Which attributes can take on different values in an individual or object or groups of individuals or groups of objects? Pause for a minute and reflect.

Now write down in your jotter at <u>least three</u> attributes of human beings that can take on different values and <u>at least one</u> attribute that cannot take on different values in an individual.

We trust that you have written them down. Now read on to find out which attributes can take on different values and those that cannot. Attributes that take on different values include height, weight, age, speed and those that do not take on different values include natural colour white and natural colour of finger nails. Any characteristic or attribute of an individual or object that can take on different values is called a variable. A value is an assigned number or label representing the attribute of a given individual or object or a group of individuals or objects. Alice's *age*, for example, can vary from 10 years through to 20 years. A group of girls can have ages such as 15 years, 18 years, 12 years, 17 years and 10 years. For a group of men, *marital status* as a variable, can be broken down into categories and given values as never married - 1, married - 2, divorced – 3. *Number of children in a family* as a variable can have values such as 0, 1, 2, 3, 4 etc. *Height* can take on values such as 1.2 metres, 1.7 metres, 2.0 metres and 2.2 metres. *Religious affiliation among a group of teachers* can be broken down into categories and given values such as: Christian – 1, Moslem – 2, Traditionalist – 3, Buddhist - 4.

3.2 Categorical and numerical variables

Variables can be classified into two main types: Categorical (qualitative) and Numerical (quantitative). Categorical (qualitative) variables have values that can only be placed into categories. For example, students in a class would have a gender as "male-1" or "female-2". Categorical (qualitative) variable cannot be added; Numbers are assigned to make analyses easier. You must however note that assigning numbers to qualitative variables do not necessarily make them quantitative in nature. They are just qualitative variables that have been assigned numbers. Numerical (quantitative) variables have values that represent quantities. These types of variables can be added. Some examples of quantitative variables are height, shoe size and age of students in your classroom. Quantitative variables can be grouped into two: discrete and continuous variables.

Write down in your jotter two variables each that you consider as categorical and numerical and bring it to face to face for discussion.

We can also further classify variables as ordered and unordered as well as discrete and continuous. The next two sub-sections explain these classifications

3.3 Ordered and unordered variables

This classification considers whether the variable has a quantitative or qualitative dimension.
Ordered variables are those variables where the attributes differ in magnitude along a quantitative dimension. These variables represent counts of the attribute and one can say that one value is more or less than the other because it has more or less of the attributes. The number of pupils in classes in a school can be counted as 20, 25, 32, 40, and 45. One can say that the class with 25 pupils has a higher enrolment than the class with 20 pupils. Likewise, the class with 32 pupils has a lower enrolment than the class with 45 pupils. If your age is 30 and your brother's age is 33, we can say that your brother is older than you or you are younger than your brother. For ordered variables it is possible to compare the values.

Unordered variables are those variables where the attributes are classified into two or more mutually exclusive categories that are qualitatively different. Numbers are used as labels or symbols to represent the categories. The numbers cannot be compared because they are only labels. In a class of pupils, their gender can be classified as female and male. Labels such as

female (1) and male (2) can be assigned to the categories. In a school, pupils can be grouped into colours such as red, blue, green and yellow for competitions. Numbers can be assigned to these colour groups as red (1), blue (2), green (3) and yellow (4).

3.4 Discrete and continuous variables

This classification considers the values that the variable assumes on the number line.

Discrete variables are those variables that have values, which in theory, assume only certain distinct values or whole numbers on a number line. These variables usually represent counts of indivisible entities. The number of pupils in each senior high school in Cape Coast can only take values that are distinct whole numbers such as 1545, 900, 850, and 1800. It is not possible to have a school with $467\frac{1}{2}$ pupils or $570\frac{3}{4}$ pupils. There can only be whole numbers or distinct numbers. In a football game the number of goals scored can only be 0, 1, 2, 3, 4, etc. It is impossible to score $3\frac{3}{8}$ goals.

Write down in your jotter three variables that you consider as discrete and bring it to face to face for discussion.

Continuous variables are those variables that have values which in theory; assume any value on a number line between two points. The values can differ by very small amounts and can be expressed as decimal fractions. Your height can be written as 1.885 metres. The weight of a table can be measured as 21.02kgs. What other continuous variables can you think of?

Pause and reflect on what continuous variables are. Now write down in your jotter three variables that you consider as continuous and bring it to face to face for discussion.

3.5 Data and data sources

In 3.1 of this unit, we provided the definition of what a variable is. We established that a variable is any characteristics or attribute of an individual or object that can take on different values. The different values associated with a variable is referred to as data. For example, the heights (in cm) of all 25 pupils in the class constitute a data (here, the height of the pupils is serving as the variable and the different values assigned for the height of the 25 pupils form the data set).

Data source is the "who" (or "what") that supplies the data (in this case, the pupils) and then the data collector is the one using the data for analysis. Data can be grouped in two categories – first-hand data called the primary data and the second-hand data called the secondary data. The primary data is provided by people who have experienced some phenomenon directly. Here, the researcher collects the data for the purpose of dealing with a problem at hand. Secondary data is an indirect account of a phenomenon. In this case, the data have already been collected for the purpose other than the problem at hand. The person performing data analysis is not the data collector.

Now, let's look at some examples of primary and secondary sources of data.

Some examples of primary data sources include: data from a political survey, data collected from an experiment, any observed data. Secondary data sources include: census data, data from print journals or data published on the internet

Pause and reflect on the advantages of using primary or secondary data. Write down at <u>least two</u> advantages each of primary and secondary data.

Now read on to find out whether your answers are included below.

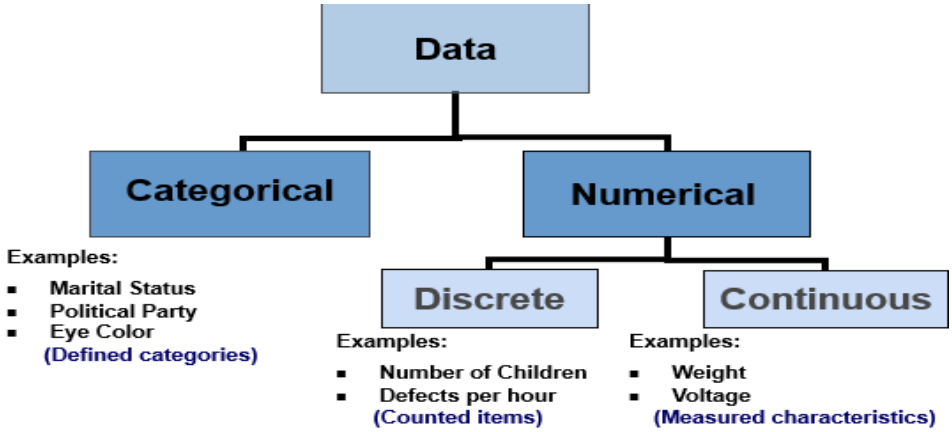
Advantages of primary data include:

1. Data is current.
2. Data is relevant and answers specific research problem.
3. Source of data is known.
4. Secrecy can be maintained.

Advantages of secondary data include:

1. It is less expensive to collect secondary data.
2. It involves less time to gather.
3. It is convenient.
4. Some information is available only from secondary data sources e.g., market shares.

In this session, you have learnt about what data is and the different sources of data. You have also learnt what variables are and the different categorizations. You learnt that data are the different values associated with a variable. The diagram below provides a summary of the types of variables learnt in this unit.



We have discussed the two (main) types of variables: categorical (qualitative) and numerical (quantitative). We have also discussed the differences between discrete and continuous variables as well as ordered and unordered variables.

Now attempt the self-assessment questions.

Self-Assessment Questions

Exercise 1.3

1. Which one of the following variables can be classified as unordered?
    A. Enrolment in a course
    B. Halls of residence
    C. Height of a student
    D. Scores in a quiz

2. Which one of the following variables can be classified as ordered?
    A. Models of mobile phones in UCC.
    B. Number of students in a class.
    C. Religious denominations in UCC.
    D. World Cup soccer groupings.

3. Which one of the following variables can be classified as unordered?
    A. Ages of students in the EPS 211 class in UCC.
    B. Names of houses of residence in a Senior High School.
    C. Number of students in each class in a Senior High School.
    D. Points obtained by houses in inter-house athletics competition.

4. Which one of the following variables can be classified as continuous?
    A. Age of students in a Psychology class.
    B. Colour of dresses students wear to class.
    C. Number of photocopy machines on a college campus.
    D. Percentage of female students in the University of Cape Coast.

5. Which one of the following variables can be classified as ordered?
    A. Parents occupation
    B. Region of country
    C. Religious affiliation
    D. Statistics achievement

6. An example of a variable that can <u>best</u> be classified as discrete is
    A. dancing ability of students.
    B. number of computers in a school.
    C. speed in a 200 metre race.
    D. study habits of students.

7. Which of the following measurement is categorical?
    A. Age of students in a class
    B. Favourite colour of students in a class
    C. Height of students in a class
    D. Scores the students obtained in a class test

For each item in 8 and 9, indicate whether the item is False or True by circling your choice.

8. A qualitative variable is non-numeric.
    A. False
    B. True

9. A quantitative variable can assume only whole number values.
    A. False
    B. True

SESSION 4: DESCRIPTIVE AND INFERENTIAL STATISTICS

You are welcome to the fourth session of Unit 1 for the Educational Statistics course. I trust that you are having a good understanding of the concepts studied so far. In the last session, we looked at the important terms we would use in the course. It is important that you have a firm grasp of these concepts. In this session, you will learn about the two major categories of Statistics.

Objectives

By the end of the session, you should be able to

(a) explain what descriptive statistics are;
(b) give examples of descriptive statistics;
(c) explain what inferential statistics are;
(d) give examples of inferential statistics.

Now read on...

4.1 Descriptive statistics

In Session 1, you studied the different definitions of the term, statistics.

Can you remember what these definitions are? Pause for a minute and reflect.

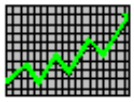
Now write down these definitions in your jotter.

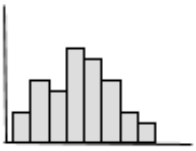
When you have finished writing the definitions, turn to Session 1 and check what you have written with the given definitions. You will notice that one of the definitions stated that statistics is a collection of two or more values computed from a set of data. This definition leads us to what descriptive statistics is.

Descriptive statistics provides summary data or information about a group. A single number or graph is often used to describe the group. Thus, descriptive data involves ccollecting, summarizing, and describing data.

1. Collect data
   - ex. Survey

2. Present data
   - ex. Tables and graphs





3. Characterize data
   - ex. Sample mean $= \frac{\sum X_i}{n}$

For instance, if there are 40 pupils in a Primary 6 class with ages that range from 11 to 17 as follows:

12	14	17	12	15	11	16	17	11	15
11	12	13	15	14	11	12	13	11	12
14	16	14	13	12	14	15	17	11	12
12	11	13	17	16	14	12	13	14	15

Then you can summarise this data by presenting them in a table or chart or by computing a single number to describe all the ages. If you want to represent all these 40 ages by one age which should be representative of all the ages, what would you do? One thing you can do is to add all the ages and divide by 40. This will give you what is known as the mean age. You can also go through the ages and find which age occurs most. This is what is known as the modal age.

Pause for a minute and calculate the mean age and obtain the modal age.

Now write down the values in your jotter.

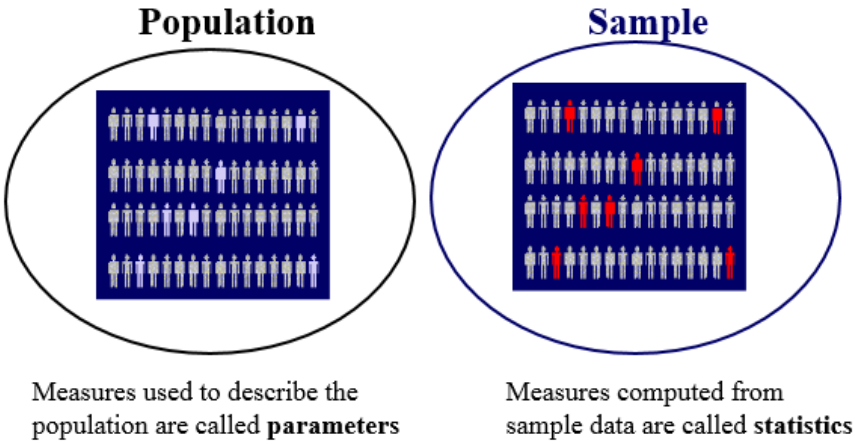
The answers are: Mean age - 13.5 years and modal age – 12 years. Have you got these values correct? If no, then go back and check your calculations. Since these values are used to describe the group of ages, they are descriptive statistics.

Other descriptive statistics include computations of proportion, median, range, standard deviation, variance and quartiles and correlation coefficients. You will learn more about how to characterize data and represent data using tables and graphs in subsequent units for this Educational Statistics course.

4.2 Inferential statistics

Sometimes, we may want to obtain information about a large group of people or objects. For example, a regional director of education may want to know about the ages of the parents of all the pupils in the basic schools in the region. This will be a difficult task to perform. The director may request the head teachers to ask the pupils to obtain the information from their parents or ask circuit supervisors to visit homes and obtain the information. However, there will be difficulties in the data collection process. Some parents may not remember or even know their ages.

To deal with a situation like this, the regional director may take a few districts and use a few circuit supervisors to obtain the information. On the basis of the information obtained, the regional director may infer the age of the parents in the region. All the parents in the region would be referred to as the population. The small group from whom the information is obtained is known as the sample. This situation is referred to as inferential statistics.



How would you define inferential statistics? Pause for a minute and reflect on the situation described above.

Now write down the definition of inferential statistics. Read on and find out if your definition is correct.

Inferential statistics is the use of information or data from a small group called a sample to make conclusions or generalizations about a much larger group called a population from which the sample is taken.

Let us look at three examples.

Example 1

A testing organization has trained a group of examiners to mark scripts. Each examiner was given a large number of scripts to mark under a leader. After marking, all the scripts are submitted to the team leader who has to make sure that there is consistency in the marking of the scripts. Because of the large number of scripts, it will not be in his interest to re-mark all the scripts. For each examiner, he chooses a few scripts from each packet and mark. Based on the closeness of the marks, the team leader may conclude as to whether there is consistency in the marking or not. Here we can say that he used information from the sample of scripts to make a conclusion on the population of scripts of the marker.

Example 2

A pharmacist has produced a drug that he claims can cure asthma. He administers the drug on 20 asthma patients and they were all cured. He therefore confirms that his drug is potent and can cure asthma. In this scenario, the pharmacist has not cured all asthma patients but just a group of 20. Based on the results from the group of 20, he makes a conclusion concerning the effectiveness of the drug to cure all asthma patients.

Example 3

A researcher wishes to know the mean age of all first year university students in all the universities in Ghana. Considering the cost involved and the risk of travelling, he obtains the ages of a group of 1400 first year students from Legon, Valley View University, All Nations University and the University of Cape Coast and obtains the mean. Based on the results, he could get an estimate of the mean age of all the first-year university students in Ghana. In this scenario, the group of 1400 students constitute the sample and the first-year university students in Ghana constitute the population. Information from the sample is used to make an inference about the mean age of the first-year students.

In this session, you have learnt about the definitions of descriptive statistics and inferential statistics and noted examples of each type. You learnt that descriptive statistics is a type of statistics that has to deal with collecting, summarizing, and describing data whereas inferential statistics deals with drawing conclusions and/or making decisions concerning a population based only on sample data.

Now attempt the self-assessment questions.

Self-Assessment Questions

Exercise 1.4

For each item in 1 - 6, select and circle the best or correct option.

1. The percentage of students in JHS 1 in Kafodidi who wear reading glasses is an example of descriptive statistics.
    A. False
    B. True

2. A researcher found that in a school 40% of the pupils in Class one had lice. He concluded that lice are very common in the whole school. This is an example of inferential statistics.
    A. False
    B. True

3. The mean height of 100 pupils in Tampa Senior High School is an example of inferential statistics.
    A. False
    B. True

4. The use of samples to make conclusions about populations from which the samples are drawn is referred to as statistical inference.
    A. False
    B. True

5. The Principal of a College of Education found 10 students at the student clinic suffering from diarrhoea. He concluded that there is an outbreak of diarrhoea in the college. This is an example of descriptive statistics.
    A. False
    B. True

6. The use of graphs to depict the performance of students in the BECE Mathematics paper is an example of inferential statistics.
    A. False
    B. True

SESSION 5: NOMINAL AND ORDINAL SCALES OF MEASUREMENT

You are welcome to the fifth session of Unit 1 for the Educational Statistics course. I trust that you did well in the self-assessment items in Session 4 where you studied the two broad categories of statistics. In this session, you will learn about scales of measurement. This concept is important in further analysis of statistical data.

Objectives

By the end of the session, you should be able to

(a) explain the term, scales of measurement;
(b) describe the characteristics of the nominal scale;
(c) describe the characteristics of the ordinal scale.

Now read on...

5.1 Scales of Measurement

The idea of scales of measurement was first developed by the Psychologist Stanley Smith Stevens in 1946. The concept, scales of measurement, refers to the categorization of variables or numbers according to specific properties. Depending on the traits/attributes/characteristics of variables and the way they are measured, different kinds of data result representing different scales of measurement. A number can only be used meaningfully for analysis if the scale of measurement is known. For example, the number 4 can be categorized in different ways. It may represent the colour blue if colours are given numbers such:

Red	1
Green	2
Yellow	3
Blue	4
Black	5

It may also represent the position of an athlete in a race like 1^{st}, 2^{nd}, 3^{rd}, 4^{th}. It is also used for the actual value of a variable such as the weight of a set of books being 4 kilograms or a score of 4 out of 10 in a quiz. Each scale of measurement has certain properties which in turn determine the appropriateness for use in certain statistical analyses.

There are 4 types of measurement scales. These are Nominal, Ordinal, Interval and Ratio.

5.2 Nominal Scales

A nominal scale classifies the elements in a variable into two or more categories in which no ranking is implied. Each category or group may be assigned a number for the purpose of identification. All those in a particular category must have a common characteristic, trait or attribute. Whatever the classification, a person/object can be in one and only one category (i.e. they must be mutually exclusive) and members of a given category have a common set of characteristics. One other property is that the categories must be exhaustive. There must be enough categories for every possible observation. The only mathematical operation that can be applied is

counting. The numbers cannot be added, subtracted, divided or multiplied. A nominal scale is the lowest of the four levels of measurement.

Here are examples of nominal scale variables.

1. Gender Female (assigned 1) Male (assigned 2).
   Note that all the females have a common human characteristic that makes them female and the males have a common human characteristic that makes them male. The numbers assigned to them cannot be added. If 1 is added to 2, the result is 3 and this is meaningless. A normal person can belong to only one category. It is not normal to have one part of your body female and another part male. The numbers assigned can also be interchanged without any impact on the data. Female can be 2 and male can be 1.
2. Regions in Ghana
3. Registered political parties in Ghana
4. Halls of residence in the University of Cape Coast
5. Mobile phone networks in Ghana
6. Marital status in a distance education class
7. Religious affiliation in a community

Write down four examples of variables with a nominal scale and state the categories for each. Bring your answers to FTF for discussion.

5.3 Ordinal Scales

An ordinal scale classifies the elements in a variable and also ranks them in terms of the degree to which they possess the characteristic or attribute of interest. An ordinal scale puts the elements in order from highest to lowest, or from most to least. The numbers represent a quality being measured and can tell us whether a case has **more** of the quality measured **or less** of the quality measured than another case. Though ordinal scales do indicate that some elements are better than others, they do not indicate how much better. The intervals between the ranks are not equal. The only mathematical operation that can be applied is ranking or order of merit. The numbers cannot be added, subtracted, divided or multiplied. If the numbers are changed, there is an impact on the data. They are used for unordered variables.

Here are some examples for you.

1. Professional qualification of teachers (Teacher, Superintendent, Senior Superintendent, Principal Superintendent, Assistant Director, Director)
2. Positions in inter-schools athletics competition (1^{st} Anfo Senior High, 2^{nd} Broni Senior High, 3^{rd} Dakye Senior High, 4^{th} Jada Senior High School)
3. Junior High BECE Performance in Kako District (1^{st} Babbu JHS, 2^{nd} Loli JHS, 3^{rd} Peto JHS, 4^{th} Lego JHS, 5^{th} Cape JHS, 6^{th} Mankess JHS, Agoga JHS, 7^{th} Abbira JHS)
4. Students' grades (A, B, C, D, F)

Write down two examples of variables with an ordinal scale and bring your answers to FTF for discussion.

In this session, you have learnt about two scales of measurement. These are the nominal scale and the ordinal scale. The properties of the nominal scale are that, elements in a variable are classified into two or more categories, elements must be mutually exclusive, members of a given category have a common set of characteristics, categories must be exhaustive, the only mathematical operation that can be applied is counting and they are used for unordered variables. The properties of the ordinal scale are that elements in a variable are ranked in terms of the degree to which they possess the characteristic or attribute of interest, numbers represent quality not quantity, numbers do not indicate how much better one member is than the other, the intervals between the ranks are not equal, the only mathematical operation that can be applied is ranking or order of merit. The numbers cannot be added, subtracted, divided or multiplied and if the numbers are changed, there is an impact on the data.

Self-Assessment Questions

Exercise 1.5

1. A district director of education measures many variables on a sample of schools. An example of a variable measured in an ordinal scale is the
    A. enrolment of the classes in each school.
    B. income in cedis of the teachers.
    C. professional qualification of the teachers.
    D. years of service for each teacher.

2. A study was conducted to see how well reading success in primary three could be predicted from various kinds of information obtained in kindergarten (reading readiness, age, gender, and socio-economic status). Which of the variables represents a nominal scale?
    A. Age
    B. Gender
    C. Reading readiness
    D. Reading success

3. Which one of the following variables **cannot** be classified as nominal?
    A. Parents occupation
    B. Region of country
    C. Religious affiliation
    D. Statistics achievement

4. One property of the nominal scale is that
    A. categories must be exhaustive.
    B. intervals between the ranks are equal.
    C. numbers represent quality not quantity.

D. they are used for unordered variables.

5. Which one of the following variables can be classified as ordinal?
   A. Ages of students in the EPS 211 class in UCC.
   B. Names of houses of residence in a Senior High School.
   C. Number of students in each class in a Senior High School.
   D. Positions of houses in inter-house athletics competition.

6. Anastacia was third in the Miss Tourism competition. This is an example of an ordinal scale.
   A. False
   B. True

SESSION 6 INTERVAL AND RATIO SCALES OF MEASUREMENT

You are welcome to the last session of Unit 1 for the Educational Statistics course. I trust that you have understood the basic statistical concepts in this unit and have done well in the self-assessment items.
In this session, you will learn about two other scales of measurement. These are the interval and ratio scales of measurement.

Objectives
By the end of the session, you should be able to
(a) describe the characteristics of the interval scale,
(b) give examples of the variables that use the interval scale,
(c) describe the characteristics of the ratio scale,
(d) give examples of the variables that use the ratio scale.

Now read on...

6.1 Interval Scales
An interval scale has all the characteristics of both nominal and ordinal scales. Elements with the same characteristics are in the same category and one value can be considered better than the other. In addition, the interval scale has equal intervals. For example, if dancing ability of primary six pupils is measured on an interval scale, then a difference between the scores of 15 and 16 is the same as the difference between 24 and 25. The numbers in the scale depict values or quantities of the elements in the variables. A value of zero is possible but it is arbitrary and does not mean the absence of the characteristic/trait. In the dancing ability example, if a primary six pupil scores zero in a dance, it does not mean that the pupil cannot dance. It may only mean that the pupils did not take the required steps needed. Values can be added and subtracted to and from each other but not multiplied or divided. When numbers are changed, there is an impact on the data.

Pause here for a minute. Read over again the properties of a variable with an interval scale.

Here are some examples for you.

1. Academic achievement in a class.
   In classroom testing, test items are based on what the teacher teaches. If a student attends class regularly and obtains a score of zero, it does not mean that the pupil does not know anything concerning the topics covered by the test items. It may mean that the pupil has not studied the content covered by the test well. The difference between a score of 45 and 48 (i.e. 3 points) is the same as the difference between a score of 72 and 75 (i.e. 3 points). However, if a pupil obtains a score of 80, it cannot be interpreted as being twice the score of 40, because interval scores do not allow for multiplication and division.
2. Fahrenheit scale for temperature.

Differences on this scale represent equal differences in temperature. The difference in temperature between 48^0 and 49^0 is the same as the difference between 72^0 and 73^0, but a temperature of 30 degrees is not twice as warm as one of 15 degrees.

3. English reading ability of Junior High School (JHS) pupils.
   JHS pupils have had more than six years of instruction in English language. A test of their ability to read English will provide a level of performance indicated by the score given to them. A pupil who scores zero will not mean that the pupil cannot read English passages at all. A pupil with a score of 15 knows he has done better than a pupil who scores 10. All pupils in with the score of 10 are in the same category.

Take a few minutes and write down two examples of variables with an interval scale. Bring your answer to FTF for discussion.

1.2 Ratio scales

A ratio scale has all the characteristics of the nominal, ordinal scales and interval scales. Elements with the same characteristic are in the same category and one value can be considered better than the other. The intervals are all equal. A value of zero is possible and it is absolute or true, meaning the complete absence of the variable. Here, none of the scale exists. Values can be added, subtracted, multiplied and divided. For example, if a lady is 30 years old, she is twice as old as a 15-year-old girl and is 10 years older than a 20-year-old girl. When numbers are changed, there is an impact on the data.

Pause here for a minute. Read over again the properties of a variable with a ratio scale.

Here are some examples.

1. Kelvin scale for temperature.
   Differences on this scale represent equal differences in temperature. This scale has an absolute zero. Thus, a temperature of 300 Kelvin is twice as high as a temperature of 150 Kelvin. The difference in temperature between 40^0 and 50^0 is the same as the difference between 70^0 and 80^0.

2. Ages (in years) of pupils.
   A class three pupil who is 10 years is twice as old as a class one pupil who is 5 years old. A person of age 0 years is not born.
3. The height of tables.
   A table which is 2 metres tall is shorter than a table which is 4 metres tall. The 4-metre table is twice as tall as the 2-metre table. A table which is 0 metre tall does not exist.

Take a few minutes and write down four examples of variables with a ratio scale. Bring your answers to FTF for discussion.

In this session, you have learnt about two additional scales of measurement. These are the interval and ratio scales. The scales have all things in common except for only one difference. The interval scale has an arbitrary zero while

the ratio scale has an absolute zero. Common properties of interval and ratio scales are that the elements with the same characteristic are in one category, values can be ordered, one value is better than another, a zero is possible, values can be added and subtracted, intervals are equal and there is an impact on data when numbers are changed.

Self-Assessment Questions

Exercise 1.6

1. A principal of a college of education measures many variables in the college. An example of a variable measured in a ratio scale is the …
    A. enrolment of students in each class.
    B. gender of the teachers.
    C. professional qualification of the teachers.
    D. teaching ability of the teachers.

2. A study was conducted to see how well reading success in primary three could be predicted from various kinds of information obtained in kindergarten (reading ability, age, gender, and socio-economic status). Which of the variables represents a ratio scale?
    A. Age
    B. Gender
    C. Reading ability
    D. Socio-economic status

3. Which one of the following variables can be classified as interval?
    A. Parents occupation
    B. Region of country
    C. Religious affiliation
    D. Statistics achievement

4. Which one of the following variables is **best** measured among senior high students using an interval scale?
    A. Favourite food
    B. Musical ability
    C. Religious affiliation
    D. Year of birth

5. The grades, A, B, C, D, E, F in a test were changed to 90, 80, 70, 60, 50 for statistical purposes. The new scores are an example of a ratio scale.
    A. False
    B. True

6. The number of cars on the University of Cape Coast campus is an example of an interval scale.
    A. False
    B. True

This is a blank sheet for your short notes on:

- issues that are not clear, and
- difficult topics, if any.

UNIT 2: DATA REPRESENTATION: ORGANIZING CATEGORICAL DATA

Unit Outline

Session 1: Purposes and principles of data representation
Session 2: Organizing categorical data: Summary table
Session 3: Organizing categorical data: Bar charts
Session 4: Organizing categorical data: Pie graphs
Session 5: Organizing categorical data: Line Graphs
Session 6: Organizing categorical data: Pareto Diagram

Hello! Welcome to the second unit of this course in Educational Statistics. I believe you enjoyed reading Unit 1 which introduced you to the important concepts and ideas in educational statistics. I trust that interest in the course has been generated in you and that you will enjoy the rest of the units.

You will realize that the classroom teacher collects a lot of data on the pupils and information is buried in the data. In this Unit, we will try to unearth information from data on pupils and students. We will focus mainly on categorical data. The major tools we shall use are summary tables, bar graphs, line graphs, pie charts as well as the pareto diagram. By observing data in a pictorial form, information is easily and better grasped and understood.

Unit Objectives

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

1. Describe the purposes and principles of data representation;
2. Construct bar graphs and state uses in education;
3. Construct pie charts and state uses in education;
4. Construct line graphs and state uses in education;
5. Construct summary tables;
6. Construct pareto diagrams.

SESSION 1: PURPOSES AND PRINCIPLES OF DATA REPRESENTATION

You are welcome to the first session of Unit 2 for the Educational Statistics course. As noted in the introduction to the Unit, representing data by graphics, pictures or tables makes it easier to grasp the information the data contains. It also allows more information to be derived from the data. In educational statistics, a lot of emphasis is placed on pictorial representations. This session deals with the purposes of representing data pictorially and the principles that should be followed in data representation.

Objectives

By the end of the session, you should be able to

(a) state and explain four purposes of pictorial representation of data,
(b) describe seven principles of representing data pictorially.

Now read on...

1.1 Purposes of pictorial representation of data

There are 4 purposes of pictorial representations of data. These four purposes are described below. Read them carefully and make sure you understand them.

1.1.1 Purpose One

Pictorial representation makes sets of data easier to keep in mind.

Educational statistics usually involves large sets of data. In a senior high school for example, enrolment of pupils may be over 1000 students in a school. There are over 19,000 pupils in primary schools in Ghana. The ages of these pupils will be a large data set. When raw data is presented to us, it often makes us wonder what we can do with it. However, if the same data set is presented as a chart, graph or table, we are able to make sense out of it and be able to keep it in mind.

1.1.2 Purpose Two

Pictorial representation helps to identify relationships and trends.

Graphs and charts show very clearly any relations that exist between figures. When raw data is available, it is difficult to identify any relationships and trends at first glance. For those who are not number-friendly, large data sets scare them but charts and graphs attract them. Generally, the larger the set of numbers the better it is to use graphs and charts to identify relationships.

1.1.3 Purpose Three

Pictorial representation makes data sets easier to understand.

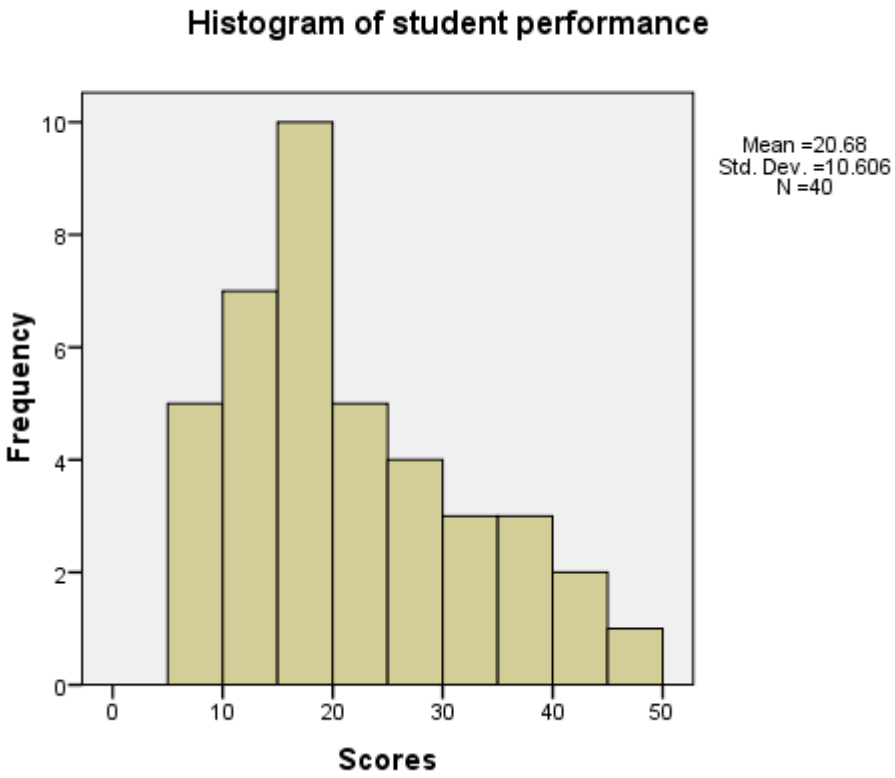
Data sets contain a lot of information. However, raw data is difficult to understand especially if you are not a statistician. For example, in a class of 40 students, the following scores may be obtained in a class test where the maximum score is 50.

18	15	12	45	20	12	17	25	42	33
30	28	10	14	20	13	28	8	15	18
20	9	16	12	40	38	32	5	17	16
27	6	18	20	35	38	8	12	15	20

Pause for a minute. Look at the data again.

What can you tell about the performance of the class? Is the class performance low, moderate or high?

Now look at the histogram below.



You can see from the histogram that a lot of the students had scores less than 20, showing that performance is low. This information is not easy to get from the raw data.

1.1.4 Purpose Four

Pictorial representation makes it easier to compare values from categories.

Educational statistics often comes in terms of groups and categories. A teacher may want to compare performances in a mathematics examination in the three form two classes he is teaching. A municipal director of education may want to compare overall performances in the West African School Certificate Examination (WASSCE) in the schools in the municipality. A headmistress may want to compare the enrolment of boys and girls in her school. These situations will best be shown by a graph or a chart rather than raw data.

1.2 Principles of pictorial representation of data

Graphs and charts are to make it easier to understand information from raw data. To be able to achieve this objective seven principles must be followed in creating graphs and charts.

1.2.1 Principle One

The chart or graph should have a title. The title provides summary information of the content of the chart/graph. It should be clear, simple, appropriate and accurate.

1.2.2 Principle Two

The chart or graph should be appropriate for the data. For example, raw scores from an examination is an interval scale data and it is not appropriate to use bar graphs.

1.2.3 Principle Three

The chart or graph should be adequately labelled. The horizontal axis as well as the vertical axes should be clearly spelt out. Without a description of the axes, it will be difficult to interpret the chart or graph.

1.2.4 Principle Four

The chart or graph should carry all the information that is necessary so that it can be interpreted meaningfully. For example, to compare performance in three classes, all the raw scores from the three classes should be available. If only two classes are provided, the comparison will not be complete.

1.2.5 Principle Five

The chart or graph should be easier to understand than the raw data they represent. They should not be loaded with so much data that it will take an expert to derive information from it. The chart or graph should be simple and clear in meaning.

1.2.6 Principle Six

The chart or graph should have a key which may be included below or within the body of the chart. The key should provide information on what the parts of the chart/graph represent.

1.2.7 Principle Seven

The chart or graph should have footnotes or source notes, when appropriate and these should be numbered (when necessary) and placed below the chart or graph.

In this session, you have learnt about four purposes of pictorial representation of raw data and seven principles to be followed in using charts and graphs to represent data. In the next session, we shall learn about the types of bar graphs you can use to represent your data so that you can obtain meaningful information for decision making.

Self-Assessment Questions

Exercise 2.1

For each of the following items, indicate whether the item is **false** or **true** by circling the correct response.

1. One purpose of using charts and graphs to represent data is to make it easy to compare groups that provide the data.
    A. False
    B. True

2. Graphs are easier to interpret if they do not have key.
    A. False
    B. True

3. One principle for charts is that it should have a high degree of validity.
    A. False
    B. True

4. Graphs are important in teaching and learning because they determine the teaching method to adopt for a class.
    A. False
    B. True

5. Graphs and charts can be used to compare the performance of boys and girls in a Basic Education Certificate Examination.
    A. True
    B. False

6. Charts and graphs should have head notes that are numbered and placed at the top of the diagram.

A. True
B. False

SESSION 2: ORGANIZING CATEGORICAL DATA: SUMMARY TABLE

For you to use and interpret data appropriately, you need to be able to recognize and understand the different types of data. In unit one, session three of this module we learnt about the sources and types of data. This session will further help you to typically identify and develop tables that are appropriate for categorical data. You have learnt that categorical variables represent types of data which may be divided into groups. Examples of categorical variables are race, sex, age group, and educational level. The purpose of presenting data in tables is to facilitate analysis of data, give effective interpretation of data, and provide rapid communication on complex issues and situations. Using tables also help to highlight cases that are clustered in a particular range of scores. This session will focus on organising categorical data using a summary table.

Objectives

By the end of the session, you should be able to

a) describe the benefits a summary table has over a raw data set,
b) explain the meaning of a summary table, and
c) investigate a data set by creating a summary table from it.

Now read on...

2.1 What is a Summary Table?

A summary table is a visualization that summarizes statistical information about data in table form. It indicates the frequency, amount, percentage of items in a set of categories so that one can see differences between categories.

It is important to understand that a summary table produces summary data of values for one categorical variable or the intersection of two or more categorical variables. In this session, we will focus on two types of summary tables:1) A summary table with only one categorical variable and 2) A summary table with exactly two categorical variables.

2.1.1 One -way Summary Table

A one-way summary table presents a categorical data with only one variable by counting the number of observations that fall into each category of the variable. For example, suppose a survey was conducted of a group of 20 students in a Senior High School to determine their religious background and the following data were collected (Here the categorical variable is the religious background or type):

Christian	Traditional	Christian	Traditional
Moslem	Christian	Moslem	Moslem
Traditional	Atheist	Traditional	Traditional
Atheist	Christian	Christian	Christian
Christian	Traditional	Moslem	Moslem

Now the question is: what can we do with this data? One thing we can do is to present the data in a table. We can create a summary table to enhance the visualization of the data. Table 2.1 is a summary table that shows the distribution of religious background of the 20 students who took part in a survey.

Table 2.1. Distribution of Religious Background

Religion	Number of Students
Christian	7
Moslem	5
Traditional	6
Atheist	2

In the simple summary table with one variable, the different categories of the variable are represented in one column of the table and the values for each category or group obtained from the raw data represented in another column. In the example in Table 2.1, there are four categories of the variable (Religion) namely: Christian, Moslem Atheist and the Traditional in one column and the number of students drawn from each category are represented in another column. Notice that, this is actually a new data that was computed from the raw data. In this format, it is easy to identify trends or patterns in the underlying data. For instance, we can observe that the religion which reported the highest number of students is the 'Christian' religion while the lowest was reported in the 'Atheist' category.

Now we have learnt what a summary table is and how to create a one-way summary table. Pause for a few minutes. What do you think would be the uses of a one -way summary table in education?

Now read on to find out whether your reasons are included below. The importance of using a one-way summary table in education includes:

- It provides a way to visualize data.
- It allows you to see things in the data you might otherwise not see.
- It allows you to manipulate and create new data.
- It helps you look at your data in new ways.

2.1.2 Contingency Table

You have learnt how to represent data on a table when only one categorical variable is involved. We will now learn about summary tables with two categorical variables. A summary table of two categorical variables is known as a two- way summary table. It is can also be termed as cross-classification table. This is because when the values for two variables intersect, the variables are said to be crossed and the process of crossing variables to form intersections is called Cross-tabulation. Another name given to such a two-way summary table is the Contingency table.

The following are true about a cross-classification (or contingency) table.

1. A cross-classification (or contingency) table presents the results of two categorical variables.
2. The joint responses are classified so that the categories of one variable are located in the rows and the categories of the other variable are located in the columns.
3. The cell is the intersection of the row and column and the value in the cell represents the data corresponding to that specific pairing of row and column categories.

Now let us look at an example. In a school the 405 students are put in Houses and Forms. A two-way table presenting the results might appear as follows:

Form (variable 2)	House (variable 1)			
	Nkrumah	Busia	Limann	Danquah
One	38	31	32	30
Two	32	42	33	30
Three	30	40	35	32

The summary table displays how the two variables: House and Form are crossed by highlighting a single value for each variable. For example, we can see from the table that the value 38 is highlighted where Nkrumah House intersects with the form One. This value is found in the left, first, topmost cell and it represents the number of students from Nkrumah House who are in Form One. This means that the number of students from Nkrumah House who are in Form One is 38. Similarly, 31 represents the number of students from Busia House who are in Form One.

Now do the following. First, draw the contingency table in a jotter. Then determine the number of students who belong to each of the Houses (i.e Nkrumah, Busia, Limann and Danquah) Repeat the exercise for the Forms. How can you confirm that the total number of students is 405? Can you describe the procedure?

Now read on to find out whether you used any of the procedure below:

1. To find the number of students in each House, add the numbers of students in each column without accounting for the effect of the other variable (in the example above, the total number of students in Nkrumah House, regardless of Form, is 38+32+30= 100).

2. To find the number of students in each Form, add the numbers of students in each row without accounting for the effect of the other variable (in the example above, the total number of students in Form One, regardless House, is 38+31+32+30 = 131).

3. The new table will then look like this:

Form (variable 2)	House (variable 1)				**Total**
	Nkrumah	Nkrumah	Nkrumah	Nkrumah	
One	38	31	32	30	**131**
Two	32	42	33	30	**137**
Three	30	40	35	32	**137**
Total	**100**	**113**	**100**	**92**	**405**

4. To find the overall total number of students add the totals of students in each column or row without accounting for the effect of the other variable (in the example above, the overall total number of students is 405. This is either 131+137+137, regardless of House or 100+113+100+92, regardless of the Form.

I hope you still remember the uses of a one-way summary table in education that we discussed earlier in this session. Similarly, we can talk about uses of the contingency table in education. Write down four uses of the contingency table in education. Bring your answers to FTF for discussion.

In this session, you have learnt about the types summary tables. We discussed a one-way summary table and a contingency table. A contingency table, also called a two-way summary table, is built to actually arrange data into two groups so they will be easier to view and analyse. Summary tables show the data that is collected in a way that makes it easier to compare. When we have a large data set, a lot of work and computation needs to be done to create a summary table, manually. However, we can use data manipulating tools like Excel to create summary tables quickly.

Self-Assessment Questions

Exercise 2.2

1. The following table provides information about the educational background of a group of people who applied to teach in a pre-school.

Education	Number
Master	12
Bachelor	23
Diploma	32
High school or less	55

a. Identify the type of variable provided by the information in the first column of the table.
b. How many people applied to teach in the pre-school?

For each of the following items, indicate whether the item is **false** or **true** by circling the correct response.

2. Summary tables are used because they automatically detect and highlight potential trends or patterns in the underlying raw data.
   A. False
   B. True

3. Summary tables are used to generate a summarized view of a large dataset which is helpful for gaining insight
   A. False
   B. True

SESSION 3: ORGANIZING CATEGORICAL DATA: BAR GRAPHS

I hope you have grasped the purposes and principles of the pictorial representation of data. Now we are going to learn how to construct specific graphs and how you use them in teaching and learning. In this session, we shall study the bar graph. We shall learn how to construct the bar graph, the types of bar graphs that are most useful in education and the strengths and limitations of the bar graph. Data that are from categorical variables (nominal scales) are represented in graphic form with the use of bar graphs. Bar graphs give a pictorial description of the data and emphasize how groups compare with one another. They are used to compare the sizes of the various parts. The height of the bars is the basis for the comparisons and not the area of the bars.

Objectives

By the end of the session, you should be able to

a) describe three types of column bar graphs,
b) construct bar graphs,
c) state and explain the strengths and limitations of bar graphs, and
d) discuss the uses of bar graphs.

Now read on...

3.1 Types of bar graphs

Bar graphs are either column or horizontal in shape. Column graphs are more popular in education. In this course, our emphasis is on column graphs. There are three common column bar graphs. These are (1) simple bar graph, (2) compound or multiple bar graph and (3) component bar graph. Examples are shown below.

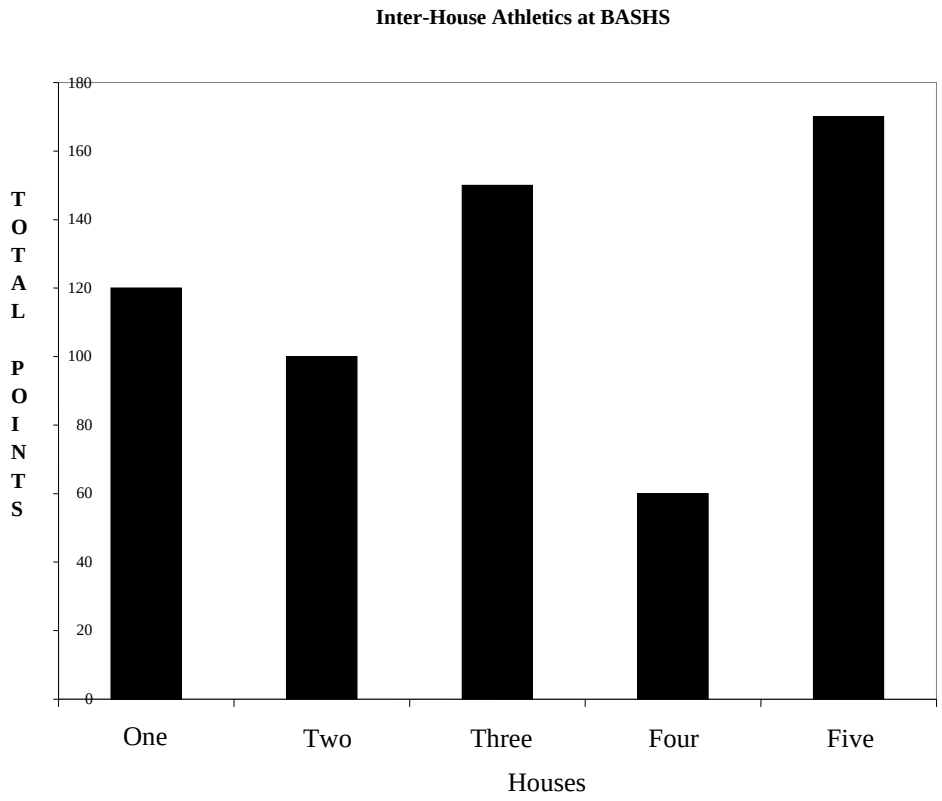
3.1.1 Simple column bar graph

In the simple bar graph, single bars are drawn to represent the values for each category or group. In the example below, points are obtained in an inter-house athletics competition at Brabo Ahenkro Senior High School (BASHS) and this is presented in Table 2.2. In the simple bar graph, bars are drawn for each house and the height of the bar corresponds to the total points obtained in the competition.

Table 2.2. Performance in Inter-House Athletics at BASHS

House	Total Points
One	120
Two	100
Three	150
Four	60
Five	170

The figure below is the **simple column** bar graph showing performance in the Inter-House Athletics competition at BASHS.



3.1.2 Compound or Multiple column bar graph

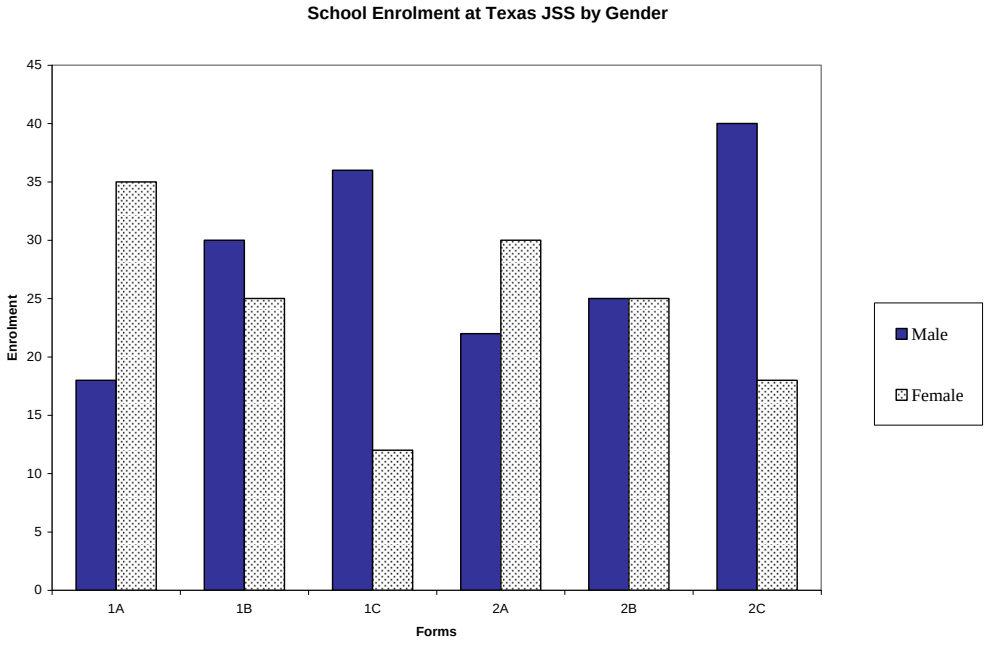
The compound or multiple column bar graph is used when there are sub-categories within a major category. For example, in the inter-house athletics competitions, there could be boys and girls in each house. In a mixed school, there would be boys and girls in a class. In a form one class, there would be representatives of various houses. If the purpose of the information desired is to compare the sub-categories, then the compound or multiple bar graph is most appropriate.

In the example below in Table 2.3, school enrolment at Texas Junior High School is provided for male and female students. In the compound or multiple bar graph, bars are drawn for the males and females for each form and the height of the bars corresponds to the enrolments. Here the total enrolment in each class is not of great concern.

Table 2.3 School Enrolment at Texas JHS

Form	Male	Female
1A	18	35
1B	30	25
1C	36	12
2A	22	30
2B	25	25
2C	40	18

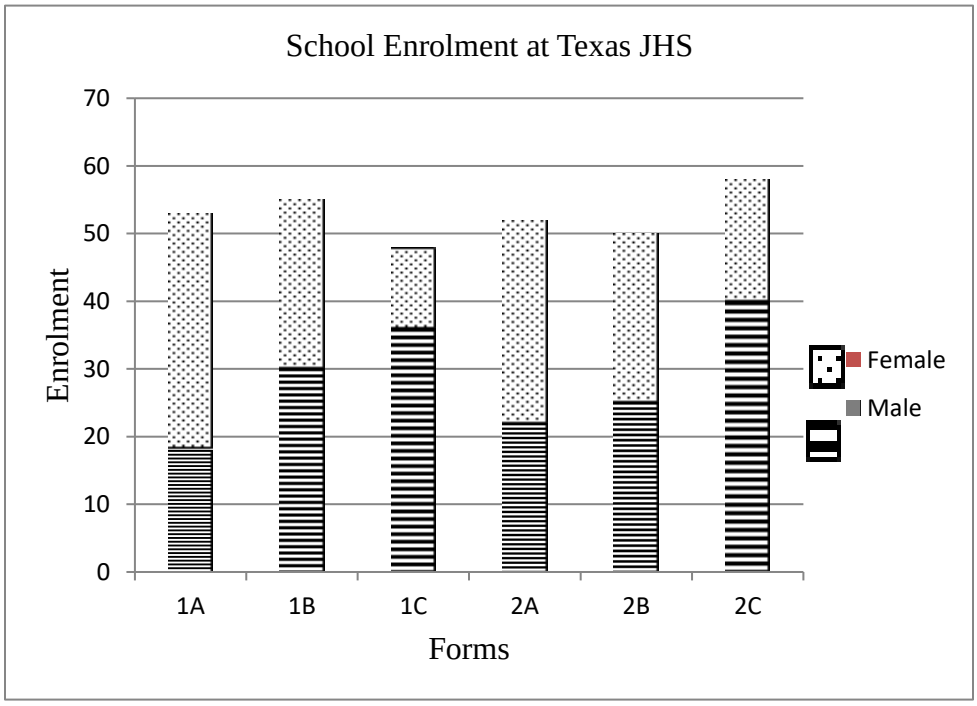
The compound or multiple column bar graph is drawn below.



3.1.3 Component Bar Graph

It is known as composite or stacked bar chart. It is used when a set of data from sub-categories combines to form a total. For example, in the University of Cape Coast, enrolment by Colleges and Schools would combine to get the total enrolment in the University. In the school districts, total pupil enrolments in each circuit would combine to get the total enrolment for the entire district. The total of the values is the length/height of the bar. Component bar graphs allow for visual comparisons between different components (i.e. how components contribute to the total of the category) as well as comparing the overall totals for the categories.

The enrolment figures in Table 2.3 are used to construct the bar graph below.



3.2 Constructing bar graphs

Four major steps are involved in the construction of bar graphs manually. It is recommended that graph sheets be used. Where computer software such as Microsoft Excel and SPSS are available, they must be used since they make the graphs look neater and more accurate.

Step 1

Draw two axes, a vertical and horizontal. Label the vertical axis by the source of the values/scores e.g. enrolment, points, etc. Label the horizontal axis by the names of the categories, e.g. houses, forms or gender.

Step 2

Divide the vertical scale into units considering the lowest value and the highest value. Choose appropriate scales such that the bars are not too tall or too short and must start with zero.

Step 3

Construct equally wide and equally spaced bars for each category with the height of the bar being the value/score for the category on the horizontal axis, which has the names of the categories as the label.

Step 4

Shade or colour the bars to differentiate bars and components. Contrasting colours should be used for the compound and component bars.

I hope you will be able to follow the steps above and draw bar graphs.

The following results were obtained by students in Chuga Senior High School. Represent the result by a compound bar graph. Use Microsoft Excel or a graph sheet. Bring the graph to FTF for discussion.

	Percentage of passes	
Subject	Boys	Girls
Mathematics	85	48
English Language	60	85
Integrated Science	75	55
Social Studies	82	90
Visual Art	90	60

Now let us look at the strengths and limitations of bar graphs.

3.3 Strengths and limitations

Bar graphs have a number of strengths and limitations. These are described below.

1. Bar graphs are easy to draw. They do not need technical expertise. In these days of computer software, drawing graphs has become much easier. I will recommend that you learn how to use Microsoft Excel to draw graphs.
2. It is easy to read values easily from the vertical axis. The vertical axis is scaled in units and this allows easy reading of the values for the corresponding categories on the horizontal axis.
3. It is easy to make comparisons between bars thus making the significance of the information obtained easily grasped.
4. Bar graphs are best used for nominal scale variables. They are not effective with variables that are in the interval and ratio scales of measurement.
5. Some data contain extreme values. These values make it difficult to construct good graphs. They make some bars too short and some very tall, thus distorting the comparisons.
6. Component and compound bar graphs may have too many subgroups. This makes the graph crowded, information not easily derived and comparisons not clear.

Now we have learnt the types of bar graphs, how to construct them and their strengths and limitations. Pause for a few minutes. What do you think would be the uses of bar graphs in education?

Take a few minutes and write down two uses of bar graphs in education. Close the module before writing down the uses in your jotter. When you are done, open the module and read on.

Now compare what you have written with the uses below.

3.4 Uses of bar graphs

Teachers and educational workers can use bar graphs in several ways. Some of the ways are listed below.

1. Enrolment of students by classes. For example, in a school, the headmaster/headmistress can draw a bar graph of the enrolment by forms say, Form 1, Form 2, Form 3, Form 4.
2. Enrolment of students by faculties. In the University of Cape Coast, there are academic Colleges, faculties and schools. A bar graph can be used to represent the enrolment by colleges, faculties and schools.
3. In the district education offices, circuit supervisors can draw bar graphs of the BECE results by schools.
4. Inter-house athletic competitions, where points are awarded. These can be represented by bar graphs.
5. Literacy rates, defined in Ghana as the percentage of people over the age of 15 who can read and write, can be represented by regions. Literacy rates can also be represented by bar graphs using the districts in each region.
6. Percentage of passes in the subjects offered in a school in the West African Senior High School Certificate Examination (WASSCE) can be represented by bar graphs.

In this session, you have learnt about the types of bar graphs and the four steps in constructing bar graphs. In addition, you have learnt about the limitations and the strengths of bar graphs as well as the uses of bar graphs. In the next session, we shall learn about the pie chart.

Self-Assessment Questions

Exercise 2.3

1. Bar graphs are most useful for representing data when the scale of measurement is
   A. interval
   B. nominal
   C. ordinal
   D. ratio

2. One strength of bar graphs is that they
   A. allow values to be read for the categories.
   B. are effective with interval scales of measurement.
   C. make provision for scores that are extreme.
   D. provide trends of performance in schools.

3. Bar graphs can be used to compare
   A. ages of heads of schools in a district.

B. attendance at one PTA meeting.
C. BECE results from 1999-2009 in a school.
D. school fees paid by each student in a school.

4. One weakness of bar graphs is that
   A. extreme values distort comparisons.
   B. they are easy to construct.
   C. they are effective with nominal scales.
   D. values can be read for categories.

5. In constructing bar graphs, spaces between bars must be equal.
   A. True
   B. False

6. For component bar graphs, many subgroups can be handled effectively for information.
   A. True
   B. False

SESSION 4 ORGANIZING CATEGORICAL DATA: PIE CHARTS

I trust that you had a good time with the bar graphs and had success in drawing the compound bar graph. In this session, we shall study the pie chart. We shall learn how to construct the pie chart, identify the strengths and limitations as well as the uses of the pie chart.

Objectives

By the end of the session, you should be able to

(a) construct a pie chart,
(b) state and explain the strengths and limitations of pie charts, and
(c) discuss the uses of pie charts.

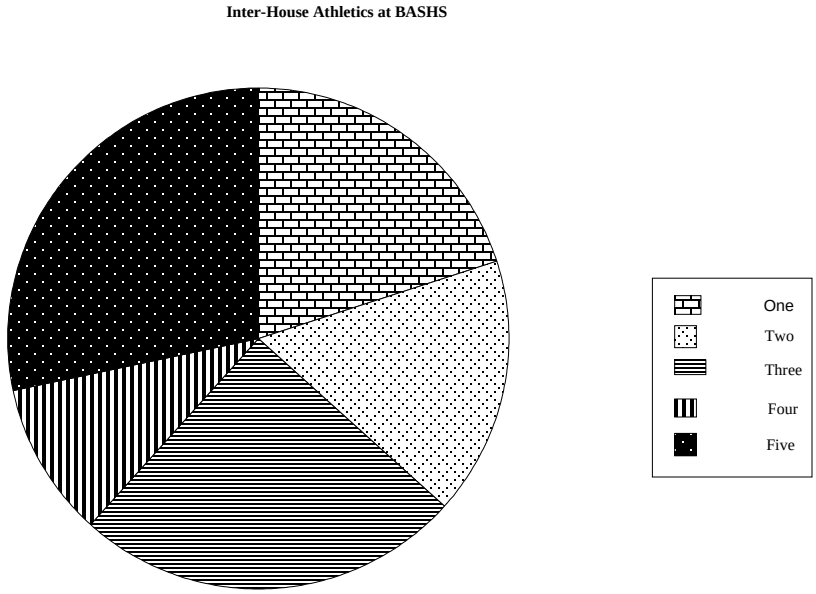
Now read on...

4.1 Nature of Pie charts

Pie charts use nominal or categorical data. Pie charts are represented in the form of a circle of 360^0 sliced into the shape of 'pies'. Each pie is cut from an angle at the centre of the circle. The angle corresponds to the data for each category or group. Pie charts give a pictorial view and the contributions of the parts that make a whole. An example is shown below based on Table 2.4.

Table 2.4 Performance in Inter House Athletics at BASHS

House	Total Points	Degrees
One	120	72
Two	100	60
Three	150	90
Four	60	36
Five	170	102
Total	600	360



4.2 Constructing pie charts

There are three simple steps to take in the construction of a pie chart. These three steps are described below. Where computer softwares such as Microsoft Excel and SPSS are available, they must be used since they make the graphs look neater and more accurate.

Step 1

Calculate the degree equivalents for the value of each category/group by dividing the total points for each group by the overall total points and multiplying the result by 360^0.

For the example above, the degrees representing the points for the houses are calculated as follows:

House One: $\frac{120}{600} \times 360^0 = 72^0$ House Two: $\frac{100}{600} \times 360^0 = 60^0$

House Three: $\frac{150}{600} \times 360^0 = 90^0$ House Four: $\frac{60}{600} \times 360^0 = 36^0$

House Five: $\frac{170}{600} \times 360^0 = 102^0$

Step 2

With the use of a pair of compasses and protractor draw a circle and measure the sectors (pies) based on the degrees calculated.

Step 3

Shade or colour the sectors to differentiate one from the other. Contrasting colours should be used to make the differences look clearer.

I trust that you will be able to follow the steps above and draw pie charts.

Try the exercise below and bring your attempt to FTF for discussion.

The following results were declared by the University of Tata in 2009. Represent the results by a pie chart. Use Microsoft Excel or a plain sheet of paper.

Class	Number of students
First Class	145
2nd Class Upper	360
2nd Class Lower	518
Third Class	100
Pass	70

Now let us look at the strengths and limitations of pie charts.

4.3 Strengths and limitations

Pie charts have a number of strengths and limitations. These are described below.

1. Pie charts are appropriate for variables that are from nominal scales.
2. In the pie chart, individual parts of the whole are seen and can be compared.
3. Pie charts provide a visual impression of the proportion that each part contributes to the overall total.
4. The angles in the chart are harder to compare, especially where they are small.
5. Pie charts are not easy to draw manually. Angles need to be calculated first and if a mistake is made wrong impressions are carried.
6. Pie charts are not suitable for data that are continuous and of ratio and interval scales of measurement.
7. The values of each category cannot be read from the chart but must be provided.
8. Pie charts are not useful where there are many parts. These come out as very small sectors and makes comparisons difficult.
9. Pie charts give only a visual impression of the raw data provided but not the details of the data.

Now we have learnt about the nature of the pie chart, how to construct it and the strengths and limitations. Pause for a few minutes. What do you think would be the uses of pie charts in education?

Take a few minutes and write down, in your jotter, two uses of pie charts in education.

Now compare what you have written with the uses below.

4.4 Uses of pie charts

Pie charts can be used by teachers and educational practitioners in several ways. Some of the ways are described below.

1. Degree classifications in the final examinations (1st class, 2nd class upper, 2nd class lower, third class and pass) in a particular year can be represented by pie charts.
2. The enrolment of students by classes in a particular school. For example, in a school, the headmaster can draw a pie chart of the enrolment by forms say, Form 1, Form 2, Form 3, and Form 4.
3. The enrolment of students by departments in a Faculty. In the University of Cape Coast, there are academic departments in the Faculty of Educational Foundations. A pie chart can be used to represent the enrolment by departments.
4. In the district education offices, circuit supervisors can draw pie charts of the BECE results by schools in each circuit.
5. Inter-house competitions, for example athletics, where points are awarded. These can be represented by pie charts.
6. Teacher professional qualifications in a district can be represented by the pie chart. The number of teachers for categories as superintendents, senior superintendents, principal superintendents, and assistant directors can be used for drawing the pie chart.

In this session, you have learnt about the pie chart. We have discussed the nature of pie charts, the steps in the construction of pie charts, the strengths and limitations of pie charts and the uses of pie charts. In the next session, we shall learn about line graphs.

Self-Assessment Questions

Exercise 2.4

1. Pie charts are most useful for representing data when the scale of measurement is
   A. interval
   B. nominal
   C. ordinal
   D. ratio

2. One strength of pie charts is that
   A. the contribution of each part can be visualized.
   B. the values of each component cannot be read.
   C. they are effective with all scales of measurement.
   D. they make provision for scores that are extreme.

3. Pie charts cannot be used to compare
    A. ages of heads of schools in a district.
    B. attendance at one PTA meeting.
    C. BECE results from 1999-2009 in a school.
    D. enrolment in the classes in a school.

4. One weakness of pie charts is that
    A. extreme values distort comparisons.
    B. they are easy to construct.
    C. they are effective with nominal scales.
    D. values cannot be read for categories.

5. In constructing pie charts, degree equivalents of the values of each category are used.
    A. True
    B. False

6. Data in continuous form are **not** appropriate for pie charts.
    A. True
    B. False

SESSION 5: ORGANIZING CATEGORICAL DATA: LINE GRAPHS

Hello! Welcome to Session 5 of Unit 2. So far, we have studied the different types of bar graphs and pie charts. In this session, you will be introduced to line graphs. We shall study the different types of line graphs and how to construct them. We shall also look at the strengths and limitations and how you can use them as an educational practitioner.

Objectives

By the end of the session, you should be able to

(a) describe two types of line graphs,
(b) construct line graphs,
(c) state and explain the strengths and limitations of line graphs,
(d) explain the uses of line graphs.

Now read on...

5.1 Types of line graphs

Data that are related to time are best used for line graphs. Time could be days, weeks, months and years. Line graphs show changes in the data over a period of time. Data from interval and ratio scales are most appropriate. Line graphs could be simple or compound. Simple line graphs give a pictorial description of the data. Compound line graphs compare group data over a period of time. Examples are shown below.

5.1.1 Simple line graph

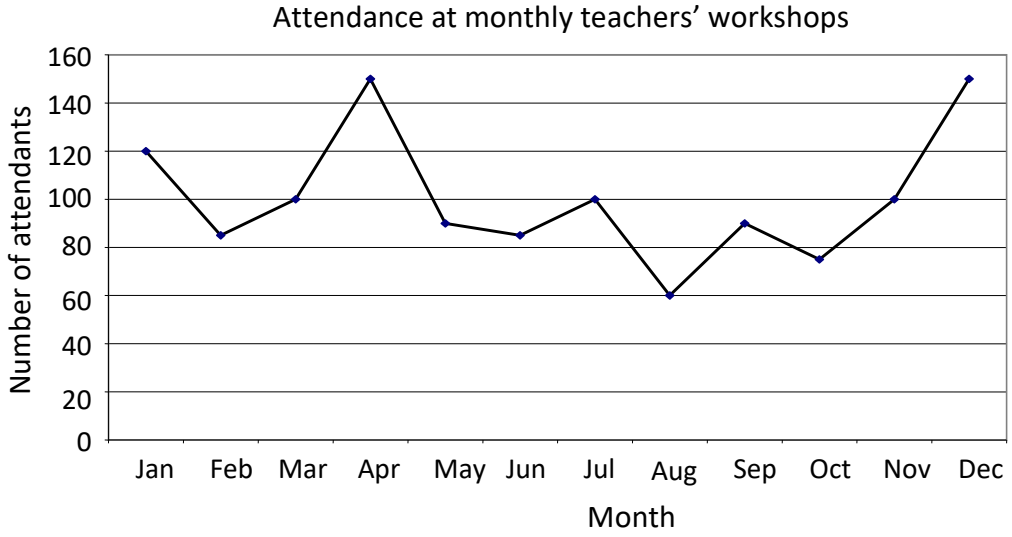
In the simple line graph, dots or points are made on the graphs to represent the values for each category or group. These dots are then joined together with straight lines. In the example below, in Table 2.5, the total numbers of teachers that attended a workshop for each month are obtained. Considering the units on the vertical scale, dots are made on a graph and then joined together with straight lines representing the number of teachers for each month.

Table 2.5 Attendance at monthly teachers' workshops

Month	Total
January	120
February	85
March	100
April	150
May	90
June	85
July	100
August	60
September	90

October	75
November	100
December	150

The figure below is a **simple** line graph showing attendance at a monthly teachers' workshop.



5.1.2 Compound line graph

The compound line graph is used when there are sub-categories within a major category. For example, in the teacher's workshop, there could be male and female teachers. An educational practitioner may be interested in comparing the attendance patterns by gender. Thus, if the purpose of the line graph is to compare trend of the sub-categories within a major category, then the compound line graph is most appropriate.

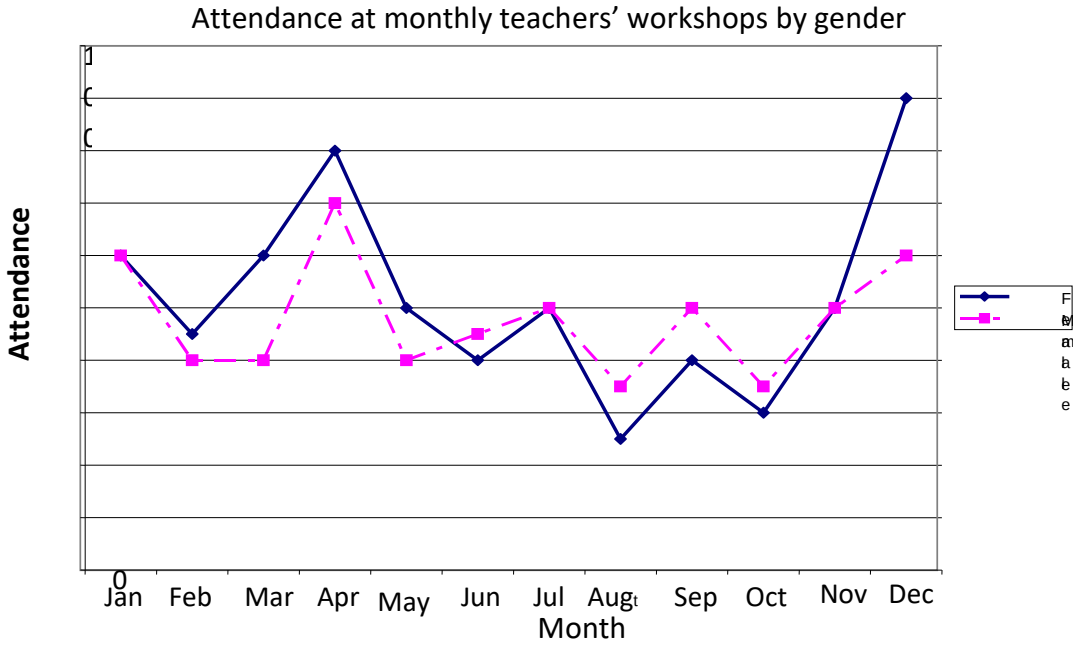
In the example below in Table 2.6, data on attendance at the workshop has been provided by gender. In the compound line graph, trends are drawn for the male and female participants. The number of males and females for each month are plotted on the graph and joined with straight lines.

Table 2.6 Attendance at monthly teachers' workshops

	Attendance	
Month	Female	Male
Jan	60	60
Feb	45	40
March	60	40
April	80	70
May	50	40
June	40	45
July	50	50
August	25	35

Sep	40	50
Oct	30	35
Nov	50	50
Dec	90	60

The line graph below shows the information provided in Table 5.2.



5.2 Constructing line graphs

Three major steps are involved in the construction of line graphs manually. It is recommended that graph sheets be used. Where computer software such as Microsoft Excel and SPSS are available, they must be used since they make the graphs look neater and more accurate.

Step 1

Draw two axes (a vertical axis and a horizontal axis). Label the vertical axis by the source of the values e.g. attendance, enrolment, scores etc. Label the horizontal axis by the time period e.g. years, months, days, weeks etc.

Step 2

Divide vertical scale by units or points, considering the lowest value and the highest value. Choose appropriate scales such that the graph is not too tall or too flat and must start with zero. Write down the time period categories on the horizontal axis.

Step 3

Plot the value or quantity for each time period on the graph and join all the points by a straight line.

I believe that you have followed the steps above and can draw line graphs. Do the exercise below and bring it to FTF for discussion.

The following results were declared by the Sunkwa Senior High School for Chemistry from 1999-2008. Represent the result by a line graph. Use Microsoft Excel or a plain sheet of paper.

Year	%age of passes
1999	58
2000	62
2001	70
2002	68
2003	55
2004	72
2005	78
2006	60
2007	70
2008	75

Now let us look at the strengths and limitations of line graphs.

5.3 Strengths and limitations of line graphs

1. Line graphs are best used for ratio and interval scale data and to some extent ordinal. They are not effective with variables that are in the nominal scale of measurement.
2. Values can be read easily from the vertical axis. The vertical axis is scaled in units and this allows for the easy reading of the values for the corresponding time periods on the horizontal axis.
3. Comparisons can be made easily among different groups and the significance of the information easily grasped.
4. Line graphs enable predictions to be made for information not yet available.
5. Some data contain extreme values. These values make it difficult to construct good graphs. They make either the graph either too short or too tall, thus distorting the comparisons.
6. Compound line graphs may have too many subgroups. This makes the graph to be crowded and information not easily derived and comparisons not clear.

Now we have learnt about the types of line graphs, how to construct a line graph and the strengths and limitations. Pause for a few minutes. What do you think would be the uses of line graphs in education?

Take a few minutes and write down two uses of line graphs in education.

Now compare what you have written with the uses below.

5.4 Uses of line graphs

Line graphs can be used by teachers and educational practitioners in several ways. Some of the ways are described below.

1. Examination results. Schools can find the trend of performance in examinations (eg BECE) over a period of time; say 2000 -2010 (a ten-year period).
2. School enrolment. Total enrolment in an institution can be studied over a period of years. Enrolment can be for the whole school or by class or subjects.
3. Attendance at workshops. The Ghana Education Service organizes workshops for teachers. To help in planning for future workshops, line graphs can be drawn to find the trend of attendance over a period of years.
4. Payment of school fees. During the first two weeks of reopening of schools, heads of institutions can obtain records of the number of students (or percentage of students) who pay school fees. This can be done over a period of years and a line graph drawn to study the trend of payment of school fees.
5. Teacher departure from the Ghana Education Service. The Ghana Education Service can obtain data on the number of teachers that leave the Service each year over a period of time and plot it on a line graph.

In this session, you have learnt about the line graph. We have discussed the nature of line graphs, the steps in the construction of line graphs, the strengths and limitations of line graphs and the uses of line graphs. In the next session, we shall learn about pareto diagrams.

Self-Assessment Questions

Exercise 2.5

1. One strength of line graphs is that
   A. extreme values cannot distort information.
   B. they are effective with all scales of measurement.
   C. trends over a period can be observed.
   D. values of each component cannot be read.

2. Line graphs can be used to compare
   A. admissions in a high school in a particular year.
   B. monthly attendance at PTA meetings.
   C. weekly BECE results in a school.
   D. teacher retention in a district in a day.

3. One weakness of line graphs is that
   A. extreme values distort comparisons.
   B. they are difficult to construct.

C. they are effective with ratio scales.
D. values can be read for categories.

4. Data on the number of teachers granted study leave with pay in 2010 can be used to construct line graphs.
    A. True
    B. False

5. Line graphs are **not** useful for representing data when the scale of measurement is ordinal.
    A. True
    B. False

SESSION 6: ORGANISING CATEGORICAL DATA: PARETO DIAGRAM

You are welcome to Session 6 of this unit. I trust that you have understood the pictorial representations of data we have studied in this unit. Remember we have described bar graphs, pie charts and line graphs. I believe that your attempt to construct these graphs and charts have been successful.

Another way of representing categorical data is by the use of pareto diagrams. In this session, we shall look at the nature of a pareto diagram, when to use it as well as how it looks like. Go gradually and re-read areas you do not understand well until you have grasped the content. Do not be in a hurry to complete the session.

Objectives

By the end of the session, you should be able to

(a) describe the nature of pareto diagram,
(b) explain when to use the pareto diagram.
(c) outline the steps involved in constructing the pareto diagram and
(d) list at least two uses of pareto diagrams in education.

Now read on...

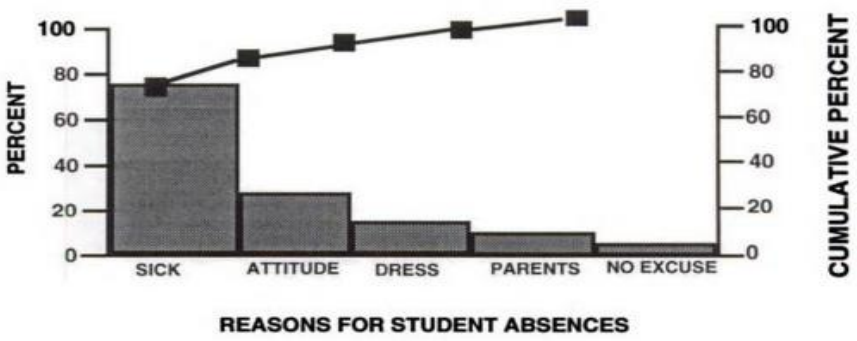
6.1 Nature of pareto diagram

A Pareto chart is a chart named after Vilfredo Pareto. It is also known as the pareto distribution diagram. I trust that remember the bar graph and line graph. If not, take some time to review this in previous sessions. You will learn in this session that the pareto chart contains both bars and a line graph. The pareto chart is a vertical bar graph in which values are plotted in decreasing order of relative frequency from left to right and the cumulative total is represented by the line (You will learn more about relative and cumulative frequencies in subsequent units). Pareto charts are extremely useful for analysing what problems need attention first because taller bars on the chart, (which represent frequency), clearly illustrate which variables have the greatest cumulative effect on a given system.

The pareto chart provides a graphic depiction of the pareto principle. The principle simply put states that: 80% of effects come from 20% of causes. In other words, most of what we do has little effect. Here is an example of what a pareto diagram looks like.

An example of a Pareto diagram

PARETO DIAGRAM



6.2 When to use a Pareto chart.

Pareto chart is used when you want to

- analyse data about the frequency of problems or causes in a process.
- analyse broad causes by looking at their specific components.
- focus on the most significant problem or cause from a list of many problems or causes.
- focus and prioritize your efforts.
- categorize data and determine the number of incidents in each category.
- communicate to others about your data.

6.3 How to construct a Pareto diagram

1. Collect your data.
2. Analyze and categorize your data.
3. Add your data to excel.
4. Sort your data in descending order.
5. Determine percentages.
6. Graph your data.
7. Polish your graph.
8. Communicate your findings.

You will have to learn the basics of excel in order to be able to construct a pareto diagram using excel. Try this in you free time. You can also construct the pareto diagram manually by following the steps below:

1. Decide what categories you will use to group items.
2. Decide what measurement is appropriate. Common measurements are frequency, quantity, cost and time.
3. Decide what period of time the pareto chart will cover: one work cycle? one full day? a week?
4. Collect the data, recording the category each time (or assemble data that already exist).
5. Subtotal the measurements for each category.
6. Determine the appropriate scale for the measurements you have collected. The maximum value will be the largest subtotal from step 5. (If you will do optional steps 8 and 9 below,

the maximum value will be the sum of all subtotals from step 5). Mark the scale on the left side of the chart.

7. Construct and label bars for each category. Place the tallest at the far left, then the next tallest to its right and so on. If there are many categories with small measurements, they can be grouped as "other."

Steps 8 and 9 are optional but are useful for analysis and communication.

8. Calculate the percentage for each category: the subtotal for that category divided by the total for all categories. Draw a right vertical axis and label it with percentages. Be sure the two scales match: For example, the left measurement that corresponds to one-half should be exactly opposite 50% on the right scale.
9. Calculate and draw cumulative sums: Add the subtotals for the first and second categories, and place a dot above the second bar indicating that sum. To that sum add the subtotal for the third category, and place a dot above the third bar for that new sum. Continue the process for all the bars. Connect the dots, starting at the top of the first bar. The last dot should reach 100 percent on the right scale

Now we have learnt about the nature of pareto chart, when to use it and how to construct. Pause for a few minutes. What do you think would be some of the application of pareto chart in education?

Take a few minutes and write down, in your jotter, two uses of pareto chart in education.

Now compare what you have written with the uses below.

6.4 Uses of Pareto Chart in education.

Pareto chart is useful to teachers and educational practitioners. A few of the uses are listed below.

1. A school administration can use pareto chart to determine the major causes of absenteeism of students to improve attendance.
2. An examination body like WAEC can use pareto chart to analyse the major causes of leakage of examination questions to solve the problem.
3. A teacher can improve the interest of students in a subject they are underperforming in by using pareto chart to determine main causes of disinterest in that particular subject.
4. Students can use pareto charts to determine the daily top activities that consume most of their time in order to prioritise well.

Take a 10 - minute break and refresh yourself. Now re-read sessions 6.2 and 6.3, going over the steps and noting the points.

Now, do the exercise below and bring to FTF for discussion.

The disciplinary committee of your school has noticed an increase in disciplinary issues at the school. The causes of indiscipline have been found to be favouritism by teachers, lack of enforcement of rules, lack of communication, lack of leadership, lack of motivation and bad habits. Given the data below, use your knowledge of pareto charts to help the committee decide on which factors to prioritise in dealing with indiscipline in order to improve discipline in your school.

Causes of indiscipline	No of cases reported per week
Favouritism by teachers ...	70
Lack of enforcement of rules	90
Lack of Communication: ...	10
Lack of leadership: ...	5
Lack of motivation: ...	25
Bad habits:	50

First, you must create a pareto chart and write down the conclusion you can draw from your graph. Second write down the advice you will give to the disciplinary committee based on the conclusions from your chart drawn.

In this session, you have learnt about pareto charts. You have noticed the nature of pareto charts, when to use a pareto chart, how to construct one and how it can be used in education. I hope the lesson has been beneficial and therefore you would be able to apply it whenever the need arises.

We have come to the end of this unit which is organizing categorical data. You have been introduced to how summary table, bar chart, pie chart, line graphs, and pareto charts are used to organise or present categorical data. In the next unit, we shall learn about how to represent numerical data. Until the next lesson, keep revising your notes and practising what you have learnt.

Self-Assessment Questions

Exercise 2.6

1. Pareto diagrams can be used in all these scenarios **except**, when analysing..……
    A. broad causes by looking at their specific components.
    B. broad problems by considering all possible causes.
    C. data about the frequency of problems or causes in a process.
    D. problems or causes and you want to focus on the most significant.

2. A pareto diagram consists of both a…….
    A. pie chart and line graph.
    B. pie chart and bar graph.
    C. line graph and bar graph.
    D. line graph and summary table.

For each of the following items, indicate whether the item is **false** or **true** by circling the correct response.

3. A pareto diagram can be used in taking decisions but not communicating data.
    A. True
    B. False

4. In a pareto diagram the order of the arrangement of bars does not matter.
    A. True
    B. False

5. The pareto principle operates based on the pareto diagram.
    A. True
    B. False

6. Pareto diagrams can be constructed both manually and with Excel.
    A. True
    B. False

UNIT 3: DATA REPRESENTATION: ORGANIZING NUMERICAL DATA

Unit Outline

Session 1: Organizing numerical data: Ordered Array
Session 2: Organizing numerical data: Stem and Leaf
Session 3: Organizing numerical data: Box and Whisker
Session 4: Organizing numerical data: Frequency distributions
Session 5: Organizing numerical data: Histogram, frequency Polygon and Ogive

Hello! Welcome to the third unit of this course in Educational Statistics. I believe you enjoyed reading Unit 2 which introduced you to how to organise categorical data. I trust that interest in the course has been generated in you and that you will enjoy the rest of the units.

In sections two to six of unit two, you were introduced to how summary table, bar chart, pie chart, line graphs, and Pareto chart are used to represent categorical data. In this section, you will be taken through how ordered array, stem-and-leaf plot, box-and-whisker plot, frequency distribution, histogram, frequency polygon and ogive are used to represent numerical data. By observing data in a pictorial form, information is easily and better grasped and understood.

Unit Objectives

By the end of this Unit, you should be able to present data using:

1. Ordered array
2. Stem-and leaf plot
3. Box-and-whisker plots
4. Frequency distribution
5. Histogram
6. Frequency polygon
7. Ogive

SESSION 1: ORGANIZING NUMERICAL DATA: OREDERED ARRAY

Hello! You are welcome to the first session of this unit. In this session, we shall study the ordered array. We shall learn what an ordered array is and how to construct it. The use of an ordered array is one of the ways to represent numerical data. Knowledge about ordered array is a fundamental concept which would be required as we advance. Therefore, pay attention and enjoy the lesson.

Objectives

By the end of the session, you should be able to

(a) explain what an ordered array is,
(b) construct an ordered array, and
(c) state strengths and limitations of ordered arrays.

Now read on...

1.1. What is an Ordered Array?

In statistics, an ordered array is a sequence of data in rank order. This means, it is a group of numbers arranged in rows and columns with the smallest at the beginning and the rest in order of size up to the largest at the end. It is important to note that in real life, data is recorded haphazardly. Therefore, whenever we want to sort numerical data neatly into rows and columns in an ascending order, we employ the ordered array plot. An ordered array therefore:

- shows range (minimum to maximum values).
- provides some signals about variability within the range.
- may help identify outliers (unusual observations).

Example

Given below are the marks (out of 25) obtained by 20 students of class 6A in mathematics in a test. 18, 16, 12, 10, 5, 5, 4, 19, 20, 10, 12, 12, 15, 15, 15, 8, 8, 8, 8, 16

The raw data when put in ascending or descending order of magnitude is called an array or arrayed data. 4, 5, 5, 8, 8, 8, 8, 10, 10, 12, 12, 12, 15, 15, 15, 16, 16, 18, 19, 20

In a diagram form, we may have:

Maths scores of class 6A	4	5	5	8	8	8
	8	10	10	12	12	12
	15	15	15	16	16	18
	19	20				

This is known as an ordered array of mathematics scores of class 6A.

1.2. Constructing an ordered array

There are two simple steps to take in the construction of an ordered array. These steps are described below. We would use the following example to illustrate the steps.

The students' record department of the University of Cape Coast surveyed the ages of some students on campus during registration. Study the data set and hence use an ordered array to represent it.

Ages of regular students	18,17,21,16,18,18,20,19,20,19,17,22,32,27,22,38,42
Ages of sandwich students	18,45,19,33,21,20,23,28,32,19,41,18

Step 1*:* Order or rank the data from lowest to highest.

In this situation, we will have:

Ages of regular students	16,17,17,18,18,18,19,19,20,20,21,22,22,27,32,38,42
Ages of sandwich students	18,18,19,19,20,21,23,28,32,33,41,45

Step 2: Create a table with columns and rows to represent the data. Remember to label it appropriately.

The ordered array for our example above will look like this:

Age of Surveyed UCC Students						
	Regular Students					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	27	32	38	42	
	Sandwich Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

This way, it is easy to explore the data. For example, we can easily identify the highest age for all the students. This is 45. The least age for the regular students is 16

Now we have learnt what an ordered array is and how to construct one. Pause for a few minutes. What do you think would be the strengths and limitations of an ordered array in?

Now read on to find out whether your reasons are included below.

1.3 Strengths and limitations of ordered array

Strengths of ordered array include:

- It provides a way to visualize data.
- It allows you to see things in the data you might otherwise not see.
- It helps you look at your data in new ways.

Limitations of ordered array include:

- If the data set is large, the ordered array is less useful.
- The order in which data was originally collected is lost.
- For a large data set, the ordered array occupies much space.

Now can you do the following exercise.
Draw an ordered array for the ages of mature students.
38, 42, 36, 28, 45, 30, 44, 32, 38, 27, 26, 42, 41, 31, 46, 38, 29, 34, 40, 48.
What age appears most? What age is the highest? What age is the lowest? Bring your answers to FTF for discussion.

In this session, you have learnt about ordered arrays. You have studied its nature, how to construct it and also its strengths and limitation. Basically we said it is about arranging values in a particular order in columns and rows. You would need the concept of ordered array in stem and leaf plots which is our next topic. I trust that you are enjoying the lesson.

Self-Assessment Questions

Exercise 3.1

1. The ordered array is used to present data which is….
   A. categorical
   B. nominal
   C. numerical
   D. qualitative

For each of the following items, indicate whether the item is **false** or **true** by circling the correct response.

2. Ordered array plots are used because they help visualise data.

A. False
B. True

3. Ordered array plots are used to generate a summarized view of a very large dataset which is helpful for gaining insight
   A. False
   B. True

4. In ordered array plots, the order in which data was originally collected is maintained
   A. False
   B. True

5. The table below provides information about the educational background of teachers in Daffodil International School.

Education	Number
High school or less	55
Diploma	32
Bachelor	23
Master	12

   i. Present the data in the table using ordered array and explain.
   ii. What is the teaching staff number of Daffodil International School?

SESSION 2 ORGANIZING NUMERICAL DATA: STEM AND LEAF PLOT

I trust that you had a good time with the ordered array and had success in drawing the ordered array tables. In this session, we shall study the stem and leaf plot. We shall learn how to read and construct the stem and leaf plot, identify the strengths and limitations as well as the uses of the stem and leaf plot.

Objectives

By the end of the session, you should be able to

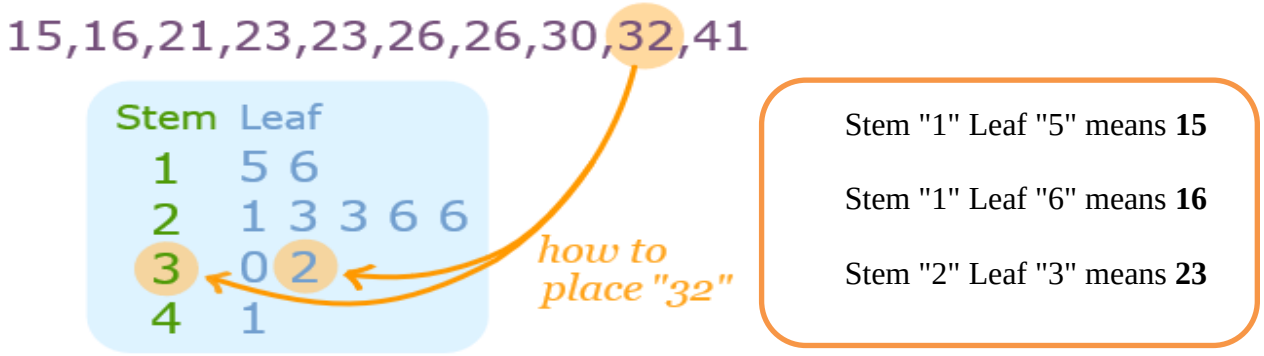
(a) read a stem and leaf plot,
(b) construct a stem and leaf plot,
(c) explain the strengths and limitations of stem and leaf plot,
(d) explain the uses of stem and leaf plot.

Now read on...

2.1 Nature of stem and leaf plot

The stem and leaf plot is an interesting way to showcase data. Stem and leaf plots use numerical data. Stem and leaf plots are a method for showing the frequency with which certain classes of values occur. It is represented in the form of a table with two columns; stem, and leaf. Each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit).
The stems are arranged vertically with their corresponding leaves ordered horizontally. Stem and leaf plots give a pictorial view of the distribution of a data set. An example is shown below.

The scores of ten students in a mathematics quiz are 15, 16, 21, 23, 23, 26, 26, 30, 32, 41.
We can illustrate the scores in a stem and leaf plot as below.



2.2 Reading Stem and Leaf plot

There is important information you can read from a stem and leaf plot. We are going to illustrate how to identify and read this information. To help us achieve this is the following example.
The ages of people at a school speech and prize giving day are represented on the following diagram.

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

Start with the key. It will guide you on how to read the other values.
The key on this plot shows that the stem is the tens place and the leaf is the ones place.

Now we can see that the oldest person at the speech and prize giving day is 66 years old.

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
(6)	3, (6)

Key: 6|3 = 63 years old

We can also see that the youngest person at the event was 01, or 1 year old.

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

Looking across the rows, we can see that there are **9 people in their 30s** and **4 people in their 40s.**

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

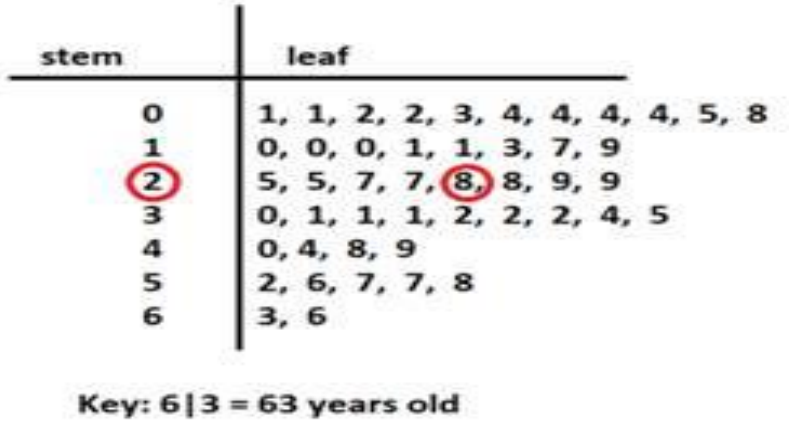
Key: 6|3 = 63 years old

With the numbers ordered on the leaf side of the plot, we can also see that there are 4 children that are **4 years old**. This represents the mode because it is the age that appears the most.

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

We can also easily get the median by finding the middle of the leaves. Here we can see that the median is 28 years old. So, half of the guests are younger than 28 and half are older than 28.



stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

2.3 Constructing stem and leaf plot

There are three simple steps to take in the construction of a stem and leaf plot. These three steps are described below. We would use the following data set to illustrate the steps.

Here is a set of data showing the science mock examination scores of students in Mr Abu's class.
56, 78, 82, 82, 90, 94, 93, 67, 67, 69, 74, 77, 92, 88, 81, 83, 84, 77, 72

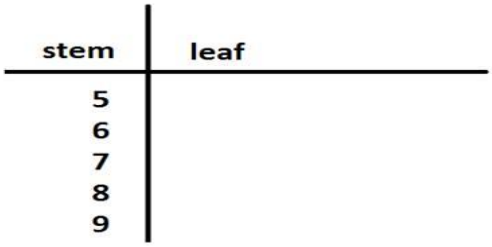
Step 1: Organize and order (lowest to highest) the data into groups.

In this situation, we will group the tests by decades.

56 67, 67, 69 72, 74, 77, 77, 78 81, 82, 82, 83, 84, 88 90, 92, 93, 94

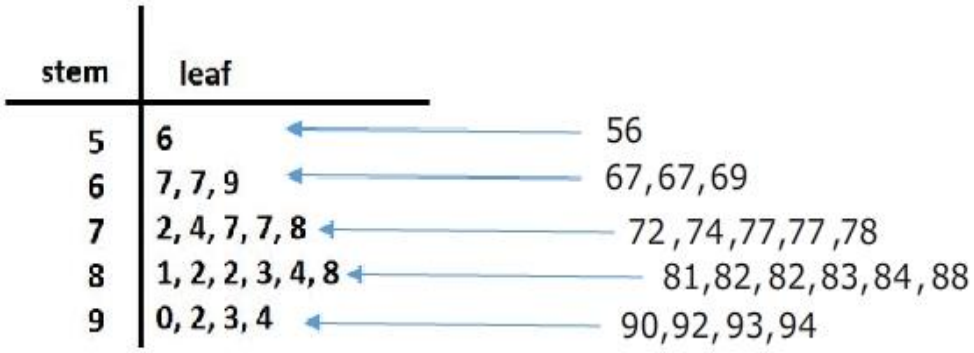
Step 2: Create the plot with the stems and the leaves identified.

For our example, we will create the plot with the stems as the tens and the leaves as the ones. The stems will be 5, 6, 7, 8 and 9

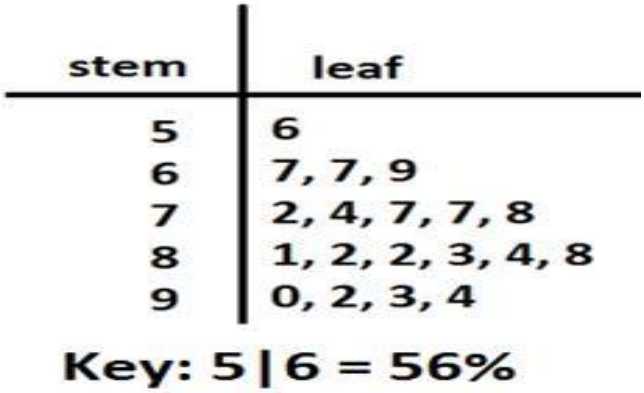


stem	leaf
5	
6	
7	
8	
9	

Now we are ready to add the ones place from each of the values in the list we made.



Step 3: Add a key to the bottom of the stem and leaf plot. This is to ensure the right interpretation of the plot.



I trust that you will be able to follow the steps above and construct stem and leaf plots.

Try the exercise below and bring your attempt to FTF for discussion.
Make a stem and leaf plot of the algebra test scores given below.
Then complete each question.
56, 65, 98, 82, 64, 71, 78, 77, 86, 95, 91, 59, 69, 70, 80, 92, 76, 82, 85, 91, 92, 99, 73

1. What type of graph does a stem and leaf plot represent when turned vertically?
2. What was the lowest score on the algebra test?
3. What was the highest score on the algebra test?
4. In which interval did most students score?

Now let us look at the strengths and limitations of stem and leaf plots.

2.3 Strengths and limitations of stem and leaf plot

Stem and leaf plots have a number of strengths and limitations. These are described below.
Strengths of the stem and leaf plot are:

1. A stem and leaf plot can be constructed quickly using pencil and paper.
2. In a stem and leaf plot, the original and specific data values of a data set can be seen and identified.
3. A stem and leaf plot allows you to clearly see the shape of the distribution of a data set.
4. In a stem and leaf plot, extreme values, data clusters and gaps are easily visible.
5. A stem and leaf plot can be used to conveniently determine the range, mode and median of a data set quickly.

Limitations of the stem and leaf plot are:

1. A stem and leaf plot is not very informative for a very small data set.
2. It is tiring in constructing stem and leaf plots for very large data sets.
3. The order in which data is originally collected is lost in a stem and leaf plot.

Now we have learnt about the nature of the stem and leaf plot, how to construct it and the strengths and limitations. Pause for a few minutes. What do you think would be the uses of stem and leaf plots in education?

Take a few minutes and write down, in your jotter, one use of stem and leaf plots in education.

Now compare what you have written with the uses below.

2.4 Uses of stem and leaf plot

Stem and leaf plots can be used by teachers and educational practitioners in the following ways.

1. Stem and leaf plots can be used by teachers to see the distribution of students' scores on a test.
2. In the district education offices, stem and leaf plots can be used to see the distribution of enrolment of students in the various schools in the district

In this session, you have learnt about stem and leaf plots. We have discussed the nature of stem and leaf plots, the steps in the construction of stem and leaf plots, the strengths and limitations of stem and leaf plots and the uses of stem and leaf plots. Do well to revise your notes and practise constructing stem and leaf plots as well as interpreting them. In the next session, we shall learn about box and whisker plots.

Self-Assessment Questions

Exercise 3.2

1. Stem and leaf plots are most useful for representing data which is…….in nature
   A. categorical
   B. nominal
   C. numerical
   D. qualitative

2. One strength of stem and leaf plots is that…
   A. the distribution of the data is visible.
   B. the order in which data was collected remains same.
   C. they are effective with all scales of measurement.
   D. they make no provision for scores that are extreme.

3. Stem and leaf plots **cannot** be used to compare…
   A. ages of heads of schools in a district.
   B. colour preference of students in a class.
   C. enrolment in the classes in a school.
   D. the skewness of two different data sets.

4. One weakness of stem and leaf plot is that …
   A. data clusters and gaps are not visible.
   B. it is easy to construct.
   C. it is very informative for a very small data set.
   D. the order in which data was collected is distorted.

For each of the following items, indicate whether the item is **false** or **true** by circling the correct response.

5. In constructing stem and leaf plot, the values must first be ordered.
   A. True
   B. False

6. A key is not important in constructing a stem and leaf plot.
   A. True
   B. False

SESSION 3: ORGANIZING NUMERICAL DATA: BOX AND WHISKER PLOT

You are welcome to Session 3 of this unit. I trust that you have understood the representations of data we have studied in this unit. Remember we have described ordered array and stem and leaf plots. I believe that your attempt to construct these plots have been successful. Today, we would study another way of presenting numeric data graphically.

Another way of presenting data graphically is by the use of box and whisker plot. In this session, we shall look at what the box and whisker plot is all about and how to construct it. Go gradually and re-read areas you do not understand well until you have grasped the content.

Objectives

By the end of the session, you should be able to

a) describe the nature of box and whisker plot,
b) construct a box and whisker plot.
c) explain the strengths and limitations of box and whisker plot and
d) state at least two uses of box and whisker plots in education.

Now read on...

3.1 The nature of box and whisker plot

A box and whisker plot is sometimes called a boxplot. It is a graph that presents information from a five-number summary. Box and whisker plots are a handy way to display data broken into four quarters (we will learn more about quartiles in subsequent units), each with an equal number of data values. The box and whisker plot does not show frequency, and it does not display each individual statistic, but it clearly shows where the middle of the data lies. It's a nice plot to use when analysing how your data is skewed. The box and central line are centered between the endpoints if data are symmetric around the median. Also, a box and whisker plot can be shown in either vertical or horizontal format.

In describing the box and whisker plot we said it presents a graphical display of the five-number summary. What is this five-number summary? You will be introduced to it soon.

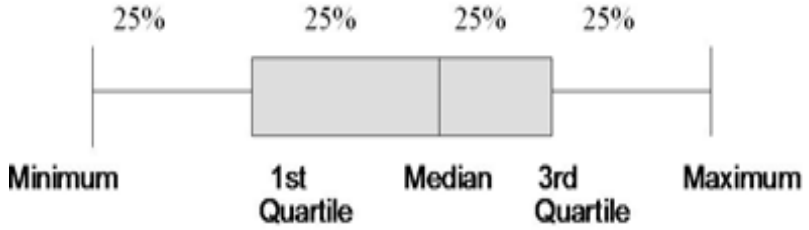
3.1.1 The five-number summary

The five-number summary is a measure that provides information about a set of observations. They describe the spread of a data. The five numbers are:

a. Minimum
b. First Quartile (Q1)
c. Median (Q2)
d. Third Quartile (Q3)
e. Maximum

Well, if this is the first time you are meeting such measures do not be overwhelmed. We have an example to clearly illustrate what each means. You can refer to Unit 4 for details on calculating median and quartiles. Before the example let us have a look at how a box and whisker plot looks like.

Example of a box and whisker plot



The minimum represents the lower extreme of the given data set and the maximum represents the upper extreme of the same data set. The two horizontal segments on each side of the box attached to the minimum and maximum values are known as the "whiskers". It is important to take note of the following in any box-and-whisker plot:

a) the left-side whisker represents where we find approximately the lowest 25% of the data
b) the right-side whisker represents where we find approximately the highest 25% of the data.
c) The box part represents approximately the middle 50% of all the data.
d) The data is divided into four regions, where each represents approximately 25% of the data.

Now that we know how a box and whisker plot looks like let us learn how to construct it.

3.2 Constructing a box and whisker plot

In constructing a box and whisker plot we would use the following example.

Suppose the scores of students on a geometry test marked out of 25 are as follows {3, 7, 8, 5, 12, 14, 21, 15, 18, 14}. Draw a box-and-whisker plot for the data.

From our example, we can find the five-number summary by following the steps:

Step 1: Order the data from least to greatest :3, 5, 7, 8, 12, 14,14,15, 18, 21

Step 2: Find the median of the data. This is also called quartile 2 (Q2) and it is 13.

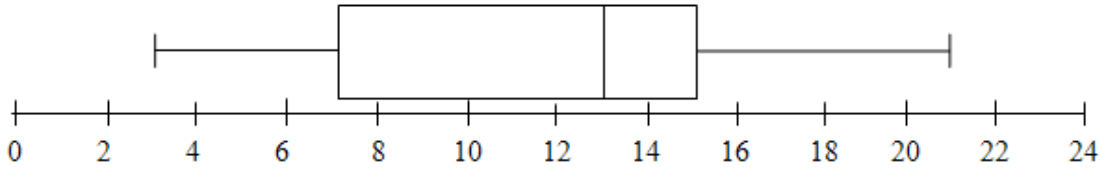
Step 3: Find the median of the data less than Q2. This is the lower quartile (Q1) and it is 7.

Step 4. Find the median of the data greater than Q2. This is the upper quartile (Q3) and it is 15.

Step 5. Find the extreme values which are the minimum and maximum data values.
Minimum value = 3 and maximum value = 21.

Step 6. Create a number line (use an equal number scale) that will contain all of the data values. It should stretch a little beyond each extreme value.

Step 7. Using the equal interval scale, draw a rectangular box with one end at Q1 and the other end at Q3. Then draw a vertical segment at the median value. Finally, draw two horizontal segments on each side of the box, one down to the minimum value and one up to the maximum value.



3.3 Strengths and limitations of box and whisker plot

i. A box and whisker plot can show whether a data set is symmetric or skewed.
ii. The shape of distribution of a data set can be seen on a box and whisker plot.
iii. Box and whisker plots allow for multiple sets of data to be displayed in a single graph.
iv. They allow for comparison of data from different categories.
v. A box and whisker plot shows the variability of a data set.
vi. A box and whisker plot does not show frequency.
vii. A box and whisker plot does not display the individual statistics.

Now we have learnt about the nature of the box and whisker plot, how to construct it and the strengths and limitations. Pause for a few minutes. What do you think would be the uses of box and whisker plots in education?

Take a few minutes and write down, in your jotter, two uses of box and whisker plots in education.

Now compare what you have written with the uses below.

3.4 Uses of box and whisker plots in education

Box and whisker plots can be used by teachers and educational practitioners in the following ways.

1. A teacher can use box and whisker plots to compare the performance of students in a particular subject from different classes.
2. The district education office can compare the performances of students of particular schools over the years.
3. A teacher can use box and whisker plots to analyse the effect of a methodology on students by comparing their scores before and after the intervention.

Take a 10-minute break and refresh yourself. Now read sessions 3.1 and 3.2 again, going over the steps and noting the examples.

Now, do the exercise below and bring to FTF for discussion.

Suppose that a Theatre Department in a University kept track of how many DVDs they rented each month for a two-year period. The numbers for each month are shown in the table below.

J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
3	5	2	8	1	5	0	3	6	4	9	15	3	6	4	1	10	3	8	7	2	9	0	11

1. Make a single box and whisker graph from this data.
2. Make separate box and whisker graphs for each year.
3. Compare the three graphs. Record your observations.

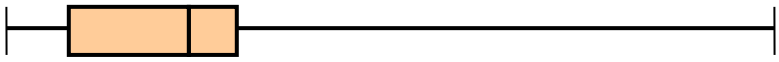
In this session, you have learnt about box and whisker plots. We have discussed the nature of box and whisker plots, the steps involved in its construction, its strengths and limitations and the uses of box and whisker plots in education. I trust that that you have enjoyed the lesson. Read over frequently and practise many other examples to enable you to apply it whenever it is applicable in your practice. In the next session, we shall learn about frequency distribution.

Self-Assessment Questions

Exercise 3.3

1. Refer to the box & whisker graph below that shows how much time was spent per night on homework for the first-year class at a certain high school during September.

Average minutes per night spent on homework



0 20 48 60 190

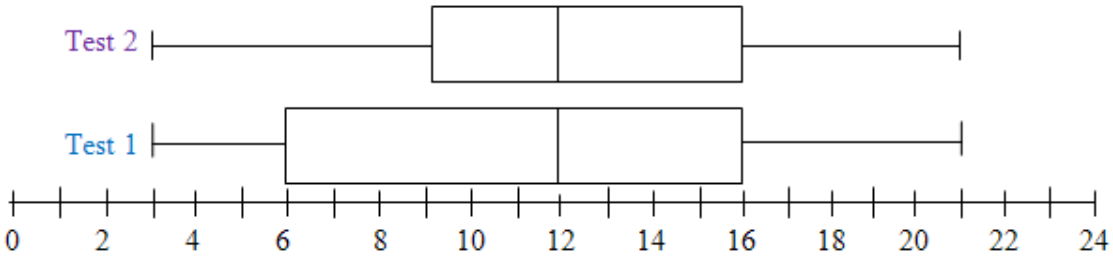
i. What percent of the first years spend more than 60 minutes on homework per night?

ii. What is the range of times that the middle 50% of the first years spend on homework per night?

iii. How many first year students do not do homework?

iv. What percent of the first year students spend less than 20 minutes per night on homework?

2. Suppose that the box-and-whisker plots below represent quiz scores out of 25 points for Test 1 and Test 2 for the same class.



i. What percentage of the students scored at or above the score of 6 for Test 1?
ii. What percentage of the students scored at or above the score of 9 or for the second test?
iii. What do these box-and-whisker plots show about how the class did on Test 2 compared to Test 1?

3. The five-number summary does **not** include……
   A. Mean
   B. Median
   C. Minimum
   D. Maximum

4. The second quartile is the same value as the median.
   A. False
   B. True

SESSION 4: ORGANIZING NUMERICAL DATA: FREQUENCY DISTRIBUTIONS

You are welcome Session 4 of this Unit. I trust that you have understood the pictorial representations of data we have studied in this Unit. Remember we have described ordered array, stem and leaf and box and whisker plots. I believe that your attempt to construct these graphs and charts have been successful.

Another way of representing data is by the use of frequency distribution tables. In this session, we shall look at the types of frequency tables and how to construct them. Go gradually and re-read areas you do not understand well until you have grasped the content. Do not be in a hurry to complete the session.

Objectives

By the end of the session, you should be able to

(a) describe the nature of frequency distributions,
(b) describe the features of frequency distribution tables,
(c) construct frequency distribution tables.

Now read on...

4.1 Nature of frequency distributions

Data normally comes in raw or ungrouped form as shown below for 40 students in a Statistics class.

The following marks were obtained by a group of 40 students in a Statistics examination.

76	88	93	75	70	93	73	62	69	75
71	80	52	76	66	54	73	80	79	89
83	62	53	79	69	56	81	75	71	72
52	65	49	80	67	59	88	87	91	82

The raw data alone does not give much information. In the example above, we can best know the highest score (93) and the lowest score (49). A lot more information can be obtained if the data are put in the form of a frequency distribution.

A frequency distribution is any arrangement of data that shows the frequency of occurrence of different values of the variable or the frequency of occurrence of values falling within arbitrarily defined ranges of the variable. A frequency distribution table is often used to show the frequency distribution which could either be ungrouped or grouped.

4.1.1 Ungrouped frequency distribution

In this type of frequency distribution, a table is drawn with two columns. In the first column, the raw scores are listed and in the second column, the number of times each score occurs is recorded.

The data on the marks obtained by the group of 40 students in a Statistics examination as given above, is presented below in Table 2.7.

Table 2.7 Ungrouped frequency distribution table

Score	Frequency
93	2
91	1
89	1
88	2
87	1
83	1
82	1
81	1
80	3
79	2
76	2
75	3
73	2
72	1
71	2
70	1
69	2
67	1
66	1
65	1
62	2
59	1
56	1
54	1
53	1
52	2
49	1
Total	40

Ungrouped frequency distributions have two major problems. Where the data set is large and the range (i.e. the difference between the highest and lowest score) is large, the table becomes too tall and occupies too much space. Secondly, several scores are often left out because they do not occur as part of the data. You will notice that scores such as 90, 85, 78, 60, 55 and others are not included. If all the scores from 93 to 49 are listed, the table will occupy a lot of space. Ungrouped frequency distributions are therefore not very useful in further statistical work.

4.1.2 Grouped frequency distributions

In this type of frequency distribution, the individual values are put into groups or classes. The values are most often put into groups or classes of 3, 5, 7, 9, and 10 as group sizes. The basic frequency distribution table has four columns. Column 1 has the classes, column 2, the class mark (the mid-points), column 3 the tallies, and column 4 the frequencies. These terms will soon become clear to you. Just relax. See an example of a grouped frequency distribution table of the Statistics scores listed above presented in Table 2.8.

Table 2.8 Grouped frequency distribution of Statistics students' performance

Class	Class Mark	Tally	Frequency
91-95	93	///	3
86-90	88	////	4
81-85	83	///	3
76-80	78	~~////~~ //	7
71-75	73	~~////~~ ///	8
66-70	68	~~/////~~	5
61-65	63	///	3
56-60	58	//	2
51-55	53	////	4
46-50	48	/	1
Total			40

4.2 Features of a grouped frequency distribution table

A grouped frequency distribution has a number of unique features. These are described below.

1. Class. This is the group of scores as shown in column 1 of Table 2.7.
2. Class interval. The range within which a group of scores lie. It has a number at the beginning and at the end. In Table 2.7 the first class from the top has the interval, 91-95. When all the class intervals have the same range (i.e. difference between the two values), the distribution is referred to as equal class interval distribution but where there are differences in the range of the intervals, the distribution is referred to as unequal class interval distribution.
3. Open-ended classes. These are classes with a value at one end, either at the beginning or the end and a description at the other end. These intervals are put either at the top or bottom of the table. Using the forty scores above a top class can be "90 and above" or "Above 90" and the bottom one can be "45 and below" or "Below 46".
4. Class limits. These are the end points of a class interval. The smaller number is the lower limit and the bigger number is the upper limit. In Table 2.7, using the bottom class of 46-50, the lower limit is 46 and the upper limit is 50.
5. Class mark: The midpoint for each class interval. They are obtained by adding the two class limits and dividing the result by 2. To get the class mark for the class, 86-90, 86 is added to 90 to obtain 186. 186 ÷ 2 gives 93.
6. Class boundaries. These are the exact or real limits of a class interval. The lower class boundaries are obtained by subtracting 0.5 from the lower class limit. The upper class boundaries are obtained by adding 0.5 to the upper class limits. A class interval with limits of 91 – 95 produces class boundaries of 90.5 - 95.5.

7. Class size/class width. These are the number of distinct or discrete scores within a class interval. They are obtained by finding the difference between successive lower class limits or upper class limits in cases of equal class intervals. They can also be obtained by finding the difference between successive class marks in cases of equal class intervals or between class boundaries for each interval. For an easier way, just count the number of scores within an interval. For example, 46-50 will give us, 46, 47, 48, 49, 50, giving us 5 numbers. The class size is then 5.
8. Frequency: This is the number of distinct scores from the given data that can be found in a class interval. They are obtained through tallying (i.e. using strokes, /// to represent the scores). To make counting easier, the strokes are often bound into bundles of 5.
9. Cumulative frequency. This is the successive sum of the frequencies starting from the frequency of the bottom class. The frequency for each class is added to the cumulative frequency below it and then recorded for the particular class. The top class has a cumulative frequency that equals the total frequency and the bottom class has a cumulative frequency that is the same as the frequency for the class. In Table 2.8, the cumulative frequency for the class, 51-55, is obtained by adding 4 (the frequency of the class) to 1 (the cumulative frequency below the class) to obtain 5. Likewise, for the class, 56-60, the cumulative frequency is obtained by adding 2 (the frequency of the class) to 5 (the cumulative frequency below the class) to obtain 7.
10. Cumulative percentage frequency. These are obtained by expressing each cumulative frequency as a percentage. The cumulative frequency of the class is divided by the total frequency and the result multiplied with 100. For example, the cumulative percentage frequency of the class 86-90 is obtained as follows: $\frac{37}{40} \times 100 = 92.5$
11. Relative frequency. This is obtained by dividing each frequency by the total frequency. In Table 2.8, the relative frequency of the class, 76-80 is obtained as follows: $\frac{7}{40} = 0.175$ The total relative frequency must always be made to add up to 1.0.
12. Cumulative relative frequency. This is the successive sum of the relative frequencies starting from the relative frequency of the bottom class. The relative frequency for each class is added to the cumulative relative frequency of the class below it and then recorded for the particular class. The top class has a cumulative relative frequency that equals 1 and the bottom class has a cumulative frequency that is the same as the relative frequency for the class. In Table 2.9, the cumulative relative frequency for the class, 71-75, is obtained by adding 0.200 (the relative frequency of the class) to 0.375 (the cumulative relative frequency below the class) to obtain 0.575. Likewise, for the class, 56-60, the cumulative relative frequency is obtained by adding 0.050 (the relative frequency of the class) to 0.125 (the cumulative relative frequency below the class) to obtain 0.175.

Table 2.9 An expanded frequency distribution table

Class	Class Mark	Tally	Frequency	Cumulative Frequency	Cumulative Percentage Frequency	Relative Frequency	Cumulative Relative Frequency
91-95	93	///	3	40	100	0.075	1.0
86-90	88	////	4	37	92.5	0.100	0.925
81-85	83	///	3	33	82.5	0.075	0.825
76-80	78	~~////~~ //	7	30	75.0	0.175	0.750
71-75	73	~~////~~ ///	8	23	57.5	0.200	0.575
66-70	68	~~////~~	5	15	37.5	0.125	0.375
61-65	63	///	3	10	25.0	0.075	0.250
56-60	58	//	2	7	17.5	0.050	0.175
51-55	53	////	4	5	12.5	0.100	0.125
46-50	48	/	1	1	2.5	0.025	0.025
Total			40			1.000	

4.3 Constructing a grouped frequency distribution table

There are nine steps in the construction of a grouped frequency distribution table. These steps are listed below.

Step 1
Draw a table with four columns with the headings – Class, Class mark, Tally, Frequency.

Step 2
Determine the range i.e., the difference between the highest score and the lowest score. For example, from the raw scores of the 40 students in the Statistics examination, the highest score is 93 and the lowest is 49. The range becomes 93 – 49 = 44.

Step 3
Decide on a class size. Popular sizes are 3, 5, 7, 10. Odd-numbered class sizes make computations easier. In Education, the most popular sizes are 5 and 10.

Step 4
Determine the approximate number of classes by dividing the range by the class size. For example, suppose a class size of 5 is taken. The approximate number of classes would be $\frac{44}{5} = 8.8$ which is rounded to 9 classes. Generally, number of classes is between 5 and 20.

Step 5
Identify the highest value or score and write it down.

Step 6
To obtain the topmost class, decide on whether to start with the lower or the upper limit of the class interval. Determine the closest numbers to the highest value identified in Step 5 that is a multiple of the class size. Choose one of the values as either the lower limit or upper limit and use the class size to determine the other limit. For example, if a class size is 5, and the highest score is 93, then the closest values are 90 and 95. A possible lower

limit is 90 which is less than 93 and a possible upper limit is 95 which is greater than 93. If the lower limit is selected, then count the numbers from the lower limit up to the class size. In this case, since the class size is five, counting from 90, gives, 90, 91, 92, 93, 94 (i.e. five scores) and the class interval becomes 90-94. If the upper limit is selected, then count the numbers from the upper limit down to the class size. In this case, since the class size is five, counting from 95, gives, 95, 94, 93, 92, 91 (i.e. five scores) and the class interval becomes 91-95.

Step 7
Complete the first column with the rest of the classes using equal class sizes by counting the number of scores that make the class size for each interval. Make sure that there are no overlaps and no gaps between successive classes. The lower limit of one class should be one discrete unit above the upper limit of the class below it. For example consider these two classes, 91-95 and 86-90 which is below it. The lower limit of 91-95 is 91 and the upper limit of 86-90 is 90. The difference between 91 and 90 is therefore 1 unit. When the first column is done complete the second column with the class marks by adding the two limits and dividing the result by 2.

Step 8
Tally the scores in the third column which is the tally column. Take the scores one by one and place slashes (/) or tally marks in the respective classes. Where the tallies are five, bind them into one unit to facilitate easy counting. See Table 2.8 above.

Step 9
Count the number of slash or tally marks for each class and write them in the frequency column. Add the frequencies and put the total at the bottom of the frequency column.

4.4 Points to note in constructing a frequency distribution table

In constructing frequency distribution tables, there are a number of important points to note. These points are described below.

1. In Education, the highest class intervals or classes are at the top so this convention must be followed in constructing the frequency distribution table.
2. Use mutually exclusive classes. Make sure that an observation falls into one and only class. Classes must not overlap at the class limits. For example, 70 – 80 and 80 – 90 contain overlapping class limits of 80.
3. There should be no class with a zero frequency. If this occurs, it is recommended that the class size is changed. Preferably increase the class size.
4. Open-ended classes should be avoided. These classes have only the lower limit if it is the class at the top, or the upper limit if it is the class at the bottom. For example, 51 and above, 20 and below.
5. Aim at classes with equal sizes or width. This facilitates easy interpretation of the information from the frequency distribution table.
6. The number of classes should not be too small (i.e. not less than 5) and not too large (i.e. not more than 20). Where the number of classes is less than 5, class size should be reduced but when the number of classes is more than 20, the class size should be increased.

Take a 10 minute break and refresh yourself. Now reread sessions 5.3 and 5.4, going over the steps and noting the examples.

Now, do the exercise below and bring to FTF for discussion.

Given the following scores of 50 students in a Statistics class, and using a class width of 5, construct a grouped frequency distribution table. Also obtain the cumulative percentage frequencies, and cumulative relative frequencies.

32	38	25	40	47	22	48	45	20	35
16	18	10	6	8	11	33	30	28	27
42	35	30	34	31	21	25	12	20	25
43	33	36	39	42	17	19	22	26	10
33	38	32	22	26	42	37	35	40	46

In this session, you have learnt about frequency distributions. You have noticed the two types of frequency distributions, the features of the grouped frequency distribution and how to construct grouped frequency distributions. To make frequency distributions more meaningful, important points must be noted and you have read these points. In the next session, we shall learn about how to use frequency distribution tables to draw histograms, frequency polygons and ogives.

Self-Assessment Questions

Exercise 3.4

1. The highest score in a distribution is 98. The class size is 10. What is the most convenient highest interval?
   A. 89 - 99
   B. 90 – 99
   C. 91 - 99
   D. 90 - 100

Study the frequency distributio  answer questions **2-5.**

Distribution of (scores for Level 100 students

Classes	Frequency
90 – 99	10
80 – 89	12
70 – 79	15
60 – 69	24
50 – 59	13
40 – 49	16
30 – 39	10
Total	100

2. The class mark for the class, 90 - 99 is
    A. 99.5
    B. 95.5
    C. 94.5
    D. 90.5

3. What is the relative frequency for the class, 80 - 89?
    A. 0.12
    B. 0.90
    C. 12.0
    D. 90.0

4. What is the cumulative frequency of the class, 60 – 69?
    A. 24
    B. 37
    C. 39
    D. 63

5. The upper-class boundary of the class, 50 – 59 is
    A. 60.5
    B. 59.5.
    C. 59.0.
    D. 50.0.

6. Which of the following statements does **not** describe the construction of a frequency distribution table in education?
    A. Classes of equal intervals are used.
    B. Open-ended classes are avoided.
    C. There can be a class with zero frequency.
    D. Number of classes is limited to 5 – 20.

SESSION 5: ORGANISING NUMERICAL DATA: HISTOGRAM, FREQUENCY POLYGON AND OGIVE

You are welcome to the last session of Unit 3 for the Educational Statistics course. I trust that you have fully grasped the previous session on frequency distributions where you learnt how to construct a frequency distribution table. The frequency distribution table alone does not provide a lot of information. It is used as a basis for obtaining further information about data. One way is to use it is to represent information from the table graphically. This graphic representation takes the form of histograms, frequency polygons and ogives. This session will take you through these graphic representations. You will learn how to construct them and how important they are in educational practice.

Objectives

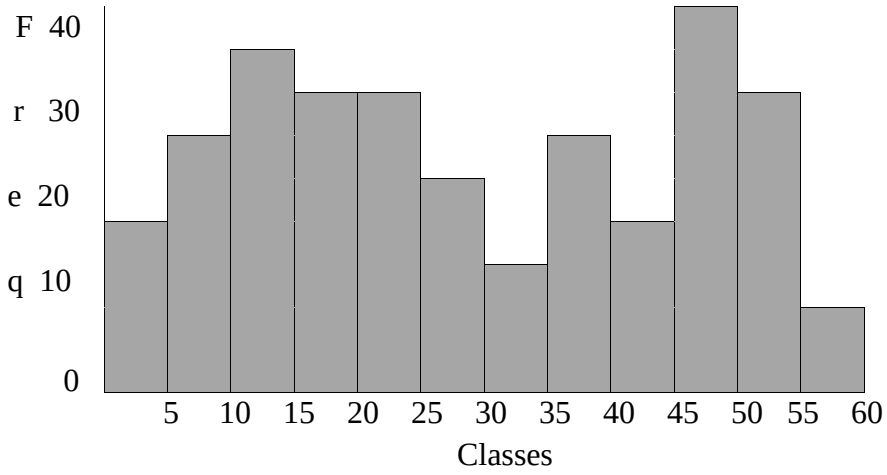
By the end of the session, you should be able to

(a) construct a histogram,
(b) explain the importance of histograms in educational practice,
(c) construct a frequency polygon,
(d) explain the importance of frequency polygons in educational practice,
(e) construct ogives.
(f) explain the importance of ogives in educational practice.

Now read on...

5.1 Histograms

You have learnt about bar graphs and can draw them easily. Well, histograms take the shape of bar graphs but with two differences. Histograms do not have spaces between the bars and use ratio and interval scales of measurement. It does not use nominal scale variables. Before a histogram is drawn, there should be a frequency distribution table. It uses the classes and the frequencies from the frequency distribution table. An example is shown below.



5.2 Constructing a histogram

The construction of a histogram involves four steps. These steps are described below. Where software such as Microsoft Excel or SPSS is available, they should be used, otherwise follow the following steps, using a graph sheet to draw the histogram.

Step 1

Draw two axes, a vertical and horizontal. Label the vertical axis by frequency and the horizontal axis by scores or classes.

Step 2

Select an appropriate scale on the vertical axis considering the highest or largest value as well as the lowest or smallest value. When using a graph sheet, the scale should be such that the bars are neither too tall nor too short.

Step 3

Use class midpoints/marks or class boundaries or class limits to label the points on the horizontal axis. It is always recommended that the label begins with the point 0. There are however situations where the lowest score is far from 0. In such circumstances, part of the horizontal axis is shrunk or moved towards the vertical axis to reduce extra unused space at the beginning of the graph.

Step 4

Draw bars of equal width representing the classes from a frequency distribution table with corresponding heights as the frequencies. There should be no spaces between the bars.

Now read over the steps again and draw a histogram for the frequency distribution table below. Discuss your histogram with your course mates during the face-to-face.

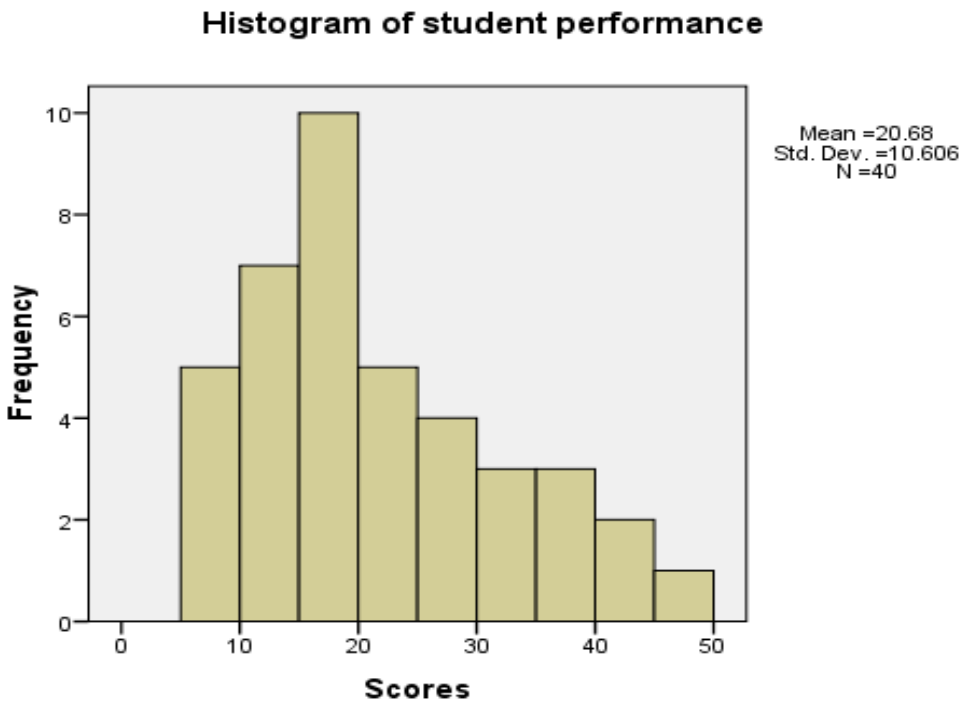
Distribution of examination scores for Level 100 students

Classes	Frequency
91 - 100	10
81 – 90	8
71 – 80	15
61 – 70	12
51 – 60	7
41 – 50	5
31 - 40	2
21 - 30	1
Total	60

5.3 Importance of histogram in educational practice

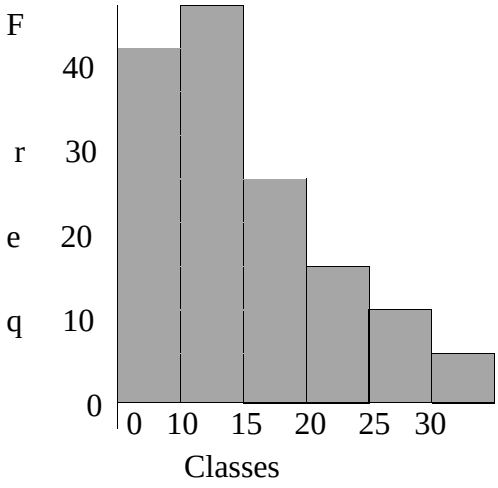
1. It gives a pictorial description of the raw data, providing information about the nature of the data. For example by observing the raw scores below it is difficult to get information about level of performance. However, if the data is presented in a histogram as shown, a better picture of the level of performance can be seen.

18	15	12	45	20	12	17	25	42	33
30	28	10	14	20	13	28	8	15	18
20	9	16	12	40	38	32	5	17	16
27	6	18	20	35	38	8	12	15	20

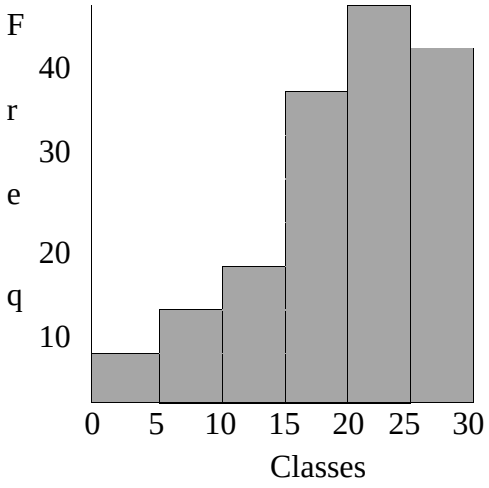


It can be observed that a lot of the students had scores between 5 and 20.

2. It gives the direction of performance in terms of academic performance (i.e. skewness).



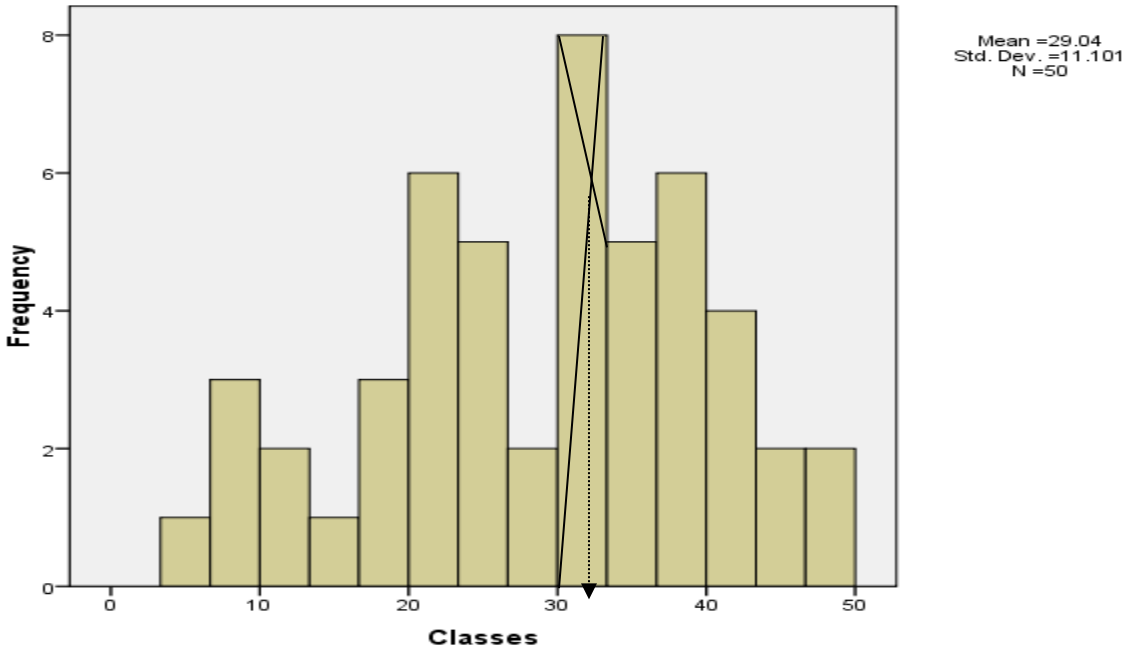
This histogram is skewed to the right implying that group performance tends to be low.



This histogram is skewed to the left implying that group performance tends to be high.

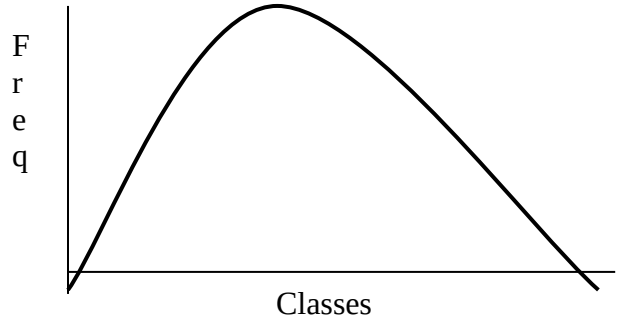
Note that the tail shows the direction of the skew. The histogram on the left is positively skewed and the one on the right is negatively skewed.

3. It provides an estimate of the most typical score. This is the intersection of the two diagonals of the tallest bar as shown in the histogram below. The most typical score (where the arrow is) can be estimated on the horizontal axis to be 32.



5.4 Frequency Polygons

A frequency polygon uses data from ratio or interval scales and depends on frequency distributions. It uses the classes and the frequencies from the frequency distribution table. An example is shown below.



5.5 Constructing a frequency polygon

The construction of a frequency polygon involves five steps. These steps are described below. Where software such as Microsoft Excel or SPSS is available, they should be used, otherwise follow the following steps, using a graph sheet to draw the frequency polygon.

Step 1

Draw two axes, a vertical and a horizontal one. Label the vertical axis by frequency and the horizontal axis scores or classes.

Step 2

Select an appropriate scale on the vertical axis considering the highest or largest value and the lowest or smallest value. When using a graph sheet, the scale should be such that the polygon is neither too pointed nor too short.

Step 3

Use class midpoints or class marks or class boundaries or class limits to label the points on the horizontal axis.

Step 4

Plot at the midpoint of each class the relevant heights as the frequencies. Join the midpoints with a straight line.

Step 5

Create two classes, one to the left and one to the right and plot on the horizontal axes the midpoints of the classes. Extend the line in Step 4 to join the horizontal axes to complete the polygon.

Now read over the steps again and draw a frequency polygon for the frequency distribution table below. Discuss your polygon with your course mates during the face-to-face.

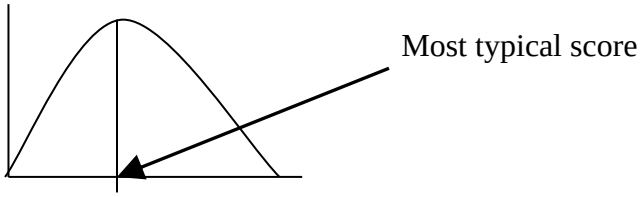
Distribution of examination scores for Level 100 students

Classes	Frequency
91 - 100	10
81 – 90	8
71 – 80	15
61 – 70	12
51 – 60	7
41 – 50	5
31 - 40	2
21 - 30	1
Total	60

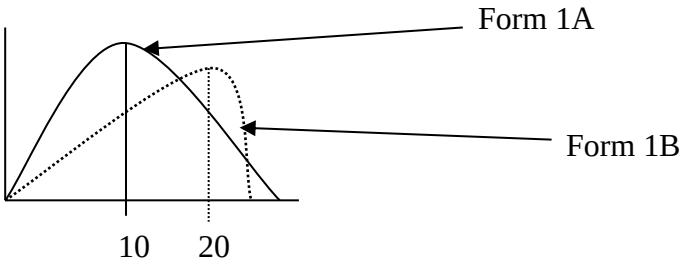
5.6 Importance of frequency polygons for educational practice

1. It gives a pictorial description of the raw data, providing information about the nature of the data. Raw data alone is difficult to study. If the raw data is transformed into a polygon, a visual impression is created and that makes information about the data easier to grasp.

2. It provides an estimate of the most typical score. This is the point on the horizontal axis where the highest point of the polygon is located. The most typical scores is used as a summary to represent the total group performance.

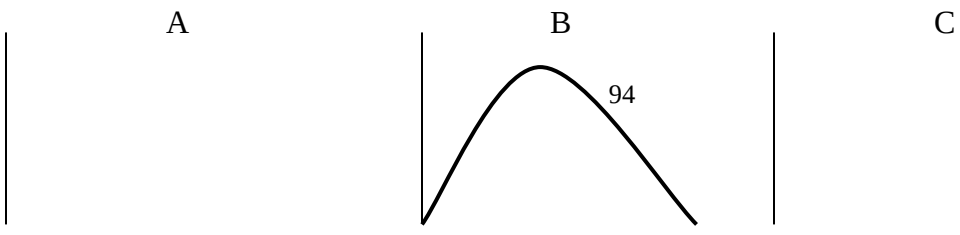


3. It is used to compare the performance of groups. For example the performance in a class test for Forms 1A and 1B can be shown as follows.



The diagram shows that Form 1B class, which is more to the right, performs better. The most typical scores, where the highest point of the polygon is located can be used to confirm the comparisons. In this case, Form 1B has 20 while Form 1A has 10. Where the total frequencies are not the same, use relative frequencies in place of the actual frequencies to draw the polygon.

4. It gives the direction of performance (i.e. skewness). Consider three classes, A, B, C.

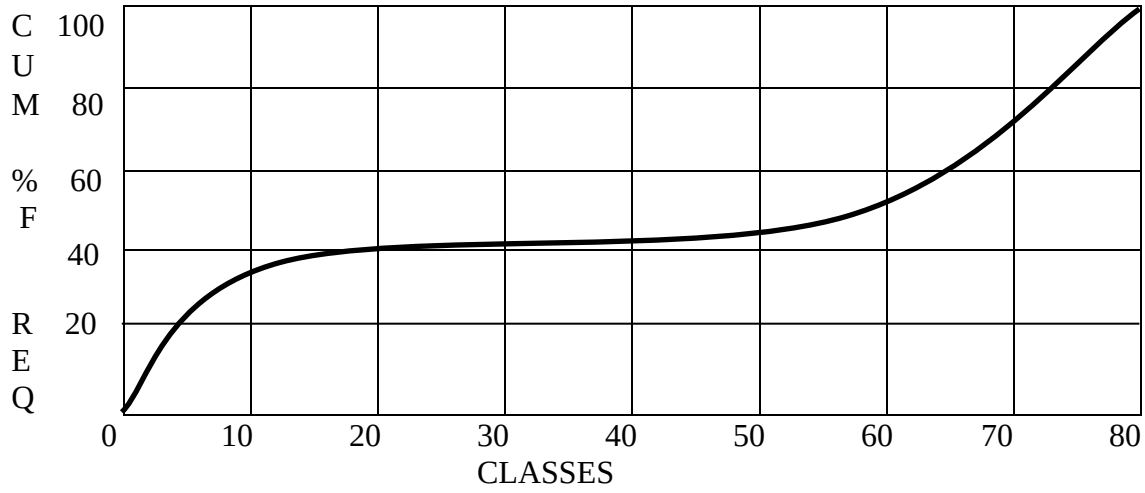


Positive skewness	Normal	Negative skewness
Skewed to the right		Skewed to the left
Tends to score low marks		Tends to score high marks

When the polygon is positively skewed, it implies that performance is low, but when it is negatively skewed, it implies that performance is high. However, when the polygon shows a normal curve, it means that the majority of the class is performing moderately (or on the "average").

5.7 Ogives

Ogives are also known as cumulative percentage frequency polygons. They are drawn from frequency distribution tables using percentage cumulative frequencies. Data from ratio or interval scales are most appropriate. There are two types of ogives. These are the "less than" and "more than" ogives. In educational practice, preference is given to the "less than" ogive and this is what we shall study in this session. They are obtained by plotting on a graph the upper class boundaries of each class against the cumulative percentage frequencies for each class. An example is shown below.



5.8 Constructing Ogives

The construction of an ogive involves six steps. These steps are described below. It is best to use graph sheets to draw the ogive.

Step 1

Draw two axes, a vertical and a horizontal one. Label the vertical axis by cumulative percentage frequency and the horizontal axis scores or classes.

Step 2
Put in the scale values from 0 to 100, preferably in steps of 10.

Step 3
Obtain cumulative percentage frequencies from a frequency distribution table.

Step 4
Put in the scale values on the horizontal axis using the upper class limits.

Step 5
Plot at the upper class boundary of each class the relevant values of the cumulative percentage frequency. Join the points with a straight line.

Step 6
Provide an extra class to the left and extend the polygon down so that it touches the horizontal axis.

Now read over the steps again and draw two ogives for the frequency distribution tables below on the same graph sheet. Discuss your polygons with your course mates during the face-to-face.

	School A		School B	
Classes	Frequency	Cum. % Freq	Frequency	Cum. % Freq
91 - 100	1	100	7	100
81 – 90	2	99	17	95.3
71 – 80	11	97	30	84
61 – 70	24	86	25	64
51 – 60	20	62	15	47.3
41 – 50	16	42	11	37.3
31 – 40	12	26	19	30
21 – 30	8	14	14	17.3
11 - 20	4	6	6	8
1 - 10	2	2	6	4
Total	100		150	

5.9 Importance of ogives in educational practice

Ogives have two main roles in educational practice.

1. It is used for comparisons of distributions of performance especially for distributions where the class totals are not the same. Generally, the graph that moves more to the right has better performance. The score obtained at the cumulative percentage frequency of 50 is often used to confirm the comparisons.

Pause for a few minutes. Look at the ogives you have drawn. Which school would you say performed better? Discuss your answer with your mates at the face to face.

2. Ogives are used to determine percentiles and percentile ranks. Percentiles and percentile ranks are measures of relative position and they describe an individual's position in relation to a known group. Percentiles are points in a distribution below which a given percent, P, of the cases lie and they divide a distribution into 100 equal parts. Percentile ranks are the percentage of cases falling below a given point on the measurement scale. It is the position on a scale of 100 to which an individual score lies.

   You may not understand what you are reading now. Relax. Do not worry at all. At Unit 6, you will learn a lot more about percentiles and percentile ranks. For now, only remember that ogives are used to determine percentiles and percentile ranks.

In this session, you have learnt about graphic presentation of data from frequency distributions. You have learnt how to draw histograms, frequency polygons and ogives. In addition, you have learnt the importance of these graphs in educational practice.

Self-Assessment Questions

Exercise 3.5

1. Histograms are most useful for representing data when the scale of measurement is

   I. interval
   II. nominal
   III. ordinal
   IV. ratio

   A. I only.
   B. IV only.
   C. I and IV.

D. I, III, IV.

2. Which one of the following variables would use a frequency polygon to represent the data.
    A. Age of students in a statistics class.
    B. Colour of dresses students wear to class.
    C. Number of photocopy machines on UCC campus.
    D. Percentage of girls in Asem Primary school.

3. One use of the histogram in the classroom is to determine
    A. the direction of student performance in a quiz.
    B. the gender distribution of students in a class.
    C. which courses students find difficult to understand.
    D. which individual students performed poorly.

4. When a frequency polygon for a quiz is positively skewed, it implies that performance is
    A. average
    B. high
    C. low
    D. normal

5. Ogives are used by teachers to
    A. compare performance of classes in the same subject.
    B. describe the performance of pupils in a test.
    C.. determine an appropriate teaching method.
    D. determine an individual's position in a test.

6. In constructing frequency polygons, the points that are marked and joined with a straight line are
    A. class marks.
    B. lower class limits.
    C. upper class boundaries.
    D. upper class limits.

This is a blank sheet for your short notes on:

- Difficult topics if any
- Issues that are not clear.

UNIT 4: MEASURES OF CENTRAL TENDENCY (LOCATION)

Unit Outline

Session 1: Purposes of the measures
Session 2: Summation
Session 3: The arithmetic mean
Session 4: The median
Session 5: The mode
Session 6: Quartiles

Congratulations! You have completed Unit 3. You are now welcome to Unit 4. So far you have studied basic statistical concepts and how to represent data graphically so that more information is obtained for decision making. In this Unit, you will learn about measures of central tendency or location. The first session discusses the purposes of these measures in education. The concept of summation is then learnt. This concept aids in understanding statistical formulae that are used in analyzing data and obtaining results. The three main measures, (i.e. the mean, median and the mode) form the crux of the Unit. You will learn how these measures are calculated, the features and how the classroom teacher can use them to enhance teaching and learning. The unit ends with a session on quartiles.

Unit Objectives

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

1. State and explain three purposes of the measures of central tendency (location);
2. Obtain the arithmetic mean from both ungrouped and grouped data;
3. Describe the properties of the mean and the uses of the mean in teaching and learning;
4. Obtain the median from both ungrouped and grouped data;
5. Describe the features of the median and the uses in teaching and learning;
6. Obtain the mode from both ungrouped and grouped data ;
7. Describe the features of the mode and the uses in teaching and learning;
8. Compute quartiles from both ungrouped and grouped data.

SESSION 1: PURPOSES OF THE MEASURES OF CENTRAL TENDENCY (LOCATION)

You are welcome to the first session of Unit 4 for the Educational Statistics course. As noted in Units 2 and 3, representing data by graphics, pictures or tables makes it easier to grasp the information the data contains. It also allows more information to be derived from the data. However, graphics and tables alone do not provide all the information a body of data contains. Measures of central tendency (location) help to reveal additional information that pictorial presentations cannot provide. In this session we shall first of all look at the general purposes of the measures.

Objectives

By the end of the session, you should be able to

(a) explain how measures of location are used to describe data,
(b) explain how the measures help to determine the levels of performance,
(c) explain how the measures give direction of student performance.

Now read on...

1.1 Nature of the Measures

The measures of central tendency are also known as measures of location. They are often referred to as averages. They are numbers that tend to cluster around the "middle" of a set of values. They provide single values which are used to summarise a set of observations or data. The three main measures that are mainly used in educational practice are the arithmetic mean, the median and the mode.

1.2 Purposes of the Measures

The measures of central tendency or location serve three main purposes. These purposes are described below.

1.2.1 Purpose One

Measures of location are used as single scores to describe data. They are typical scores that are used to represent a set of data. Using a typical value to represent a set of scores is a daily occurrence in different forms. For example, in an inter-class spelling competition, one student may be chosen to represent a class. Several factors may be considered before choosing the particular student. Once selected, the student becomes a representative (or a typical value) of the class. In a similar sense, a single score may be needed to represent a set of scores. One of the ways of selecting a single (typical) score is to use the measures of central tendency.

Suppose you have 200 students in your class. If you give them a quiz and you mark, you will have 200 scores before you. What meaning can you give to the performance of the class? Has the class performed well? Has the performance of the class been poor? To answer these questions, it will

not be wise to start calling out the names of the individual students and their scores. It will be a tiring and fruitless exercise. The best thing to do is to compute a typical score. This typical score would either be the mean, median or the mode.

Consider the following scores obtained by students in an examination.

52	69	58	42	30	65	85	50	72	59	48	25
74	52	68	75	40	59	59	65	70	65	40	66
72	70	85	90	56	60	45	40	72	80	50	55
70	45	48	60	65	58	52	65	60	64	75	80
70	45	40	35	40	58	55	60	54	40	42	48
70	40	50	58	50	80	42	45	60	20	25	40
60	45	80	45	20	60	40	48	52	45	60	40
45	38	30	50	59	75	45	50	40	60	68	60
56	58	50	42	65	40	80	60	45	20	50	40

Pause for a few minutes. What do you think would be a typical score to represent the scores? Write it down in the box below.

For the set of scores above, a typical score is 55. This is provided by both the mean and the median. How close were you to this score? What basis did you use to obtain your typical score? We shall soon learn how to compute the mean and the median as well as the mode.

1.2.2 Purpose Two

They help to know the level of performance by comparing with a given standard of performance. Very often teachers are asked about the general performance of their students. The answers are often like "The performance this year is very poor", "This year the students did not do well at all", "Oh, I tell you, my students did extremely well this year". These responses are based on subjective comparisons. Phrases such as, very poor, did not do well, extremely well, do not have any scientific basis. One teacher's perception of "poorness" may be different from another.

To solve the problem of subjectivity and ambiguity, measures of central tendency are obtained for a group and these measures are compared with a known standard. Therefore instead of saying the performance is poor, a teacher can say, the performance is above average, or average, or below average, where average would be the known standard or criterion.

In a school where the grading system is A, B, C, D and E, the average performance could be C, and the midpoint of the C range can be taken to be the standard or criterion. For example, if the C range is from 60 – 70, then a possible standard would be 65 [i.e. 60 + 70)/2]. In some situations, there is a pass or fail category, based on a pass mark. Suppose the pass mark is 55, those who score 55 and above have passed and those below 55, have failed. The standard or criterion is therefore 55.

For individual cases, a measure of central tendency or location can be taken as the standard or criterion for comparison. Instead of an individual responding to a question about his/her performance as poor, very good, excellent, it is better to say performance is above average, far above average, below average or far below average. In this instance, there is no subjectivity. Performances are being compared to actual values that are taken as an average.

Consider the following set of scores for 40 pupils in a Social Studies class.

68	42	58	45	60	72	80	50	70	90
75	80	45	60	72	85	60	75	58	62
48	65	60	65	55	48	74	68	66	59
36	90	54	58	62	68	44	90	65	78

The mean for these scores is 64.0 and the median is 63.50. If it is assumed that the average or standard or criterion performance is 60, then one can say that the performance of this class is above average since 64 and 63.5 are above the standard of 60. In sessions 3, 4, and 5, you will learn when to use the mean, median or mode as the 'average'.

For the individual number 1 who obtained 68, we can say that the performance is above average if we use either the mean or the median. For individual number 2, who obtained 42, we could say that the performance is far below average.

1.2.3 Purpose Three

They give the direction of student performance. In Unit 2, Session 6 you learnt that when a frequency polygon is positively skewed, it implies that performance is low, but when it is negatively skewed, it implies that performance is high. However, when the polygon shows a normal curve, it means that the majority of the class is performing moderately (or on the "average"). Do you remember this? If not, then go to Session 5.6 of Unit 3 and read over.
The information that is provided by the frequency polygon is also provided by the measures of central tendency or location. Instead of drawing a frequency polygon, you can compute the values of the mean, median and the mode and make comparisons.

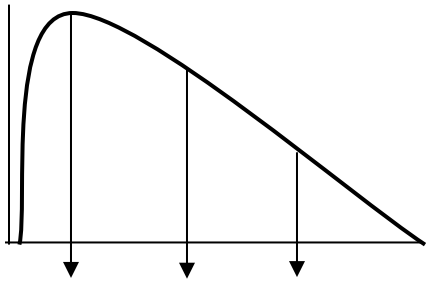
Let us see how this works.

1. Where Mean >Median, or Mean >Mode or Median > Mode, the distribution is skewed to the right (positive skewness) showing that performance tends to be low.

For example, in a class test, the following values may be obtained for the measures of central tendency,

Mean = 50 Median = 45 Mode = 38

You can observe that the Mean is greater than the Median and is also greater than the Mode. Also the Median is greater than the Mode. This information implies that the performance tended to be low in this class test. The frequency polygon below illustrates the point.

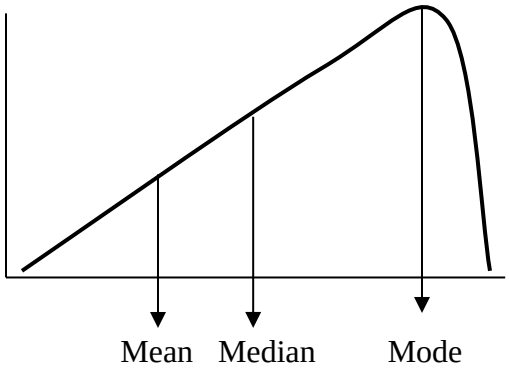


2. Where Mean < Median, or Mean < Mode or Median is < Mode, the distribution is skewed to the left (negative skewness) showing that performance tends to be high.

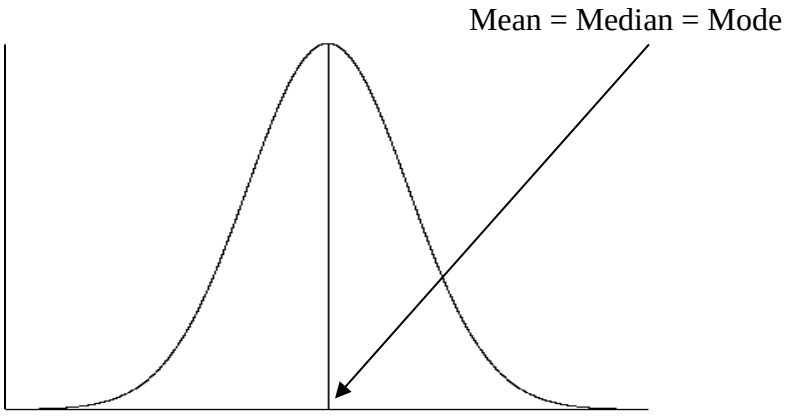
For example, in a class test, the following values may be obtained for the measures of central tendency:

Mean = 55 Median = 60 Mode = 75

You can observe that the Mean is less than the Median and is also less than the Mode. Also the Median is less than the Mode. This information implies that the performance tended to be high in this class test. The frequency polygon below illustrates the point.



However, if the mean, median and mode have the same value, then the distribution of the values is normal. This is illustrated below.



In this session, you have learnt about what the measures of central tendency are and the three purposes that they serve. You have noted that the three main measures of central tendency or location are the mean, median and mode. These three measures are used as typical scores to represent a set of data and they also help to show the level of performance when compared with a given standard. In addition, they give the direction of student performance. I believe that you have grasped this introductory concept.

Self-Assessment Questions

Exercise 4.1

1. In a class quiz, a mean of 48 was obtained with a median of 62. How would the performance of the class be described?
    A. Average
    B. Below average
    C. High
    D. Low

2. Measures of location can be used to determine the direction of student performance.
    A. False
    B. True

3. In a class test, the mean was 55 and the mode was 68. Performance is therefore high.
    A. False
    B. True

4. The median in an entrance examination was 62 with a mode of 54. The performance of the group was low.
    A. True
    B. False

5. When the mean is equal to the median, performance is skewed to the right.
    A. False
    B. True

6. When a distribution is negatively skewed, the mode is greater than the mean.
    A. True
    B. False

SESSION 2: SUMMATION

You are welcome to the second session of Unit 4 for the Educational Statistics course. I trust that you had a good understanding of the purposes that measures of central tendency serve to improve teaching and learning. In this session, we shall review some basic algebra. This review will help you to understand the formulae that you will learn to compute further statistics in subsequent lessons. Algebra is a very interesting subject and I believe you will enjoy the session.

Objectives

By the end of the session, you should be able to

(a) explain the concept, summation, with example,
(b) solve algebra problems involving addition using summation,
(c) solve algebra problems involving subtraction using summation,
(d) solve algebra problems involving multiplication using summation,
(e) solve algebra problems involving division using summation.

Now read on...

2.1 Summation

Generally, the letter, X, is used with a subscript to differentiate numbers in algebra as follows.

15	12	10	10	9	20	14	11	13	16
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}

Very often we wish to make arithmetical operations such as addition, subtraction, multiplication and division with the numbers. To make it easier for us to make these operations, the algebra notations are put into formulae. The first operation we will look at is summation. Summation is the addition of numbers.

For example, you may want to add the first six numbers, 15 12 10 10 9 20
You may also want to add the numbers starting from 9, i.e. 9 20 14 11 13 16. Or you may be interested in adding up all the numbers.

Well, the Greeks have introduced a sign which is used to represent summation. This sign is below.

$$\sum$$

When you see this sign, it means that add all the numbers that come after it. But as indicated earlier, actual figures are not used but letters, especially the letter X which represents the numbers. So we write, $\sum X$. This means that you add all the numbers represented by the letter X. Note that to differentiate between the numbers, we use subscripts like X_1 X_2 X_3. However, to use the letters, we cannot just list them $\sum X_1$ X_2 X_3. First we use the letter, i, to represent the subscripts. For the three X subscripts, we write i = 1, i = 2, i = 3. To add the first three numbers, we use the symbols:

$$\sum_{i=1}^{3} X_i$$

To add the 2nd to the 8th numbers, we use the symbols:

$$\sum_{i=2}^{8} X_i$$

When you want to add all the ten numbers, then you use:

$$\sum_{i=1}^{10} X_i$$

Note also that where all the numbers are used, the subscript can be ignored.

$$\sum_{i=1}^{10} X_i \text{ can be written as } \sum X$$

In general, where the total number of figures is not yet determined, we use

$$\sum_{i=1}^{n} X_i$$

where 1 is the X_1 number and n is the subscript for the last number, i.e. X_n

Pause for a minute. Read over the material. Does it become clearer after the second reading? I hope you have grasped the concept of summation.

Now, using the numbers above, attempt to write the following in summation notation.

Add: 14 11 13 16

Are you done? Ok. The answer is

$$\sum_{i=7}^{10} X_i$$

I hope you got it right. We added the numbers from the 7th to the 10th . Let us try another one. Write in summation form the sum of the numbers:

10 9 20 14 11 13

Have you finished? Ok.

The answer is

$$\sum_{i=4}^{9} X_i$$

I trust that you got it right. We added the numbers from the 4th to the 9th.
Good. I believe you have clearly understood this. Now let us go on to do the actual addition.

2.2 Addition, using the summation notation

We have learnt above that the summation notation can be used to write the addition of numbers in a summary form. For example, to add the numbers 2nd to the 8th numbers, we write:

$$\sum_{i=2}^{8} X_i$$

Now, let us do the actual addition.
Given:

15 12 10 10 9 20 14 11 13 16

What is the value of:

$$\sum_{i=1}^{4} X_i \ ?$$

This means that we add the first four numbers, 15 12 10 10 . This gives us the value 47. We therefore write,

$$\sum_{i=1}^{4} X_i = 47$$

This is an easy and simple concept. I do hope you have understood it.

Now I want you to try your hand at the following, using the space available.

1. $\sum_{i=2}^{6} X_i$
2. $\sum_{i=3}^{9} X_i$
3. $\sum_{i=1}^{7} X_i$

4. $\sum_{i=5}^{10} X_i$

5. $\sum_{i=4}^{8} X_i$

Now, check the answers below.

1. 61 2. 87 3. 90 4. 83 5. 64

2.3. Subtraction, using the summation notation

We will now look at subtraction, where we use the summation notation. This is an extension of the addition concept. You need to add the numbers first before the subtraction. Make sure that you do not make mistakes during the addition. If the addition is wrong, the subtraction will also be wrong. Note that the concept does not apply to individual numbers but the sum of the numbers.

Let us look at a few examples, based on the numbers below.

115	122	80	70	99	120	140	121	136	162
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}

1. $\sum_{i=2}^{5} X_i - \sum_{i=1}^{3} X_i = 491 - 317 = 174$

2. $\sum_{i=7}^{9} X_i - \sum_{i=3}^{5} X_i = 397 - 249 = 148$

Pause for a few minutes. Have you followed the operations? I trust that you have clearly understood it. However, if you have not understood it yet, then go over it again.

Now I want you to try your hand at the following, using the space available.

1. $\sum_{i=7}^{9} X_i - \sum_{i=3}^{5} X_i$

2. $\sum_{i=8}^{10} X_i - \sum_{i=3}^{5} X_i$

3. $\sum_{i=1}^{4} X_i - \sum_{i=9}^{10} X_i$

4. $\sum_{i=6}^{10} X_i - \sum_{i=2}^{5} X_i$

Now check the answers with what you got.

1. 148 2. 170 3. 89 4. 308

2.4. Multiplication and Division, using the summation notation

As noted with addition and subtraction, multiplication and division operations can also be done using the summation notation.

Let us look at a few examples based on the numbers below.

115 122 80 70 99 120 140 121 136 162

1. $\sum_{i=7}^{9} X_i \times \sum_{i=3}^{4} X_i = 397 \times 150 = 59550$

2. $\sum_{i=2}^{6} X_i \div \sum_{i=2}^{3} X_i = 491 \div 202 = 2.43$

3. $\sum_{i=8}^{10} X_i \times 10 = 419 \times 10 = 4190$

4. $\sum_{i=7}^{10} X_i \div 34 = 559 \div 34 = 16.4$

Pause for a few minutes. Have you followed the operations? I trust that you have clearly understood it. However, if you have not understood it yet, then go over it again.

Now I want you to try your hand at the following, using the space available.

1. $\sum_{i=6}^{10} X_i \times \sum_{i=3}^{4} X_i$

2. $\sum_{i=1}^{10} X_i \div \sum_{i=6}^{7} X_i$

3. $\sum_{i=1}^{4} X_i \times 15$

4. $\sum_{i=5}^{10} X_i \div 18$

Now check the answers with what you got.

1. 101,850 2. 4.48 3. 5,805 4. 43.22

In this session, you have learnt about the summation sign and the operations in arithmetic based on summation. The operations include addition, subtraction, multiplication and division. The skills you have acquired here will be used in the sessions on measures of central tendency and measures of variability. I trust that you have grasped the computations well.

Self-Assessment Questions

Exercise 4.2

The following scores were obtained by students in an examination. Use the scores to compute the required values.

45, 82, 75, 87, 60, 48, 90, 72, 65, 80, 65, 49, 52, 56, 68, 72, 64, 80, 70, 58,

1. $\sum_{i=4}^{20} X_i$

2. $\sum_{i=4}^{9} X_i + \sum_{i=12}^{18} X_i$

3. $\sum_{i=1}^{15} X_i - \sum_{i=16}^{20} X_i$

4. $\sum_{i=2}^{8} X_i \times \sum_{i=17}^{20} X_i$

5. $\sum_{i=1}^{20} X_i \div 20$

6. $\sum_{i=8}^{17} X_i \div \sum_{i=1}^{3} X_i$

SESSION 3: THE ARITHMETIC MEAN

You are welcome to the third session of Unit 4 for the Educational Statistics course. You remember that in Session 1, we discussed the three uses of the measures of central tendency or location. Remember that measures of central tendency tell us where the centre of the distribution is located and it is the point at which scores of observations cluster. In this session, we shall take one of the measures, the mean and consider how helpful it is to improve teaching and learning.

Objectives

By the end of the session, you should be able to:

(a) compute the arithmetic mean from a set of scores,
(b) describe the properties of the arithmetic mean,
(c) explain the strengths of the mean,
(d) explain the weaknesses of the mean,
(e) explain the uses of the mean in teaching and learning.

Now read on...

3.1 Computing the mean

In Statistics, there are three types of the mean. These are arithmetic, geometric and harmonic means. In Education however, the arithmetic mean is the most useful. In this session, we shall adopt the term, Mean, to represent the Arithmetic mean. The Arithmetic Mean (or the Mean) is the sum of the observations in a set of data divided by the total number of observations.

The arithmetic mean is often represented by the symbol, $\overline{X}$ pronounced X bar. The mean can be computed from raw data (ungrouped data) and grouped data. It can also be easily obtained from Microsoft Excel, SPSS and other statistical software.

3.1.1 Computing from raw data (ungrouped data)

Given the following scores, 15, 12, 10, 10, 9, 20, 14, 11, 13, 16, to obtain the mean, all the scores are added and divided by the total number of observations.

The mean for the scores above is:

$$\overline{X} = \frac{15+12+10+10+9+20+14+11+13+16}{10} = \frac{130}{10} = 13$$

The above expression can be written in the algebraic form as learnt in Session 2 as:

$$\overline{X} = \frac{\sum_{i=1}^{10} X_i}{10} = \frac{130}{10} = 13$$

In general, the equation will be written as, $\overline{X} = \frac{\sum X}{N}$ where N is the total number of observations.

Computing the mean from ungrouped data is simple. But you need to be careful with your additions because if a mistake is made, the final answer will be wrong.

Now compute the mean for the following set of scores:

45, 82, 75, 87, 60, 48, 90, 72, 65, 80, 65, 49, 52, 56, 68, 72, 64, 80, 70, 58

Compare your answer with the following result. $\overline{X} = 66.9$

Have you got it right? Good. If you got it wrong, check your addition and division.

Now let us compute the mean from grouped data.

3.1.2 Computing from grouped data

Two methods are generally used. These are the conventional or long method and the coding method. The methods are used with frequency distributions tables.

The long method uses the following formula: $\overline{X} = \frac{\sum fx}{n}$ OR $\overline{X} = \frac{\sum fx}{N}$ where f is the frequency and x, the class marks or class midpoints. Note that n and N refer to the number of observations and can also be written here as $\sum f$, which is the total frequency from the frequency distribution table.

The following steps are used, when given a frequency distribution table.

Step 1. Obtain the class marks or class midpoints.

Step 2. Multiply the class marks with frequencies and complete a new column, fx

Step 3. Add the values in the fx column

Step 4. Divide the result in Step 3 with total frequency to obtain the mean.

Now follow the example in Table 3.1.

Table 3.1 Computing the mean using conventional method

Scores	Midpoint x	Freq f	fx
46 – 50	48	4	192
41 – 45	43	6	258
36 – 40	38	10	380
31 – 35	33	12	396
26 – 30	28	8	224

21 – 25	23	7	161
16 – 20	18	3	54
Total		50	1665

Applying the formula gives us: $\overline{X} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{n} = \frac{1665}{50} = 33.3$

I trust that you have followed the steps and understood it well. Now let us try the coding method.

In the coding method, the following steps are used. Before you use the coding method, make sure that all class intervals are of equal size.

Step 1. Obtain the class midpoints or class marks.

Step 2. Create a new column after the frequency column and give it a heading, d.

Step 3. Choose the class that is in the middle of the distribution, but if there is no class exactly in the middle choose one of the two middle classes (preferably the class with the higher frequency). Under the column, d, code this class with '0'(zero).

Step 4. Give a code of 1 to the class immediately above the class coded 0. The next higher class is given a code of 2, the next higher one, a code of 3. Continue until you reach the topmost class.

Step 5. Give a code of −1 to the class immediately below the class coded 0. The next lower class is given a code of −2, the next lower one, a code of −3. Continue until you reach the bottom class.

Step 6. Create another column, fd, where you put in the values of the product of the frequencies and the codes.

Step 7. Add the values in the fd column.

Step 8. Divide the result in Step 7 with the total frequency and multiply the result with the class size, i.

Step 9. Add the result in Step 8 to the midpoint of the class coded 0 and obtain the final answer. This midpoint is called the assumed mean (AM).

The nine steps above are summarized in the formula for the coding method as:

$\overline{X} = AM + \left[\frac{\sum fd}{\sum f}\right]i$, or $\overline{X} = AM + \left[\frac{\sum fd}{N}\right]i$ where AM, is the assumed mean, f, is the frequency, d is the code for each class, $\sum f$ is the total frequency or N and i, the class size.

Now follow the example in Table 3.2.

Table 3.2 Computing the mean using the coding method

Scores	Midpoint	Freq	code	
	X	f	d	fd
46 – 50	48	4	3	12
41 – 45	43	6	2	12
36 – 40	38	10	1	10
31 – 35	33	12	0	0
26 – 30	28	8	-1	-8
21 – 25	23	7	-2	-14
16 – 20	18	3	-3	-9
Total		50		3

Applying the formula gives us: $\overline{X} = AM + \left[\frac{\sum fd}{\sum f}\right] i = 33 + \left[\frac{3}{50}\right] 5 = 33 + \frac{15}{50} = 33.3$

You will notice that both methods give the same result. The coding method is more appropriate where the frequencies are large in value. It is also easier to use when the midpoints have fractions such that multiplying them with the frequencies produces large values.

I believe you have understood the procedure for obtaining the mean through the coding method. I want you to try both methods to obtain the mean for the frequency distribution below.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

The answer is 43.1. I hope you have got both answers correct. Good. Now let us move on to the properties of the mean.

3.2 Properties of the mean

The mean has features that distinguish it from the other measures of central tendency. These features or properties are listed below.

1. The mean is influenced by every score or value that makes it up. If a score is changed, the values of the mean change.
   For example, for the scores, 3, 4, 2, 4, 7, the mean is 4. However, if we change 2 to 7 to get 3, 4, 7, 4, 7, the mean changes to 5. Thus a change in just one value changes the mean.
2. The mean is very sensitive to extreme scores which are called outliers.

The mean for the following scores, 4, 2, 3, 6, 5 is 4. If 3 is changed to 23, we now have, 4, 2, 23, 6, 5. The new value of 23 is an extreme score considering the fact that all the scores are below 7. The mean changes from 4 to 8. The value of 8 is greater than the majority of the scores.

3. The mean is a function of the sum (or aggregate or total) of the scores. This implies that one cannot obtain the mean without knowing the sum of the scores. If one number is missing, the mean cannot be obtained. Of the three measures it is the only one that is a function of the sum of the scores.

   This property also makes it possible to calculate the mean for a combined group if only the means and number of scores (N) are available since the $N\overline{X} = \sum X$.

   For example, Mr. Mensah's class has a mean of 5 with a class size of 20 while Ms Addo's class has a mean of 6 with a class size of 30. The mean for the combined class can be obtained by finding the sum for Mr. Mensah's class and the sum for Ms Addo's class. The results are added and divided by the total number of students. The calculation is shown below.

   Mean for the total group: $\overline{X} = \frac{(5 X 20) + (6 X 30)}{50} = \frac{280}{50} = 5.6$

4. If the mean is subtracted from each individual score and the differences are summed, the result is 0. Given the scores, 4, 2, 3, 6, 5 with a mean of 4, if we subtract the mean from each individual score and we sum up the results we will get 0. This is illustrated below.

   $4 - 4 = 0$
   $2 - 4 = -2$
   $3 - 4 = -1$
   $6 - 4 = 2$
   $5 - 4 = 1$

   The distance of the score from the mean is known as the deviation. The values of 0, -2, -1, 2, and 1 are called deviations and the sum of the deviations is 0.

5. If the same value is added to or subtracted from every number in a set of scores, the mean goes up or goes down by the value of the number.

   For example, given the scores, 8, 2, 10, 4, the mean, $\overline{X} = 6$. If we add 2 to each score we obtain, 10, 4, 12, 6, which gives a mean of $\overline{X} = 8$ which is the original mean plus the value added to each score i.e. 6 + 2

6. If each score is multiplied or divided by the same value, the mean increases or decreases by the same value.

   For example, given the scores, 8, 2, 10, 4, the mean, $\overline{X} = 6$. If we multiply each score by 3 we obtain, 24, 6, 30, 12, which gives a mean of $\overline{X} = 18$ which is the original mean times 3 i.e. 6 x 3

3.3 Strengths and weaknesses of the mean

The mean has a number of strengths and weaknesses. These are stated below.

3.3.1 Strengths of the mean

1. It uses every score in the data set. Thus every score contributes to obtaining the mean. This is not so with the median and mode as we shall see later.
2. It is the best summary score for a set of observations which is normal and there are no extreme scores.
3. It is used a lot for further statistical analysis. As we shall see later, the two other measures, median and mode, have limited statistical use.

3.3.2 Weaknesses of the mean

1. It is influenced by extreme scores. These extreme scores distort the value of the mean and results in wrong interpretation of the data.
2. It is very sensitive to a change in the value of any score. Since the mean is based on all scores, the moment one score changes, the mean will also change.
3. It cannot be computed if a score is missing and the sum of the scores or observations cannot be obtained.

3.4. Uses of the mean

As a measure of central tendency or location, the classroom teacher will find the mean useful in improving teaching and learning.

1. It is useful when the actual magnitude of the scores is needed to get an average. For example, to select a student to represent a whole class in a statistics competition, the student's total performance in statistics is used for the selection.
2. Several descriptive statistics are based on the mean. These descriptive statistics such as the standard deviation, variance, correlation coefficients, z-scores and T-scores are very useful in teaching and learning. Without the mean, they cannot be computed.
3. It is the most appropriate measure of central tendency when the scores are symmetrically distributed (i.e. normal). A symmetrical or normal distribution does not have extreme scores to influence the mean.
4. It provides a direction of performance when compared with the other measures of location especially the median. Where Mean >Median, the distribution is skewed to the right (positive skewness) showing that performance tends to be low and where Mean < Median, the distribution is skewed to the left (negative skewness) showing that performance tends to be high.
5. It serves as a standard of performance with which individual scores are compared. For example, for normally distributed scores, where the mean is 56, an individual score of 80 can be said to be far above average. Also performance can be described as just above average or far below average or just below average considering the individual scores.

In this session, you have learnt about the arithmetic mean. You have learnt how to compute the mean from both grouped and ungrouped data. In addition, you have also learnt about the properties of the mean, the strengths and weaknesses

of the mean as well as the uses of the mean. The mean is the most widely used measure of central tendency and I trust that you have grasped it well.

Self-Assessment Questions

Exercise 4.3

1. The mean score obtained by 10 students in a statistics quiz was 20 out of a total of 25. It was found later that a student who obtained 5 actually had 20. How would the discovery affect the mean?
    A. More information is needed.
    B. New mean is greater than old mean.
    C. Old mean is greater than the new mean.
    D. There is no change in the old mean.

2. A group of 20 students earned a class mean of 30 on a quiz. A second group of 30 students had a mean score of 45 on the same test. What is the mean score of the 50 students?
    A. 32.5
    B. 39.0
    C. 41.0
    D. 45.0

3. One strength of the mean as a measure of location is that it is
    A. appropriate for nominal scale variables.
    B. limited in further statistical analysis.
    C. not affected by extreme scores.
    D. useful for symmetrical distributions.

The table below gives the distribution of the ages of teachers in a school district.

Age	Number of teachers
45 - 49	25

40 - 44	36
35 - 39	77
30 - 34	47
25 - 29	15
Total	200

4. What is the value of the mean age?
    A. 35.2
    B. 36.2
    C. 37.2
    D. 38.2

5. One weakness of the mean as a measure of central tendency is that it
    A. cannot be used when data is complete.
    B. is influenced largely by extreme scores.
    C. is most appropriate for normal distributions.
    D. uses few values in a distribution.

SESSION 4: THE MEDIAN

You are welcome to the fourth session of Unit 4 for the Educational Statistics course. You remember that in Session 3, you learnt how to compute the mean. You also learnt the properties, the strengths, weaknesses and the uses of the mean. We noted that the mean is the most widely used measure of central tendency. Well, another measure of central tendency is the median and in this session you will learn how to compute the median, know the properties, the strengths and weaknesses as well as the uses.

Objectives

By the end of the session, you should be able to

(a) compute the median from a set of scores,
(b) describe the properties of the median,
(c) explain the strengths of the median,
(d) explain the weaknesses of the median,
(e) explain the uses of the median in teaching and learning.

Now read on...

4.1 Nature of the median

The median is a score for a set of observations such that approximately one-half (50%) of the scores are above it and one-half (50%) are below it when the scores are arranged sequentially. It is regarded as the 'middle score' in a distribution after the scores have been arranged from the highest to the lowest value or from the lowest value to the highest value. It is often represented by the symbol, Mdn.

4.2 Computing the median

The median can be computed from both ungrouped and grouped data.

4.2.1 Computing from ungrouped data

Suppose you have the scores, 8, 4, 9, 1, 3. To obtain the median, you arrange the scores in a sequential order, say, from the lowest score to the highest score. In the given set of scores, this gives, 1, 3, 4, 8, 9. Locate the score in the middle. This gives us 4.
In several instances, the data set you will have may not be as few as this. You may have about 40, 80, 200 scores or values. Simple formulae have been derived to help us locate the median.

To find the median, first arrange the scores in an ascending or descending order. Then for

odd set of numbers, locate the median at the $\left[\frac{n+1}{2}\right]$th position. For the even set of numbers, locate the median by adding the two middle numbers and dividing by 2. You can also find the median to be the number at the $\left[\frac{n+1}{2}\right]$th position.

Let us look at a few examples.

1. For odd set of numbers

Suppose you are given a set of observations as: 8 11 26 7 12 9 6 20 14.
There are 9 observations so this is an odd number of scores.

1. Rearrange the scores in a sequential order: 6 7 8 9 11 12 14 20 26
2. Find $\left[\frac{n+1}{2}\right]$th position i.e. $\left[\frac{9+1}{2}\right]\text{th} = \frac{10}{2}\text{th} = 5\text{th}$ position
3. The score at the 5^{th} position is 11.

The advantage with this procedure is that you do not need to rearrange the entire set of scores. When you locate the score at the required position, you stop.

2. For even set of numbers

Suppose you are given a set of numbers as: 48 52 36 54 62 71 69 45 58 32
There are 10 observations so this is an even number of scores.

1. Rearrange the scores in a sequential order: 32 36 45 48 50 54 58 62 69 71
2. Locate the two middle scores, add them and divide by two i.e. $\frac{50+54}{2} = \frac{104}{2} = 52$

   Alternatively, you can find the $\left[\frac{n+1}{2}\right]$th position, ie $\left[\frac{10+1}{2}\right]\text{th} = \frac{11}{2}\text{th} = 5\frac{1}{2}\text{th}$ position. This means that the median lies half-way between the 5^{th} and 6^{th} positions.
3. The score at the 5^{th} position is 50 and at the 6^{th} position is 54. Half-way between 50 and 54 is $\frac{(50+54)}{2} = \frac{104}{2} = 52$. The median score is therefore 52.

You notice that for the even set of numbers, both methods provide the same results.

Now obtain the median for the following set of numbers.

45, 82, 75, 87, 60, 48, 90, 72, 65, 80, 65, 49, 52, 56, 68, 72, 64, 80, 70, 58

The answer is 66.5. Have you got it right? I hope so. Congratulations. If you got it wrong, please check your calculations again. Now let us look at the grouped data.

4.2.2 Computing the median from grouped data

Computing the median involves 4 simple steps. These steps are described below.

Step 1. Obtain cumulative frequencies for the frequency distribution.

Step 2. Identify the median class. It is the class that will contain the middle score or the median. Find the value of $\frac{\sum f}{2}$,or $\frac{N}{2}$ where $\sum f$ (or N) is the total frequency. This is the position of the middle score or median. Checking from the cumulative frequency column, find the value that is equal to the position or the smallest value that is greater than the position.

Step 3. Identify the lower class boundary of the median class and the class size.

Step 4. Apply the formula below by substituting the respective values into the formula.

$$\text{Mdn} = L_1 + \left[\frac{\frac{N}{2} - cf}{f_{mdn}}\right] i \quad \text{where}$$

L_1 is the class boundary of the median class

N is the total frequency

cf is the cumulative frequency of the class just below the median class

i is the class size/width

f_{mdn} is the frequency of the median class

Now follow the example in Table 3.3

Table 3.3 Computing the median from grouped data

Classes	Midpoint	Freq	Cum Freq
	X	f	cf
46 – 50	48	4	50
41 – 45	43	6	46
36 – 40	38	10	40
31 – 35	33	12	30
26 – 30	28	8	18
21 – 25	23	7	10
16 – 20	18	3	3
Total		50	

The total frequency is 50 therefore $\frac{50}{2} = 25$ .Now there is no 25 in the cumulative frequency column so we select the smallest value that is greater than 25. This value is 30, which belongs to the 31 – 35 class. The median class therefore is 31 – 35. The lower class boundary is 30.5 and the class size is 5.

Substituting the values in the table in the formula above, we have:

$$\text{Mdn} = 30.5 + \left[\frac{\frac{50}{2} - 18}{12}\right]5 = 30.5 + \left[\frac{25-18}{12}\right]5 = 30.5 + \left[\frac{7}{12}\right]5 = 30.5 + [0.58]5 = 30.5 + 2.9 = 33.4$$

I believe that you have understood the procedure for obtaining the median from grouped data.

I want you to try and obtain the median for the distribution in the table below.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

The answer is 44.5. Have you got it right? I hope so. If you got it wrong, check the steps and the calculations. Now let us discuss the properties of the median.

4.3 Properties of the median

The median has a number of properties that distinguishes it from the other measures of central tendency. These properties are listed below.

1. It is often not influenced by extreme scores as the mean does. For example, the median for the following numbers, 2, 3, 4, 5, 6 is 4. If 6 changes to 23 as an extreme score, the median remains 4.
2. It does not use all the scores in a distribution but uses only one value.
3. It has limited use for further statistical work.
4. It can be used when there is incomplete data at the beginning or end of the distribution.
5. It is mostly appropriate for data from interval and ratio scales.
6. Where there are very few observations, the median is not representative of the data.
7. Where the data set is large, it is tedious to arrange the data in an array for ungrouped data computation of the median.

4.4 Strengths and weaknesses of the median

The Median has a number of strengths and weaknesses. These are described below.

4.4.1 Strengths of the median

1. It is not affected by extreme scores.
2. It is the most appropriate measure of central tendency when the distribution of scores is skewed.

3. It can be obtained even if data is incomplete. If data is missing at the beginning and end of the sequential arrangement, the median can still be obtained.

4.4.2 Weaknesses of the median

1. It has limited use in further statistical work. Most statistical distributions are assumed normal so the median does not come into focus much.
2. Where there are very few scores or an odd pattern of scores, the median may not be accurate. For example in a class of 20, where 15 students had 10 and 4 students had 18 and 1 student had 20, the distribution of scores looks like this: 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 18, 18, 18, 18, 20. What would be the middle score? In a situation like this, an estimate of the median may not be accurate.
3. It uses very little of the information available in the set of scores. It depends on only one score and ignores information at the ends of the distribution. It does not use all the scores in the distribution.
4. It cannot be used where the variables are from the nominal scale of measurement.
5. It is not sensitive to changes in the distribution, except where the changes occur in the middle of the distribution.

4.5 Uses of the median

As a measure of central tendency, the classroom teacher and other educational practitioners will find the median as a useful measure of central tendency (location).

1. It is used as the most appropriate measure of central tendency or location when there is reason to believe that the distribution is skewed. For skewed distributions, the best measure of central tendency, which provides a summary score or the typical score is the median.
2. It is used as the most appropriate measure of location when there are extreme scores to affect the mean. For example in an establishment of senior and junior staff, the best measure of the 'average' or 'typical' salary is the median because the senior staff salaries will inflate the mean.
3. It is useful when the exact midpoint of the distribution is wanted.
4. It provides a standard of performance for comparison with individual scores when the score distribution is skewed. For example, if the median score is 60 and an individual student obtains 55, performance can be said to be below average/median. Also performance can be described as just above average or far below average or just below average.
5. It can be compared with the mean to determine the direction of student performance.
   Where Median < Mean, the distribution is skewed to the right (positive skewness) showing that performance tends to be low and where Median > Mean, the distribution is skewed to the left (negative skewness) showing that performance tends to be high.

In this session, you have learnt about the median. You have learnt how to compute the median from both grouped and ungrouped data. In addition, you have also learnt about the properties of the median, the strengths and weaknesses of the median as well as the uses of the median. I trust that you have grasped the concept of the median well.

Self-Assessment Questions

Exercise 4.4

1. One strength of the median as a measure of location is that it is
   A. appropriate for nominal scale variables.
   B. limited in further statistical analysis.
   C. not affected by extreme scores.
   D. useful for symmetrical distributions.

2. The following scores were available for 9 students in a Statistics class.
   18 20 15 12 12 10 8 17 20
   The score for the 10^{th} student was missing but it was known to be the second highest score. What would be the median for the distribution?
   A. 12
   B. 15
   C. 16
   D. 17

3. The median score for a group of 19 students was 58. A 20^{th} student who had a score of 45 joined the group. What is the new median score?
   A. 10.5
   B. 45.0
   C. 58.0
   D. It cannot be determined

4. The median score for 15 students in a test was 67. Fourteen of the students had a median score of 66. What was the score for the 15^{th} student?
   A. 66
   B. 67
   C. 68
   D. More information is required.

5. One limitation of the median as a measure of location is that it
   A. can be used when data is incomplete.
   B. depends largely on extreme scores.
   C. is inappropriate for skewed distributions.
   D. uses few values in a distribution.

6. Compute the median age in the following distribution.

Age	Number of teachers
45 - 49	25
40 - 44	36
35 - 39	77
30 - 34	47
25 - 29	15
Total	200

SESSION 5: THE MODE

You are welcome to the fifth session of Unit 4 for the Educational Statistics course. You remember that in Sessions 3 and 4, you learnt how to compute the mean and median. You also learnt the properties, the strengths, weaknesses and the uses of the mean and median. We noted that the mean is the most widely used measure of central tendency and is appropriate for distributions that are normal. The median is well suited for distributions that are skewed. Well, another measure of central tendency is the mode and in this session you will learn how to compute the mode and know the strengths and weaknesses.

Objectives

By the end of the session, you should be able to

(a) compute the mode from a set of scores,
(b) explain the strengths of the mode,
(c) explain the weaknesses of the mode,
(d) explain the uses of the mode in teaching and learning.

Now read on...

5.1 Nature of the mode

The mode is the number in a distribution that occurs most frequently. Some data sets are such that the mean and median may not be appropriate as measures of central tendency. For example, the following scores may be obtained in a test:

50, 50, 50, 50, 50, 50, 50, 50, 80, 100

The median is not clear-cut. Which score lies in the middle of the group? This is not clearly seen. The mean however is 58 but this cannot be a score representing the group because 8 scores are below it and 2 scores are above it. In this case the mode is the most representative value for the group. The symbol used for the mode is Mo.

A distribution can have only one mode. Such distributions are called unimodal. A distribution may not have a mode at all. There are also multi-mode distributions like 2 modes (bi-modal), 3 modes (tri-modal) etc. Let us look at the following sets of data.

Set 1: 14, 14, 15, 15, 18, 18, 18, 22, 24
Set 2: 21, 24, 25, 18, 32, 50, 45, 26, 35
Set 3: 42, 42, 50, 50, 62, 62, 68, 68, 70

Pause for a moment. Which numbers occur most frequently in each of the sets above? Write them down.

Now compare what you have written with the following answers.

In Set 1, 18 occurred most frequently. It occurred 3 times. Therefore there is only one mode.
In Set 2, no number occurred most frequently. Therefore there is no mode.
In Set 3, 42, 50, 62 and 68 occurred the same number of times, i.e. 2 times. There are therefore 4 modes.

5.2 Computing the mode

The mode can be computed from both ungrouped (raw) data and grouped data (frequency distributions).

5.2.1 Computing from raw data (ungrouped data)

Suppose you are given the following set of scores:
12, 18, 21, 56, 45, 75, 48, 45, 21, 36, 35, 38, 45, 65, 72, 45, 48, 21, 45, 21
To obtain the mode, you do a visual search to determine the number that occurs most frequently. This method however, wastes a lot of time. To reduce the amount of search and the degree of errors, a tally method is recommended. Here you list the numbers and as each appears you represent it with a slash. At the end, find the value that has the most number of slashes.

The data above can be represented as follows:

Number	12	18	21	56	45	75	48	36	35	38	65	72
Tally	/	/	////	/	~~////~~/	/	//	/	/	/	/	/
Frequency	1	1	4	1	5	1	2	1	1	1	1	1

From the distribution of raw data, the mode is 45. It appeared 5 times, which is more than the others.

5.2.2 Computing from grouped data

The mode can be obtained from grouped data by three simple steps. These steps are outlined below.

Step 1. Determine the modal class i.e. the class with the highest frequency.
Step 2. Determine the lower class boundary of the modal class and the class size or width.
Step 3. Apply the following formula.

$$\text{Mode} = L_1 + \left[\frac{f_{mod} - f_{-1}}{(f_{mod} - f_1) + (f_{mod} - f_{-1})} \right] i$$

where

L_1 is the class boundary of the modal class
f_{-1} is the frequency of the class below the modal class
f_1 is the frequency of the class above the modal class
i is the class size/width
f_{mod} is the frequency of the modal class

Now follow the example in Table 3.4

Table 3.4 Computing the mode from grouped data

Classes	Freq
46 – 50	4
41 – 45	6
36 – 40	10
31 – 35	12
26 – 30	8
21 – 25	7
16 – 20	3
Total	50

The class with the highest frequency is 31 – 35. This gives the modal class. The lower class boundary of the modal class is 30.5

Applying the formula:

$$\text{Mode} = L_1 + \left[\frac{f_{mod} - f_{-1}}{(f_{mod} - f_1) + (f_{mod} - f_{-1})}\right] i = 30.5 + \left[\frac{12-8}{((12-10)+(12-8))}\right] x\ 5 = 30.5 + \left[\frac{4}{6}\right] x\ 5 = 33.8$$

If you have not understood it, go over it again and find the mode for the following distribution.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

The answer is 43.6. Have you got it right? I hope so. If you got it wrong, check the steps and the calculations.

Now let us discuss the strengths and weaknesses of the mode.

Pause for a few moments. Think about the mode. What do you think would be the strengths and weaknesses of the mode? Close your module. Write down 2 strengths and 2 weaknesses in your jotter.

Now open your module and compare what you have written with the following strengths and weaknesses.

5.3 Strengths and weaknesses of the mode

The mode has a number of strengths and weaknesses. These are stated below.

5.3.1 Strengths of the mode

1. It is easy to find from raw or ungrouped data.
2. It is not affected by extreme values or outliers.
3. It is the most appropriate measure of central tendency when the variable is nominal.
4. It is not affected by the shape of the distribution. It does not depend on whether the distribution of scores is normal or skewed.
5. When a distribution is normal, the mode is of the same value as the mean and the median.

5.3.2 Weaknesses of the mode

1. It can be absent in a distribution. i.e. a distribution may not have a mode at all.
2. A distribution can have two or more modes and it is difficult to select one as the measure of central tendency or location.
3. It does not take into account all the values in a distribution.
4. For a frequency distribution, where the data involved is discrete, it is difficult to obtain a mode which has a discrete value.
5. It has limited statistical use. For further statistics, the mode is not used because a distribution may not have a mode or may have a multi-mode.

5.4 Uses of the mode

As a measure of central tendency, the mode has limited use in improving teaching and learning due to the apparent weaknesses listed above. However, in some cases, the mode proves useful.

1. It is useful when there is the need for a rough estimate of the measure of central tendency or location. Computing the mean and median will take more time, so the mode gives a quick estimate of a summary score for a group.
2. It is useful when there is the need to know the most frequently occurring value. For example in the fashion world there may be the need to know the most common dress style. The mode provides the answer.
3. When a unique mode is available, it provides a standard of performance for comparison with individual scores. For example, if the modal score is 48 and an individual student obtains 55, performance can be said to be above average. Also performance can be described as just above average or far below average or just below average.
4. It can be compared with the mean to determine the direction of student performance.

Where Mode < Mean, the distribution is skewed to the right (positive skewness) showing that performance tends to be low and where Mode > Mean, the distribution is skewed to the left (negative skewness) showing that performance tends to be high.

5. It can be compared with the median to determine the direction of student performance. Where Mode < Median, the distribution is skewed to the right (positive skewness) showing that performance tends to be low and where Mode > Median, the distribution is skewed to the left (negative skewness) showing that performance tends to be high.

In this session, you have learnt about the mode. You have learnt how to compute the mode from both grouped and ungrouped data. In addition, you have also learnt about the strengths and weaknesses of the mode as well as the uses of the mode. I trust that you have grasped the concept of the mode well.

Self-Assessment Questions

Exercise 4.5

1. One strength of the mode as a measure of location is that it is
   A. affected by extreme scores.
   B. appropriate for nominal scale variables.
   C. not limited in further statistical analysis.
   D. sensitive to every individual score.

2. One weakness of the mode as a measure of central tendency for a distribution is that it
   A. is appropriate for nominal data.
   B. is used if there is incomplete data.
   C. provides more than one modal score.
   D. uses every score in the distribution.

3. The mode for a group of 19 students was 58. A 20^{th} student had a score of 57. What is the new mode?
   A. 20
   B. 57
   C. 58
   D. It cannot be determined
4. The mode for a group of 30 students in a test was 55. For twenty-nine of the students, the mode was 54. What was the score for the 20^{th} student?
   A. 1
   B. 54

C. 55
D. More information is required.

5. Compute the modal age in the following distribution.

Age	Number of teachers
45 - 49	25
40 - 44	36
35 - 39	77
30 - 34	47
25 - 29	15
Total	200

SESSION 6: QUARTILES

You are now in the last session of Unit 4 for the Educational Statistics course. You are doing very well by reaching this session. This Unit contains sessions on the measures of central tendency which include the mean, median and mode. For the past sessions in this Unit you have learnt about the mean, median and mode. Now you can compute the mean, median and mode. You can also state the strengths, weaknesses and the uses of the three measures of central tendency. I believe that you have clearly understood the concept of the measures of central tendency and that you can now use them to improve teaching and learning in your classroom.

In this session, you will learn about quartiles. They are not measures of central tendency but they are scores that can be used to represent sets of observations. The main use of the quartiles is to provide the needed computational skills to be able to compute a measure of variability called the quartile deviation.

Objectives

By the end of the session, you should be able to

(a) define a quartile,
(b) compute the lower quartile from a set of scores,
(c) compute the upper quartile from a set of scores.

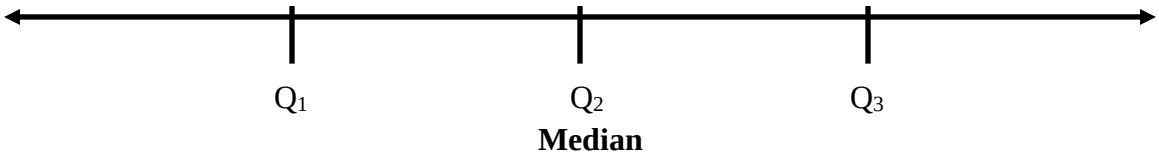
Now read on...

6.1 Nature of quartiles

Quartiles are individual scores of location that divide a distribution into 4 equal parts such that each part contains 25% of the data. Practically there are 3 quartiles – the first (lower) quartile, the second (middle) quartile and the third (upper) quartile. The second (middle) quartile is the median which you studied in Session 4. The symbols used to represent the quartiles are:

Q_1 ─ First (lower), quartile; Q_2 ─ Second (middle) quartile; Q_3─ Third (upper) quartile

The quartiles are illustrated below



6.2 Computing the Quartiles

Quartiles can be computed from both ungrouped and grouped data. Our focus is on the lower quartile and the upper quartile since we have studied the middle quartile (median) already.

6.2.1 Computing from ungrouped data.

There are two methods in computing quartiles from ungrouped data. These are the median method and the formula method.

6.2.1.1 The median method

1. First arrange the scores in a sequential order (either ascending or descending).
2. Find the overall median (i.e. score at the $\left[\frac{n+1}{2}\right]$ th position) for the data set. The overall median divides the distribution in two equal parts.
3. Find the median for the first half/part. This median becomes Q_1, the first quartile.
4. Find the median for the second half/part. This median becomes the Q_3, third quartile.

Let us look at an example.

Suppose you are given the following scores: 8, 10, 12, 7, 6, 13, 18, 25, 4, 22, 9.

1. Arrange the scores in ascending order as, 4, 6, 7, 8, 9, 10, 12, 13, 18, 22, 25
2. Median: The score at the $\left[\frac{n+1}{2}\right] = \left[\frac{11+1}{2}\right] = \frac{12}{2} = 6\text{th position}$ which is 10.

4, 6, 7, 8, 9, 10, 12, 13, 18, 22, 25

3. Find the median for the first part: 4, 6, 7, 8, 9. This gives Q_1 as 7.
4. Find the median for the second part: 12, 13, 18, 22, 25. This gives Q_3 as 18.

Now let us look at the formula method

6.2.1.2 The formula method

1. First arrange the scores in a sequential order (either ascending or descending).
2. Locate Q_1 at the $\frac{1}{4}(n+1)$ th position
3. Locate Q_3 at the $\frac{3}{4}(n+1)$ th position

Let us look at an example.

Suppose you are given the following scores: 8, 10, 12, 7, 6, 13, 18, 25, 4, 22, 9.

1. Arrange the scores in ascending order as, 4, 6, 7, 8, 9, 10, 12, 13, 18, 22, 25
2. Find the $\frac{1}{4}(11+1)$ th position. This gives us $\frac{12}{4} = 3\text{rd position}$. The score at the 3rd position is 7 which is Q_1.

3. Find the $\frac{3}{4}(11+1)$ th position. This gives us $\frac{36}{4} = 9\text{th position}$.The score at the 9th position is 18 which is Q_3.

For an even set of numbers, the positions may end up with fractions.
Let us look at an example.

Suppose you are given a set of observations as: 8 11 26 7 12 9 6 20 14 18 11 22.
There are 12 observations so this is an even number of scores.

To find the quartiles:

1. Arrange the scores in an ascending order as, 6, 7, 8, 9, 10, 11, 12, 14, 18, 20, 22, 26
2. To obtain Q_1, find the $\frac{1}{4}(12+1)$th position. This gives us $\frac{13}{4} = 3\frac{1}{4}$th position.This means that Q_1 lies between the 3rd and 4th positions. Now, at the 3rd position is 8 and the 4th position is 9. Multiply the difference between 8 and 9 with $\frac{1}{4}$. This gives you $\frac{1}{4} \text{ x } 1 = \frac{1}{4}$. Add the answer $\frac{1}{4}$ to 8 to obtain Q_1 as $8\frac{1}{4}$ or 8.25
3. To obtain Q_3, find the $\frac{3}{4}(12+1)$th position. This gives us $\frac{39}{4} = 9\frac{3}{4}$th position.This means that Q_3 lies between the 9th and 10th positions. Now, at the 9th position is 18 and the 10th position is 20. Multiply the difference between 18 and 20 with $\frac{3}{4}$. This gives you $\frac{3}{4} \text{ x } 2 = 1\frac{1}{2}$. Add the answer $1\frac{1}{2}$ to 18 to obtain Q_3 as $19\frac{1}{2}$ or 19.5

I know this will be too much for some of you. Take a 5-minute break and refresh your mind.

Welcome back. Go over Session 6.2 again. Take your time and proceed slowly.

Now obtain the lower quartile and the upper quartile for the following set of numbers using both the median method and the formula method.

45, 82, 75, 87, 60, 48, 92, 72, 65, 80, 65, 49, 52, 56, 68, 72, 64, 80, 70, 58

The answers are: Median method $Q_1 = 57$ $Q_3 = 77.5$
Formula method $Q_1 = 56.5$ $Q_3 = 78.75$

Have you got them right? I hope so. Congratulations. If you got them wrong, please check your calculations again. You will notice that the differences in the results are not big. Any of them will be accepted as correct. However, when the data set is very large, the formula method is more convenient.

Now let us look at the grouped data.

6.3 Computing the quartiles from grouped data

Computing the quartiles involves 4 simple steps. These steps are described below.

Step 1. Obtain cumulative frequencies from the frequency distribution.

Step 2. Identify the quartile classes.

For Q_1, it is the class that will contain the lower quartile. Find the value of $\frac{1}{4}\sum f$,or $\frac{N}{4}$ where $\sum f$ (or N) is the total frequency. This is the position of the lower quartile. Checking from the cumulative frequency column, find the value that is equal to the position or the smallest value that is greater than the position.

For Q_3, it is the class that will contain the upper quartile. Find the value of $\frac{3}{4}\sum f$,or $\frac{3}{4}N$ where $\sum f$ (or N) is the total frequency. This is the position of the upper quartile. Checking from the cumulative frequency column, find the value that is equal to the position or the smallest value that is greater than the position.

Step 3. Identify the lower class boundary of the lower quartile and the upper quartile classes and the class size.

Step 4. Apply the formula below by substituting the respective values into the formula.

$$Q_1 = L_1 + \left[\frac{\frac{N}{4} - cf}{f_{Q1}}\right] i \quad \text{where}$$

L_1 is the lower class boundary of the lower quartile class
N is the total frequency
cf is the cumulative frequency of the class just below the lower quartile class
i is the class size/width
f_{Q_1} is the frequency of the lower quartile class

$$Q_3 = L_3 + \left[\frac{\frac{3N}{4} - cf}{f_{Q3}}\right] i \quad \text{where}$$

L_3 is the lower class boundary of the upper quartile class
N is the total frequency
cf is the cumulative frequency of the class just below the upper quartile class
i is the class size/width
f_{Q_3} is the frequency of the upper quartile class

Now follow the example in Table 3.5

Table 3.5 Computing the quartiles from grouped data

Classes	Midpoint X	Freq f	Cum Freq cf
46 – 50	48	4	50
41 – 45	43	6	46
36 – 40	38	10	40
31 – 35	33	12	30
26 – 30	28	8	18
21 – 25	23	7	10
16 – 20	18	3	3
Total		50	

The total frequency is 50, therefore for Q_1, $\frac{50}{4} = 12.5$ .Now there is no 12.5 in the cumulative frequency column so we select the smallest value that is greater than 12.5. This value is 18, which belongs to the 26 – 30 class. The lower quartile (Q_1) class therefore is 26 – 30. The lower class boundary is 25.5 and the class size is 5. Substituting the values in the table in the formula above, we have:

$$Q_1 = 25.5 + \left[\frac{\frac{50}{4} - 10}{8}\right]5 = 25.5 + \left[\frac{12.5 - 10}{8}\right]5 = 25.5 + \left[\frac{2.5}{8}\right]5 = 25.5 + [0.31]5 = 25.5 + 1.56 = 27.1$$

For Q_3, $\frac{3}{4}$ x $50 = 37.5$ .Now there is no 37.5 in the cumulative frequency column so we select the smallest value that is greater than 37.5. This value is 40, which belongs to the 36 – 40 class. The upper quartile (Q_3) class therefore is 36 – 40. The upper class boundary is 35.5 and the class size is 5. Substituting the values in the table in the formula above, we have:

$$Q_3 = 35.5 + \left[\frac{\frac{150}{4} - 30}{10}\right]5 = 35.5 + \left[\frac{37.5 - 30}{10}\right]5 = 35.5 + \left[\frac{7.5}{10}\right]5 = 35.5 + [0.75]5 = 35.5 + 3.75 = 39.3$$

I believe that you have understood the procedure for obtaining the quartiles from grouped data.

I want you to try and obtain Q_1 and Q_3 for the distribution in the table below.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

The answers are: $Q_1 = 31.7$ and $Q_3 = 55.5$. Have you got them right? I hope so. If you got them wrong, check the steps and the calculations.

As indicated earlier, this session is to provide you with the computational tools to be used in the next session. We shall discuss the educational use of the quartiles in the next session.

In this session, you have learnt about quartiles, which are measures of location that divide a distribution into four equal parts. The quartiles are Q_1, Q_2 and Q_3. Q_2 is also known as the Median. The emphasis of this session has been on the methods of obtaining the lower quartile (Q_1) and the upper quartile (Q_3). I trust that you have acquired the computational tools to be used in Unit 4.

Self-Assessment Questions

Exercise 4.6

1. What is the lower quartile in the following distribution?

   82 90 66 78 88 72 60 80

   A. 69
   B. 78
   C. 79
   D. 85

2. What is the third quartile in the following distribution?

   14 22 8 56 46 28 30 17 29 10 60 40 33

   A. 10.5
   B. 15.5
   C. 42.0
   D. 43.0

3. What is the value of the first quartile in the following distribution?

   48 88 98 76 78 68 54 60 90 65 94

   A. 60
   B. 62.5
   C. 76
   D. 90

4. What is the third quartile in the following distribution?

   12 18 10 19 22 25 17 20 14

A. 8
B. 13
C. 18
D. 21

The distribution below is the ages of teachers. Use it to answer questions 5 and 6.

Age	Number of teachers
45 - 49	25
40 - 44	36
35 - 39	77
30 - 34	47
25 - 29	15
Total	200

5. What is the value of the first quartile?

6. Compute the value of Q_3.

This is a blank sheet for your short notes on:

- Difficult topics if any
- Issues that are not clear.

UNIT 5: MEASURES OF VARIATION (VARIABILITY)

Unit Outline

Session 1: Nature of the measures
Session 2: The range
Session 3: The variance
Session 4: The standard deviation
Session 5: Coefficient of variation
Session 6: Quartile deviation

Congratulations! You have completed Unit 4. You are now welcome to Unit 5. In Unit 4 you have studied the measures of central tendency or location. These measures provide single values to represent a set of observations or scores. The major measures you studied were the arithmetic mean, the median and the mode. In addition you learnt how to compute the lower and the upper quartiles which will be used here. In this Unit, you will learn about measures of variation or variability. These measures are also referred to as measures of dispersion, spread or scatter. The first session discusses the nature and purposes of these measures in education. The measures we will learn are the range, variance and standard deviation, coefficient of variation and the quartile deviation. You will learn how these measures are calculated, the features and how the classroom teacher can use them to enhance teaching and learning.

Unit Objectives

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

1. State and explain the purposes of the measures of variation,
2. Obtain the range from ungrouped and grouped data,
3. State the strengths, weaknesses and the uses of the range in teaching and learning,
4. Obtain the variance and standard deviation from both ungrouped and grouped data;
5. State the features of the variance and standard deviation and the uses in teaching and learning,
6. Obtain the coefficient of variation from both ungrouped and grouped data;
7. Obtain the quartile deviation from both ungrouped and grouped data;
8. State the strengths, weaknesses and the uses of the quartile deviation in teaching and learning.

SESSION 1: NATURE OF THE MEASURES OF VARIATION

You are welcome to the first session of Unit 5 for the Educational Statistics course. As noted in Unit 4, the measures of central tendency serve three main purposes. They are used as single scores to describe data, they help to show the level of performance by comparing with a given standard of performance and they also show the direction of performance.

In this session we shall first of all look at the nature of the measures of variation and the purposes they serve.

Objectives

By the end of the session, you should be able to

(a) describe the nature of the measures of variation,
(b) explain how the measures help to describe individual differences in terms of achievement,
(c) explain how the measures provide tools for further statistical analysis.

Now read on...

1.1 Nature of the measures of variation

The measures of variation are also called measures of variability, spread, dispersion or scatter. The main measures that are used mainly in educational practice are:

1. The Range
2. The Variance
3. The Standard Deviation
4. The Quartile Deviation (also known as the semi-interquartile range).

The variance and the standard deviation are closely related. The variance is the square of the standard deviation and the standard deviation is the square root of the variance. Thus if the variance is 144, the standard deviation will be 12. If the standard deviation is 9, the variance is 81.

Measures of variation provide the degree of differences within a set of observations. Let us consider the following situation.

Set 1 scores: 48, 51, 47, 50 Total = 196
Set 2 scores: 30, 72, 90, 4 Total = 196

Pause for a moment. What are the means of the two sets of scores? Write them down in the boxes below.

Set 1 Set 2

The answers are 49 and 49. Have you got them right? I believe so. Well done.

Now take a close look at the two sets of data. Have you noticed anything? Write down what you have noticed.

Now compare what you have noticed with this observation.

You will notice that in the first data set the scores are close to each other. All the scores are close to the mean of 49, or they cluster around the mean, which serves as the centre point. In the second set, the scores are far from each other. For example, 4 is so far from 90 but both sets have the same mean.

The measures of variation tell us how far the scores are from each other. This information is important for teaching and learning. If there is a big variation within a class, the teacher needs to adopt a method to suit the wide dispersion of abilities. However, if the variation is small, this means that all the students are at about the same level of performance, which may be low, moderate or high. Again the teacher needs to adopt the appropriate teaching method to suit the class.

1.2 Purposes of the measures

The measures of variation or variability serve two main purposes. These purposes are described below.

1.2.1 Purpose One

Measures of variation are used as single scores to describe differences within a set of data. They are scores that are used to indicate whether there are wide variations within a group of scores or the scores cluster around a particular value. Where there are wide variations, the group is believed to be heterogeneous and where the scores cluster around a typical value the group is homogeneous.

Let us consider the following example.

Set 1. 44, 40, 40, 42, 42, 48, 42, 42, 45, 43, 40, 40, 41, 40, 40, 41, 40, 40, 42, 46

Set 2. 20, 48, 50, 50, 50, 48, 12, 10, 55, 54, 48, 58, 59, 35, 24, 56, 30, 51, 30, 52

Pause for a moment. Calculate the mean score for both sets of data. What value have you got? Is it 32 or 42 or 48 or 50?
Your answer is correct if you got 42.

Now let us look at the highest score and the lowest score for each of the sets.

Set 1. Highest score = 48. Lowest score = 40. The difference is 8.
Set 2. Highest score = 59. Lowest score = 10. The difference is 49

You will notice that though both sets of scores have mean scores of 42, the difference between the highest and lowest scores differ. In Set 1, it is 8 units and in Set 2 it is 49 units. Set 1 can be taken to be a homogeneous group while Set 2 is a heterogeneous group.

For a heterogeneous class, the classroom teacher will notice that there are high achievers as well as low achievers. It is a mixed ability group. As a teacher, you need to adopt a method to cater for the high achievers, moderate achievers as well as the low achievers.

On the other hand, where the class is homogeneous, the teacher has to find out the level of performance by computing the measures of central tendency. In our example above, the mean is 42. Assume that the proficiency level is 30. Since 42 is higher than 30, the class can be described as performing above the proficiency level.

1.2.2 Purpose Two

They provide tools for further statistical analysis.

In Unit 1, Session 4, you learnt about descriptive and inferential statistics.

Do you remember these? Can you write down the definitions? Ok. I want you to make an attempt. Write down the definitions of descriptive and inferential statistics.
Are you done?
Good. Now turn to Unit 1, Session 4 and check your definitions. How did you do? I hope you did well.

Measures of variability are descriptive statistics. They are single numbers that are used to describe a group. In Unit 7, you will learn about correlation which is a measure of the relationship among variables and has a number of educational uses. However to be able to know the correlation or relationship between groups, you need to obtain a measure of variability. Here the most appropriate measures are the variance and the standard deviation. Knowledge of the standard deviation or variance will help you to understand the formula used in computing the correlation coefficient, which is a measure of the relationship between variables.

In Unit 6, we shall study standard scores. These standard scores (Z, T) tell us about an individual's position in relation to others. This is a very important concept in education and has applications in the classroom. As we shall see later, the formula for obtaining a Z-score is given as:

$$Z = \frac{X - \overline{X}}{s}$$

This formula implies that the value of s, the standard deviation, a measure of variability must be known before Z can be obtained.

In inferential statistics, we use information from a sample to make a general statement about a population. Inferential statistics uses a lot of statistical tools such as t and F. You will get to know

about these tools later. These tools depend on the measures of variation especially the variance and the standard deviation. Without information from the measures of variability, it will be difficult to use these inferential statistics tools.

In this session, you have learnt about what the measures of variability and the purposes that they serve. You have noted that the four main measures of variability are the range, variance, standard deviation, quartile deviation. These four measures are used to describe the extent to which there are differences within a set of observations and their use in further statistical analysis. I believe that you have grasped this introductory concept.

Self-Assessment Questions

Exercise 5.1

1. The variance for a set of scores is 25. What is the standard deviation?
    A. 2
    B. 5
    C. 25
    D. 625

2. Measures of dispersion can be used to determine the direction of student performance.
    A. False
    B. True

3. In a class homogeneous class, students perform at the same ability level in a subject.
    A. False
    B. True

4. The standard deviation for a set of scores is 9. The variance for the same set of scores is 3.
    A. True
    B. False

5. The standard deviation is indispensable in the computation of the z-score.
    A. True
    B. False

6. When a class is homogeneous, performance is always above average.
    A. True
    B. False

SESSION 2: THE RANGE

You are welcome to the second session of Unit 5 for the Educational Statistics course. You remember that in Session 1, we discussed the two uses of the measures of variation. Remember that measures of variation tell the degree of individual differences that are within a set of observations and other statistical tools depend on the measures of variation to function. In this session, we shall take one of the measures, the range and consider how helpful it is to improve teaching and learning.

Objectives

By the end of the session, you should be able to

(a) define the range,
(b) compute the range from ungrouped and grouped data,
(c) explain the strengths of the range,
(d) explain the weaknesses of the range ,
(e) explain the uses of the range in teaching and learning.

Now read on...

2.1 Nature of the range

The range is defined as the difference between the highest (largest) and the lowest (smallest) values in a set of data. For example for the set of data, 48, 51, 47, 50, the largest value is 51 and the smallest value is 47. The range is therefore 51 – 47 =4. It is the simplest of all the measures of variation.

2.2 Computing the range

The range can be computed for both raw (ungrouped) data and grouped data. The procedures are described below.

2.2.1 Computing the range from raw (ungrouped) data

Three simple steps are involved in computing the range from raw data. These steps are:

1. Determine the highest (largest) value (H) in the data set.
2. Determine the lowest (smallest) value (L) in the data set.
3. Find the difference between the two values i.e. H - L

Let us look at an example.

Given the following set of observations, determine the range.

14 22 8 56 46 28 30 17 29 10 60 40 33

The highest value (H) is 60 and the lowest value (L) is 8. The range is H – L = 60 – 8 =52.

I believe you have understood the computation of the range. It is very simple and easy.

Now obtain the range for the following set of data.

82 90 66 78 88 72 60 80

What value did you get? 20? 30? 18? 28? Well if you got 30, then your answer is correct. Congratulations. If you did not get 30, crosscheck your calculations.

Now let us look at grouped data.

2.2.2 Computing from grouped data

This method is also very easy. There are three steps.

1. Determine the lower class boundary of the bottom class and denote it as L.
2. Determine the upper class boundary of the topmost class and denote it as H.
3. Compute the range by finding the difference between H and L.

Let us work an example.

Given the following frequency distribution table, obtain the range.

Table 4.1 Computing the range from grouped data

Classes	Freq
46 – 50	4
41 – 45	6
36 – 40	10
31 – 35	12
26 – 30	8
21 – 25	7
16 – 20	3

The bottom class is 16 – 20 and the lower class boundary (L) is 15.5. The topmost class is 46 – 50 and the upper class boundary is 50.5 (H). The range is H – L = 50.5 – 15.5 = 35

The computation of the range is simple and easy and I believe you do not have any problems with it.

Pause for a moment. Reflect on the range. Close this page. Write down two strengths and two weaknesses of the range as a measure of variation in your jotter.

Now open your module and compare what you have written with the following strengths and weaknesses.

2.3 Strengths and weaknesses of the range

The range has a number of strengths and weaknesses. These are listed below.

2.3.1 Strengths of the range

1. It is easy to compute.
2. It is easy to interprete.
3. It is simple.
4. It can be used when data is incomplete and knowledge of the missing data is available.

2.3.2 Weaknesses of the range

1. It does not take into account all the data/scores. It uses only two values.
2. It ignores the actual spread of all the scores. It may therefore give a distorted picture of the variation in the data.
3. It does not consider how the scores relate to each other.
4. It does not consider the typical observations in the distribution but concentrates only on the extreme values.
5. Different distributions can have the same range which would give misleading conclusions.
6. It is only a crude or rough measure of variation

2.4 Uses of the range

Due to the numerous weaknesses, the range has limited use.

1. When data is too scanty or too scattered to justify the computation of a more precise measure, the range provides a fair estimate of the extent of variability available.
2. It may be necessary to require knowledge of only the extreme scores or total spread in a set of observations. In a test, a teacher may be interested in only the highest score and the lowest. The range will conveniently serve that purpose.

In this session, you have learnt about the range. You have learnt how to compute the range from both ungrouped and grouped data. In addition, you have also learnt about the strengths and weaknesses of the range as well as the uses of the range. I trust that you have grasped the concept of the range well.

Self-Assessment Questions

Exercise 5.2

Compute the range for the following sets of data.

1. 18, 22, 48, 45, 90, 93, 65, 62, 28, 75, 15, 30, 35, 80, 82

2. 44, ─ 8, 14, ─ 14, 24, 28, ─ 30, 52, 58, 40, 42, 48, 50, ─ 1

3. ─ 4, ─ 15, ─ 18, ─ 56, ─ 52, ─ 40, ─ 75, ─ 18, ─ 36, ─ 19, ─ 50, ─ 55, 0,

4. What is the range in the following distribution?

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

5. One strength of the range as a measure of variation is that it
   A. can be used when data is incomplete.
   B. depends largely on extreme scores.
   C. disregards the actual spread of the scores.
   D. uses few values in a distribution.

SESSION 3 THE VARIANCE

You are welcome to the third session of Unit 5 for the Educational Statistics course. You remember that in Session 2, we discussed the range which is the difference between the highest value and the lowest value in a set of scores. It is used mostly as a measure of variation when there are very few observations and when knowledge of extreme scores is all that is needed. It has limited use in educational practice. In this session, we shall learn about the variance which is one of the most popular measures of variation.

Objectives

By the end of the session, you should be able to

(a) define the variance,
(b) compute the variance from ungrouped and grouped data,
(c) explain the strengths of the variance,
(d) explain the weaknesses of the variance ,
(e) explain the uses of the variance in teaching and learning.

Now read on...

3.1 Nature of the variance

The variance is the mean square deviation. It is defined as the mean of the squares of the deviations of the scores from the mean of the distribution. The symbols used are σ^2 or $S^2$for population variance and s^2 for sample variance.

If the definition is not clear, do not worry. As we move on and we work examples, it will become clearer.

3.2 Computing the variance

The variance can be computed from both the raw (ungrouped) data and grouped data.

3.2.1 Computing from raw data (ungrouped data)

The variance can be computed from raw data by using two formulae. These are the conventional formula and the computational formula. The procedures are described below.

3.2.1.1 Using the conventional formula.

Five easy steps are involved in this method.

1. Compute the arithmetic mean. $\overline{X}$
2. Find the deviation of the scores from the mean. $X - \overline{X}$
3. Square the deviations. $\left[X - \overline{X}\right]^2$
4. Find the sum of the squared deviations. $\sum\left[X - \overline{X}\right]^2$

5. Find the mean of the squared deviations. $\frac{\sum\left[X - \overline{X}\right]^2}{N}$

In general, the procedure gives us the formula: $s^2 = \frac{\sum\left[X - \overline{X}\right]^2}{N}$

Let us work an example.

Given the set of scores, find the variance.
15, 12, 10, 10, 9, 20, 14, 11, 13, 16
The mean is 13.

Score	Deviation $X - \overline{X}$	Deviation squared $\left[X - \overline{X}\right]^2$
15	15-13=2	4
12	12-13=-1	1
10	10-13=-3	9
10	10-13=-3	9
9	9-13=-4	16
20	20-13=7	49
14	14-13=1	1
11	11-13=-2	4
13	13-13=0	0
16	16-13=3	9
Total		102

$$s^2 = \frac{\sum\left[X - \overline{X}\right]^2}{N} = \frac{102}{10} = 10.2$$

3.2.1.2 Using the computational formula

For this method, six steps are involved.

1. Find the sum of the individual scores. $\sum X$
2. Divide the sum by the total frequency. $\frac{\sum X}{N}$
3. Square the result. $\left[\frac{\sum X}{N}\right]^2$
4. Square the individual scores, X^2
5. Find the sum of the squared individual scores and divide by N. $\left[\frac{\sum X^2}{N}\right]$

6. Find the difference between the result in Step 5 and the result in Step 3. This gives the variance.

The procedure is summarized in the formula below.

$$s^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2$$

Let us work an example using this formula.

Given the set of scores, find the variance.

15, 12, 10, 10, 9, 20, 14, 11, 13, 16

Observation Number	X	X^2
1	15	225
2	12	144
3	10	100
4	10	100
5	9	81
6	20	400
7	14	196
8	11	121
9	13	169
10	16	256
Total	130	1792

$$s^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 = \frac{1792}{10} - \left(\frac{130}{10}\right)^2 = 179.2 - 13^2 = 179.2 - 169 = 10.2$$

You will notice both methods have produced similar results. One advantage of the computational method over the conventional method is that fractions are not involved much in the computations. In the conventional method, when the mean is a fraction, the deviations are more difficult to work with.

Now go over the methods again and make sure you have understood every step.

Compute the variance for the data set below using both methods in your jotter.

18, 22, 16, 12, 15, 19, 17, 20, 15, 11

What are your answers? If your answer is 10.65, then you are right. If you got it wrong, then check your answers again. Which of the methods did you find easier?

Now let us compute the variance from the grouped data.

3.2.2 Computing from grouped data

Three methods are generally used. These are the conventional method, computational method and the coding method. These methods are used with the frequency distribution table.

3.2.2.1 Using the conventional method

Six steps are involved with this procedure. These are listed below.

1. Compute the arithmetic mean. $\overline{X}$
2. Find the deviation of the class marks (X) from the mean. $X - \overline{X}$
3. Square the deviations. $\left[X - \overline{X}\right]^2$
4. Multiply the squared deviations with the frequency. $f\left[X - \overline{X}\right]^2$
5. Find the sum of the product of frequency and squared deviations. $\sum f\left(X - \overline{X}\right)^2$
6. Divide the result in Step 5 by the total frequency, N. $\frac{\sum f\left(X - \overline{X}\right)^2}{N}$ to get the variance.

In general, the procedure gives us the formula: $s^2 = \frac{\sum f\left(X - \overline{X}\right)^2}{N}$

This method is generally long and can be tedious.

3.2.2.2 Using the computational method

For this method, six steps are involved.

1. Multiply the class marks with the frequency. fx
2. Find the sum of the product in Step 1. $\sum fx$
3. Divide the sum in Step 2 by the total frequency and square the result. $\left(\frac{\sum fx}{N}\right)^2$
4. Square the class marks, multiply with the frequencies and find the sum. $\sum fx^2$
5. Divide the result in Step 4 by total frequency, N. $\frac{\sum fx^2}{N}$.
6. Find the difference between result in 5 and result in 3. This gives the variance.

The procedure is summarized in the formula below.

$$s^2 = \frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2$$

Now let us apply this formula in working through an example.

Given the following frequency distribution table, obtain the variance.

Table 4.2 Computing the variance from grouped data

Classes	Class Mark X	Frequency f	X^2	fX^2	fX
46-50	48	4	2304	9216	192
41-45	43	6	1849	11094	258
36-40	38	10	1444	14440	380
31-35	33	12	1089	13068	396
26-30	28	8	784	6272	224
21-25	23	7	529	3703	161
16-20	18	3	324	972	54
Total		50		58765	1665

Applying the formula gives:

$$s^2 = \frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2 = \frac{58765}{50} - \left(\frac{1665}{50}\right)^2 = 1175.3 - 33.3^2 = 1175.3 - 1108.89 = 66.4$$

3.2.2.3 Using the coding method

This method involves a number of steps. Before you use the coding method make sure that all class intervals are of equal size.

Step 1. Obtain the class midpoints or class marks.

Step 2. Create a new column after the frequency column and give it a heading, d.

Step 3. Choose the class that is in the middle of the distribution, but if there is no class exactly in the middle choose one of the two middle classes, preferably the class with the higher frequency. Under the column, d, code this class with '0'(zero).

Step 4. Give a code of 1 to the class immediately above the class coded 0. The next higher class is given a code of 2, the next higher one, a code of 3. Continue until you reach the topmost class.

Step 5. Give a code of −1 to the class immediately below the class coded 0. The next lower class is given a code of −2, the next lower one, a code of −3. Continue until you reach the bottom class.

Step 6. Create another column, fd, where you put in the values of the product of the frequencies and the codes.

Step 7. Add the values in the fd column, divide by the total frequency, N and square the result.

Step 8. Create another column, d^2 and fill it with the squares of the codes.

Step 9. Create an fd^2 column and fill it with the product of the frequency and d^2

Step 10. Sum up the values in the fd^2 column and divide by the total frequency, N

Step 11. Find the difference between the result in Step 10 and Step 7.

Step 12. Multiply the result in Step 11 with the square of the class size, i^2.

The 12 steps above are summarized in the formula for the coding method as:

$$s^2 = \left[\left(\frac{\sum fd^2}{N}\right) - \left(\frac{\sum fd}{N}\right)^2\right] x\, i^2$$

Now follow the example in Table 4.3.

Given the following frequency distribution table, obtain the variance by using the coding method.

Table 4.3 Computing the variance from grouped data using coding method

Classes	Class Mark X	Frequency f	d	fd	d^2	fd^2
46-50	48	4	3	12	9	36
41-45	43	6	2	12	4	24
36-40	38	10	1	10	1	10
31-35	33	12	0	0	0	0
26-30	28	8	-1	-8	1	8
21-25	23	7	-2	-14	4	28
16-20	18	3	-3	-9	9	27
Total		50		3		133

Applying the formula gives:

$$s^2 = \left[\left(\frac{\sum fd^2}{N}\right) - \left(\frac{\sum fd}{N}\right)^2\right] x\, i^2 = \left[\left(\frac{133}{50}\right) - \left(\frac{3}{50}\right)^2\right] x\, 25 = (2.66 - 0.0036)\, x\, 25 = 66.4$$

You will notice that both methods yielded the same results.

Pause for a moment. Reflect. Have you clearly followed the procedures for computing the variance? If No, go over the procedures again.

Now compute the variance for the following data using both the computational and coding methods.

Classes	Frequency
61 - 70	15
51 – 60	20

41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

What answers have you got? Are they the same as 238.24? If yes, well done. If no, please go back and check your calculations.

Ok. Let us move on to the properties of the variance.

3.3 Properties of the variance

The variance has features that make it unique as a measure of variability. These properties are outlined below.

1. The variance of a constant is zero. For example, given the scores, 8 8 8 8 8 8 8 8 8 8, the variance is zero because there are no differences in the scores.
2. It is not resistant. It is affected by extreme scores or outliers. For example, given the scores, 5, 8, 10, 6, 7, the variance is 3.7. However, if the scores become, 5, 8, 10, 6, 24, where 7 is replaced with 24 as an extreme score, the variance changes from 3.7 to 59.8
3. The variance is independent of change of origin. If each score in a set of data is reduced or increased by the same amount, the variance of the new set of data does not change. For example, given the data 5, 8, 10, 6, 7, with a variance is 3.7, if 10 points is added to each score to obtain 15, 18, 20, 16, 17 the variance remains unchanged.
4. The variance is not independent of change of scale. If each score in a set of data is multiplied or divided by the same amount, say a constant k, the resulting variance equals k^2 multiplied by the old variance. Suppose you are given the data, 48, 51, 47, 50 with a variance of 2.5. If each score is multiplied by 10 points to obtain 480, 510, 470, 500, the variance becomes 10^2 x 2.5=250.

3.4 Strengths and Weaknesses of the variance

The variance has a number of strengths and weaknesses. These are listed below.

3.4.1 Strengths of the variance

1. It uses every score in the data set. Thus every score contributes to obtaining the variance. This is not the same with the range and the quartile deviation
2. It is used a lot for further statistical analysis. As we shall see later, some of the measures of variability like the range and the quartile deviation have limited use in further statistical analysis.
3. It is most appropriate for scores that are normally distributed.

3.4.2 Weaknesses of the variance

1. It is influenced by extreme scores. It gives more weight to these extreme scores resulting in a wrong interpretation of the results.
2. It is sensitive to a change in the value of any score in the distribution. If one score changes, the variance also changes.

3. It cannot be computed if missing data is reported since the variance depends on every individual score.
4. It is not appropriate for judging the variation within a set of observations. This is because the variance is obtained in terms of squared units and thus the measurement units like years, metres, kilograms, litres are computed in squares.

3.5 Uses of the variance

1. The usefulness of the variance in educational practice is based on the standard deviation. You may recall that the variance is the square of the standard deviation. Therefore the uses of the standard deviation in educational practice are related to the variance. We shall learn more about this in the next session.
2. The variance is used a lot in inferential statistics. As we noted in Session 1 of this Unit, in inferential statistics, we use information from a sample to make a general statement about a population. Statistical tools used include t-tests and F-tests. These tests use the variance in their computations. Without information from the variances it will be difficult to obtain results for interpretation.

Now let us revise what we have learnt in this session. You have learnt about the nature of the variance and how to compute the variance from both ungrouped and grouped data. In addition, you have also learnt about the properties, strengths and weaknesses of the variance as well as the uses of the variance. I trust that you have a good understanding of this important measure of variation in statistics.

Self-Assessment Questions

Exercise 5.3

1. One weakness of the variance as a measure of dispersion is that it is
   A. affected by extreme scores.
   B. dependent on all scores.
   C. useful for normal distributions.
   D. useful for further analysis.

2. Compute the variance for the following set of scores.

   82 90 66 78 88 72 60 80

3. Compute the variance for the following distribution.

Classes	Frequency

61 - 70	15
51 – 60	20
41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

SESSION 4 THE STANDARD DEVIATION

You are welcome to the fourth session of Unit 5 for the Educational Statistics course. You remember that in Session 3, we discussed the variance which is the mean of the squared deviations for a set of scores. You also learnt how to compute the variance from both ungrouped and grouped data. The properties, strengths, weaknesses and uses were discussed. In this session, we shall learn about the standard deviation which is the most popular measure of variation. As you remember from Session 1, a special relationship exists between the variance and the standard deviation. We can say that they are "bed-fellows". They always walk together. When you see the standard deviation, you can see the variance. This relationship is that the variance is the square of the standard deviation and the standard deviation is the square root of the variance.

Objectives

By the end of the session, you should be able to

(a) define the standard deviation,
(b) compute the standard deviation from ungrouped and grouped data,
(c) explain the strengths of the standard deviation,
(d) explain the weaknesses of the standard deviation ,
(e) explain the uses of the standard deviation in teaching and learning.

Now read on...

4.1 Nature of the standard deviation

The standard deviation is the most used measure of variation. It is the square root of the mean square deviation. It is defined as the square root of the mean of the squares of the deviations of the scores from the mean of the distribution. It is a more stable measure of variability than the range and the quartile deviation. The symbols used are σ or S for population standard deviation and s for sample standard deviation.

I trust that this definition is clear since you have had a good study of the variance.

4.2 Computing the standard deviation

The standard deviation can be computed from both the raw (ungrouped) data and grouped data.

4.2.1 Computing from raw data (ungrouped data)

The standard deviation can be computed from raw data by using two formulae. These are the conventional formula and the computational formula. The procedures are described below.

4.2.1.1 Using the conventional formula.

Six easy steps are involved in this method.

1. Compute the arithmetic mean. $\overline{X}$

2. Find the deviation of the scores from the mean. $X - \overline{X}$
3. Square the deviations. $\left[X - \overline{X}\right]^2$
4. Find the sum of the squared deviations. $\sum\left[X - \overline{X}\right]^2$
5. Find the mean of the squared deviations. $\frac{\sum\left[X - \overline{X}\right]^2}{N}$
6. Find the square root of the result in Step 5. $\sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{N}}$

In general, the procedure gives us the formula: $s = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{N}}$

Let us work an example.

Given the set of scores, find the standard deviation.
15, 12, 10, 10, 9, 20, 14, 11, 13, 16
The mean is 13.

Score	Deviation $X - \overline{X}$	Deviation squared $\left[X - \overline{X}\right]^2$
15	15-13=2	4
12	12-13=-1	1
10	10-13=-3	9
10	10-13=-3	9
9	9-13=-4	16
20	20-13=7	49
14	14-13=1	1
11	11-13=-2	4
13	13-13=0	0
16	16-13=3	9
Total		102

$$s = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{N}} = \sqrt{\frac{102}{10}} = \sqrt{10.2} = 3.19$$

4.2.1.2 Using the computational formula

For this method, seven steps are involved.

1. Find the sum of the individual scores. $\sum X$

2. Divide the sum by the total frequency. $\frac{\sum X}{N}$
3. Square the result. $\left[\frac{\sum X}{N}\right]^2$
4. Square the individual scores, X^2
5. Find the sum of the squared individual scores and divide by N. $\left[\frac{\sum X^2}{N}\right]$
6. Find the difference between the result in Step 5 and the result in Step 3.
7. Find the square root of the result in Step 6.

The procedure is summarized in the formula below.

$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}$$

Let us work an example using this formula.

Given the set of scores, find the standard deviation.

15, 12, 10, 10, 9, 20, 14, 11, 13, 16

Observation Number	X	X^2
1	15	225
2	12	144
3	10	100
4	10	100
5	9	81
6	20	400
7	14	196
8	11	121
9	13	169
10	16	256
Total	130	1792

$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\frac{1792}{10} - \left(\frac{130}{10}\right)^2} = \sqrt{179.2 - 13^2} = \sqrt{179.2 - 169} = \sqrt{10.2} = 3.19$$

You will notice both methods have produced similar results. One advantage of the computational method over the conventional method is that fractions are not involved much in the computations. In the conventional method, when the mean is a fraction, the deviations are more difficult to work with.

Now go over the methods again and make sure you have understood every step.

Compute the standard deviation for the data set below using both methods in your jotter.

18, 22, 16, 12, 15, 19, 17, 20, 15, 11

What are your answers? If your answer is 3.26, then you are right. If you got it wrong, then check your answers again. Which of the methods did you find easier?

Now let us compute the standard deviation from the grouped data.

4.2.2 Computing from grouped data

Three methods are generally used. These are the conventional method, computational method and the coding method. These methods are used with the frequency distribution table.

4.2.2.1 Using the conventional method

Seven steps are involved with this procedure. These are listed below.

1. Compute the arithmetic mean. $\overline{X}$
2. Find the deviation of the class marks (X) from the mean. $X - \overline{X}$
3. Square the deviations. $\left[X - \overline{X}\right]^2$
4. Multiply the squared deviations with the frequency. $f\left[X - \overline{X}\right]^2$
5. Find the sum of the product of frequency and squared deviations. $\sum f\left(X - \overline{X}\right)^2$
6. Divide the result in Step 5 by the total frequency, N. $\dfrac{\sum f\left(X - \overline{X}\right)^2}{N}$
7. Find the square root of the result in Step 6. $\sqrt{\dfrac{\sum f\left(X - \overline{X}\right)^2}{N}}$

In general, the procedure gives us the formula: $s = \sqrt{\dfrac{\sum f\left(X - \overline{X}\right)^2}{N}}$

This method is generally long and can be tedious.

4.2.2.2 Using the computational method

For this method, seven steps are involved.

1. Multiply the class marks with the frequency. fx
2. Find the sum of the product in Step 1. $\sum fx$

3. Divide the sum in Step 2 by the total frequency and square the result.
$\left(\frac{\sum fx}{N}\right)^2$
4. Square the class marks, multiply with the frequencies and find the sum.
$\sum fx^2$
5. Divide the result in Step 4 by total frequency, N. $\left(\frac{\sum fx^2}{N}\right)$
6. Find the difference between the result in Step 5 and the result in Step 3.
$\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2$
7. Find the square root of the result in Step 6.

The procedure is summarized in the formula below.

$$s = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

Now let us apply this formula in working through an example.

Given the following frequency distribution table, obtain the standard deviation.

Table 4.3 Computing the standard deviation from grouped data

Classes	Class Mark X	Frequency f	X^2	fX^2	fX
46-50	48	4	2304	9216	192
41-45	43	6	1849	11094	258
36-40	38	10	1444	14440	380
31-35	33	12	1089	13068	396
26-30	28	8	784	6272	224
21-25	23	7	529	3703	161
16-20	18	3	324	972	54
Total		50		58765	1665

Applying the formula gives:

$$s = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2} = \sqrt{\frac{58765}{50} - \left(\frac{1665}{50}\right)^2} = \sqrt{1175.3 - 33.3^2} = \sqrt{1175.3 - 1108.89} = \sqrt{66.41} = 8.15$$

4.2.2.3 Using the coding method

This method involves a number of steps.
Before you use the coding method, make sure that all class intervals are of equal size.

Step 1. Obtain the class midpoints or class marks.
Step 2. Create a new column after the frequency column and give it a heading, d.
Step 3. Choose the class that is in the middle of the distribution, but if there is no class exactly in the middle choose one of the two middle classes preferably the class with the higher frequency. Under the column, d, code this class with '0'(zero).
Step 4. Give a code of 1 to the class immediately above the class coded 0. The next higher class is given a code of 2, the next higher one, a code of 3. Continue until you reach the topmost class.
Step 5. Give a code of −1 to the class immediately below the class coded 0. The next lower class is given a code of −2, the next lower one, a code of −3. Continue until you reach the bottom class.
Step 6. Create another column, fd, where you put in the values of the product of the frequencies and the codes.
Step 7. Add the values in the fd column, divide by the total frequency, N and square the result.
Step 8. Create another column, d^2 and fill it with the square of the codes.
Step 9. Create an fd^2 column and fill it with the product of the frequency and d^2
Step 10. Sum up the values in the fd^2 column and divide by the total frequency, N
Step 11. Find the difference between the result in Step 10 and Step 7.
Step 12. Find the square root of the result in Step 11.
Step 13. Multiply the result in Step 12 with the class size, i.

The 13 steps above are summarized in the formula for the coding method as:

$$s = i\,x\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Now follow the example in Table 4.4.

Given the following frequency distribution table, obtain the standard deviation by using the coding method.

Table 4.4 Computing the standard deviation from grouped data using coding method

Classes	Class Mark X	Frequency f	d	fd	d^2	fd^2
46-50	48	4	3	12	9	36
41-45	43	6	2	12	4	24
36-40	38	10	1	10	1	10

31-35	33	12	0	0	0	0
26-30	28	8	-1	-8	1	8
21-25	23	7	-2	-14	4	28
16-20	18	3	-3	-9	9	27
Total		50		3		133

Applying the formula gives:

$$s = \text{i}\, x \sqrt{\frac{\sum \text{fd}^2}{\text{N}} - \left(\frac{\sum \text{fd}}{\text{n}}\right)^2} = 5\sqrt{\frac{133}{50} - \left(\frac{3}{50}\right)^2} = 5\sqrt{2.66 - 0.0036} = 8.15$$

You will notice that both methods yielded the same results.

Pause for a moment. Reflect. Have you clearly followed the procedures for computing the standard deviation? If No, go over the procedures again.

Now compute the standard deviation for the following data using both the computational and coding methods.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

What answers have you got? Are they the same as 15.4? I hope so. Good job.

Ok. Let us move on to the properties of the standard deviation.

4.3 Properties of the standard deviation

The standard deviation has features that make it unique as a measure of variability. These properties are outlined below.

1. The standard deviation of a constant is zero. For example, given the scores, 10, 10, 10, 10, 10, 10, 10, 10, the standard deviation is zero because there are no differences in the scores.

2. It is not resistant. It is affected by extreme scores or outliers. For example, given the scores, 5, 8, 10, 6, 7, the standard deviation is 1.9. However, if the scores become, 5, 8, 10, 6, 24, where 7 is replaced with 24 as an extreme score, the standard deviation changes from 1.9 to 7.7
3. The standard deviation is independent of change of origin. If each score in a set of data is reduced or increased by the same amount, the standard deviation of the new set of data does not change. For example, given the data 5, 8, 10, 6, 7, with a standard deviation of 1.9, if 10 points is added to each score to obtain 15, 18, 20, 16, 17 the standard deviation remains unchanged.
4. The standard deviation is not independent of change of scale. If each score in a set of data is multiplied or divided by the same amount, say a constant k, the resulting standard deviation equals k multiplied by the old standard deviation. Suppose you are given the data, 48, 51, 47, 50 with a standard deviation of 1.58. If each score is multiplied by 10 points to obtain 480, 510, 470, 500, the standard deviation becomes 10 x 1.58 = 15.8.

4.4 Strengths and Weaknesses of the standard deviation

The standard deviation has a number of strengths and weaknesses. These are listed below.

4.4.1 Strengths of the standard deviation

1. It uses every score in the data set. Thus every score contributes to obtaining the standard deviation. This is not the same with the range and the quartile deviation.
2. It is used a lot for further statistical analysis. As we shall see later, some of the measures of variability like the range and the quartile deviation have limited use in further statistical analysis.
3. It is the most appropriate measure of variability for scores that are normally distributed.
4. It is easy to interpret. It represents the mean or 'average' of the individual amounts of variability. Thus the larger this mean, the greater the variation within the data set.

4.4.2 Weaknesses of the standard deviation

1. It is influenced by extreme scores. It gives more weight to these extreme scores resulting in a wrong interpretation of the results.
2. It is sensitive to a change in the value of any score in the distribution. If one score changes, the standard deviation also changes.
3. It cannot be computed if missing data is reported since the standard deviation depends on every individual score.

4.5 Uses of the standard deviation

1. It is used as the most appropriate measure of variation or variability when there is reason to believe that the distribution is normal. For a distribution to be normal, the mean, median and mode must be exactly or approximately equal. In a situation like this, to know the extent of the individual differences within the group, the standard deviation has to be computed.
2. It helps to determine whether a group is homogeneous with respect to a variable or it is heterogeneous. In terms of classroom achievement, the standard deviation gives us information about whether the class is of the same ability group or a mixed ability group. If

the standard deviation is small, the class is homogeneous, but if the standard deviation is large the class is heterogeneous.

As a rough estimate, if scores have a maximum of 50, a standard deviation of 5 or less is considered small, producing a homogeneous class. On the other hand, for scores more than 50, a standard deviation of 10 or less gives a homogenous class. However, to be more exact, the coefficient of variation is computed. Where the coefficient of variation (CV) is 33 or less than 33, the group is homogeneous, but where it is greater than 33, the group is heterogeneous. We shall learn more about the coefficient of variation in the next session.

The nature of the group has implications for teaching and learning. For a heterogeneous class, abilities are mixed. In the teaching, the teacher has to cater for the high-achievers as well as the low-achievers. Attention should also be paid to the middle group. In this case, ability group teaching may be considered. For a homogeneous group, the teacher needs to find the level of performance by examining the measures of central tendency. This will determine whether the group is a high-achieving group or a low-achieving group. In all situations, the teacher needs to adopt a teaching method to suit the class.

4. It is helpful in computing other statistics, which have educational importance. For example standard deviations are needed to obtain standard scores and correlation coefficients. The importance of these statistics will be known in later sessions of this course.
5. It is useful in determining the reliability of test scores. Reliability is the degree of consistency of assessment results. Reliability helps us to know how much confidence to put in the interpretations of students assessment results. Methods of computing reliability include the split-half correlation, Kuder-Richardson and coefficient alpha. These methods depend on the standard deviation of the scores.

Now let us revise what we have learnt in this session. You have learnt about the nature of the standard deviation and how to compute it from both ungrouped and grouped data. In addition, you have also learnt about the properties, strengths and weaknesses of the standard deviation as well as the uses of the standard deviation. I trust that you have a good understanding of this important measure of variation in statistics.

Self-Assessment Questions

Exercise 5.4

1. When a distribution is normal, the most appropriate measure of variability is the standard deviation.

    A. True
    B. False.

2. One weakness of the standard deviation as a measure of dispersion is that it is
    A. affected by extreme scores.
    B. dependent on all scores.
    C. useful for normal distributions.
    D. useful for further analysis.

3. In a Computer Studies class, a mean of 30 and a standard deviation of 10 were obtained in a quiz. The instructor later multiplied each score by 2. What is the new standard deviation of the scores?

    A. 10
    B. 20
    C. 40
    D. 60

4. Compute the standard deviation for the following set of scores.

    82 90 66 78 88 72 60 80 50 70

5. Compute the standard deviation for the following distribution.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	35
31 – 40	27
21 – 30	12
11 – 20	11

SESSION 5: COEFFICIENT OF VARIATION

You are welcome to the fifth session of Unit 5 for the Educational Statistics course. You remember that in Sessions 3 and 4, we discussed the variance and standard deviation as measures of variation. These two measures are the most popular ones in educational practice. Now you can compute them and state the properties, strengths, weaknesses and educational uses. We have also noted that the value of the standard deviation is a guide to determine if a class is homogeneous or heterogeneous. A large standard deviation shows a heterogeneous class and a small standard deviation indicates a homogeneous class. In discussing homogeneity and heterogeneity, we mentioned the coefficient of variation. In this session we shall know more about the coefficient of variation.

Objectives

By the end of the session, you should be able to

(a) define the coefficient of variation,

(a) compute the coefficient of variation from ungrouped and grouped data,

(c) explain the uses of the coefficient of variation in teaching and learning.

Now read on...

5.1 Nature of the coefficient of variation (CV)

It is considered a relative measure of variation. It is defined as the ratio of the standard deviation to the mean. It is often expressed as a percentage, so that the value is multiplied by 100. It is only defined for non-zero means, and is most useful for variables that are always *positive*. It is appropriate for ratio and interval scales of measurement.

5.2 Computing the coefficient of variation

The coefficient of variation (CV) can be computed for both raw (ungrouped) data and grouped data. The procedures are described below.

5.2.1 Computing CV from raw (ungrouped) data

Three simple steps are involved in computing the coefficient of variation from raw data. These steps are:

1. Compute the mean.
2. Compute the standard deviation.
3. Divide the standard deviation by the mean and multiply with 100.

This is summarized in the formula; $CV = \frac{S}{\bar{X}} \times 100$

Let us look at an example.

Given the following set of observations, determine the coefficient of variation.

14, 22, 8, 56, 46, 28, 30, 17, 29, 10, 60, 40, 30, 20, 25

$$CV = \frac{S}{\overline{X}} \text{ x } 100 = \frac{15.1}{29} \text{ x } 100 = 52.06\%$$

I believe you have understood the computation of the coefficient of variation. It is very simple and easy.

Now obtain the coefficient of variation for the following set of data.

82 90 66 78 88 72 60 80

What value did you get? The answer is 12.6%. Did you get it right? Congratulations. If you did not get 12.6, crosscheck your calculations. The mean is 77 and the standard deviation is 9.7.

Now let us look at grouped data.

5.2.2 Computing from grouped data

This method is also very easy. There are three steps as in ungrouped data.

1. Compute the mean.
2. Compute the standard deviation.
3. Divide the standard deviation by the mean and multiply with 100.

This is summarized in the formula; $CV = \frac{S}{\overline{X}} \text{ x } 100$.

Let us work an example.
Given the following frequency distribution table, obtain the coefficient of variation.

Table 4.5 Computing the coefficient of variation from grouped data

Classes	Freq
46 – 50	5
41 – 45	8
36 – 40	12
31 – 35	15
26 – 30	6
21 – 25	3
16 – 20	1

First compute the mean. This gives you 35.8. Secondly compute the standard deviation. This gives you 7.05. The coefficient of variation becomes:

$$CV = \frac{S}{\bar{X}} \text{ x } 100 = \frac{7.05}{35.8} \text{ x } 100 = 19.7\%$$

The computation of the coefficient of variation is simple and easy and I believe you do not have any problems with it.

Pause for a moment. Reflect on the coefficient of variation. Close this page. Write down one strength and one weakness of the coefficient of variation as a relative measure of variation in your jotter.

Now open your module and compare what you have written with the following strengths and weaknesses.

5.3 Strengths and weaknesses of the coefficient of variation

The coefficient of variation has a number of strengths and weaknesses. These are listed below.

5.3.1 Strengths of the coefficient of variation

1. It is easy to compute.
2. It is unitless and this makes it possible to compare variability for different distributions. It does not, for example, have metres, kilogrammes, years etc attached to it. The standard deviation, range, variance etc have units attached to them.
3. Where the distribution is normal, it is based on every score in the distribution.
4. It is easy to interpret. The larger the value of the CV, the greater the variability.

5.3.2 Weaknesses of the coefficient of variation

1. It is affected by extreme values. When extreme values affect the mean and the standard deviation, this in turn affects the coefficient of variation.
2. It cannot be used when the mean is negative, zero or near zero. It is only used for distributions with positive values.
3. It is sensitive to a change in the value of any score in the distribution. If one score changes, both the mean and the standard deviation change and so does the CV.
4. It cannot be computed for a variable that is normally distributed if missing data is reported. Since the mean and the standard deviation depend on every individual score it will not be possible to obtain these values.

5.4 Uses of the coefficient of variation

The coefficient of variation has three main applications.

1. It is used to determine whether a group is homogeneous or heterogeneous. If the value of the CV is 33% or less, then the group is homogeneous, otherwise, it is heterogeneous. As we saw in Session 4, the class teacher needs to adopt a suitable method for each group.

2. It is used to compare variations within or between groups where there are different units of measurement. For example, the mean height of 40 students is 150 cm with a standard deviation of 12 cm. The mean weight of the students is 60 kg with a standard deviation of 7 kg. Looking

at the standard deviations, we would say that there is more variation in height than weight. But this will be like comparing yams to fish. The best comparison will be using the CVs. The CV for height is 8% and for weight is 11.7%. It is therefore more reasonable to say that there is more variation in weight than height.

3. It is used to compare variations within or between groups where there are different means but with the same unit of measurement. For example, the mean height of 40 students in a Statistics class is 150 cm with a standard deviation of 12 cm. In another class of 35 students, the mean height is 187 cm with a standard deviation of 15 cm. Looking at the standard deviations, we would say that there is more variation in the class of 40 than the class of 35. The CV for the class of 40 is 8% and for the class of 35 is 8%. It is therefore more reasonable to say that there is equal variation in height between the two groups.

In this session, you have learnt about the coefficient of variation. You have learnt how to compute the coefficient of variation from both grouped and ungrouped data. In addition, you have also learnt about the strengths and weaknesses of the coefficient of variation as well as the uses of the coefficient of variation. I trust that you have grasped the concept of the coefficient of variation well.

Self-Assessment Questions

Exercise 5.5

Compute the coefficient of variation for the sets of data in items 1-3.

1. 18, 22, 48, 45, 90, 93, 65, 62, 28, 75, 15, 30, 35, 80, 82

2. 44, ─ 8, 14, ─ 14, 24, 28, ─ 30, 52, 58, 40, 42, 48, 50, ─ 1

3.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	25
31 – 40	17
21 – 30	12
11 – 20	11

4. The mean age of a class of statistics students is 28 years with a standard deviation of 8 years. In a class quiz, the mean score obtained was 18, with a standard deviation of 6. Which variable shows more dispersion? Age or achievement? Support your answer.

5. One strength of the coefficient of variation as a relative measure of variation is that it
    A. can be used when data is incomplete.
    B. can be used when units of measurement are different.
    C. disregards the actual spread of the scores.
    D. uses few values in a distribution.

6. The coefficient of variation for a set of scores is 75%. It is known that the standard deviation is 15. What is the mean of the distribution?
    A. 15
    B. 20
    C. 75
    D. 100

SESSION 6: QUARTILE DEVIATION

You are welcome to the last session of Unit 5 for the Educational Statistics course. You remember that in Unit 4 Session 6, we learnt how to compute quartiles. We noted that quartiles divide a distribution into four equal parts and that practically, there are three quartiles, Q_1, Q_2, Q_3. In this session, we shall see the application of quartiles in educational practice.

Objectives

By the end of the session, you should be able to

(a) define the quartile deviation,
(b) compute the quartile deviation from ungrouped and grouped data,
(c) explain the strengths of the quartile deviation,
(d) explain the weaknesses of the quartile deviation,
(e) explain the uses of the quartile deviation in teaching and learning.

Now read on...

6.1 Nature of the quartile deviation (QD)

The quartile deviation (QD) is also called the semi-inter quartile range and it depends on quartiles. Quartiles divide distributions into 4 equal parts. Practically there are 3 quartiles.
The QD is half the distance between the first quartile (Q_1) and the third quartile (Q_3).

6.2 Computing the quartile deviation

The quartile deviation can be computed for both raw (ungrouped) data and grouped data. The procedures are described below.

6.2.1 Computing the quartile deviation from raw (ungrouped) data

Three simple steps are involved in computing the quartile deviation from raw data. These steps are:

1. Compute the lower quartile (Q_1).
2. Compute the upper quartile (Q_3).
3. Find the difference between $Q_3$and $Q_1$and divide by 2.

The procedure is summarized in the formula:

$$QD = \frac{Q_3 - Q_1}{2}$$

Let us look at an example.

Given the following set of observations, determine the quartile deviation.

14 22 8 56 46 28 30 17 29 10 60 40 33

Rearranging the scores gives us:

8 10 14 17 22 28 29 30 33 40 46 56 60

$Q_1 = 15.5$ $Q_3 = 43$

I trust that you understand how these values are obtained. Good. If you have forgotten, then go to Unit 4 Session 6 and review the methods.

$$QD = \frac{Q_3 - Q_1}{2} = \frac{43 - 15.5}{2} = \frac{27.5}{2} = 13.75$$

I believe you have understood the computation of the quartile deviation. It is very simple and easy.

Now obtain the quartile deviation for the following set of data.

82 90 66 78 88 72 60 80

What value did you get? 8? 13? 16? 26? Well if you got 8, then your answer is correct. Congratulations. If you did not get 8, crosscheck your calculations.

Now let us look at grouped data.

2.2.2 Computing from grouped data

This method is also very easy. There are three steps.

1. Compute the first quartile, Q_1.
2. Compute the third quartile (Q_3).
3. Find the difference between $Q_3$and $Q_1$and divide by 2.

The procedure is summarized in the formula:

$$QD = \frac{Q_3 - Q_1}{2}$$

Now you already know how to compute the quartiles from grouped data.

To refresh your memory, go to Unit 4, Session 6 and review how to compute the quartiles from grouped data. When you are done, compute the quartile deviation for the data below. Keep your answers and discuss with your colleagues at the FTF.

Given the following frequency distribution table, obtain the quartile deviation.

Table 4.6 Computing the quartile deviation from grouped data

Classes	Freq
46 – 50	8
41 – 45	10
36 – 40	10
31 – 35	8
26 – 30	5
21 – 25	5
16 – 20	4

Pause for a moment. Reflect on the quartile deviation. Close this page. Write down one strength and one weakness of the quartile deviation as a measure of variation in your jotter.

Now open your module and compare what you have written with the following strengths and weaknesses.

6.3 Strengths and weaknesses of the quartile deviation.

The quartile deviation has a number of strengths and weaknesses. These are listed below.

6.3.1 Strengths of the quartile deviation

1. It is easy to compute when the lower and upper quartile values are available.
2. It is not influenced by extreme scores. These scores are usually out of the first or third quartile values.
3. It can be computed when data is missing at the beginning or end of the set of observations after a sequential arrangement of the data.
4. It is the most appropriate measure of variation when there are reasons to believe that the score distribution is skewed.
5. It is easy to interpret. It represents the mean or 'average' of the interquartile range. Thus the larger this mean, the greater the variation within the data set.
6. It is less sensitive to changes in the values of the scores in the distribution. If one score changes, the effect on the quartile deviation is minimal.

6.3.2 Weaknesses of the quartile deviation

1. It does not take into account all the data or scores. It is based only on two values, the first quartile and the third quartile.
2. It ignores the actual spread of all the scores and may therefore give a distorted picture of the variation in the data. This is because different patterns of dispersion may yield the same quartile deviation.
3. It has limited statistical use. It is not used much in inferential statistics.

6.4 Uses of the quartile deviation

1. It is used as the most appropriate measure of variation or variability when there is reason to believe that the distribution is skewed. For a skewed distribution the mean is not equal to the median. It is either greater or smaller by a significant amount.
2. It helps to determine whether a group is homogeneous with respect to a variable or it is heterogeneous when the distribution of scores is skewed. In terms of classroom achievement, the quartile deviation gives us information about whether the class is of the same ability group or a mixed ability group. If the quartile deviation is small, the class is homogeneous, but if the quartile deviation is large the class is heterogeneous.

   As a rough estimate, if scores have a maximum of 50, a quartile deviation of 5 or less is considered small, producing a homogeneous class. On the other hand, for scores more than 50, a quartile deviation of 10 or less gives a homogenous class. However, to be more exact, the coefficient of variation is computed. Where the coefficient of variation (CV) is 33 or less the group is homogeneous, but where it is greater than 33, the group is heterogeneous.

   The nature of the group has implications for teaching and learning. For a heterogeneous class, abilities are mixed. In the teaching, the teacher has to cater for the high-achievers as well as the low-achievers. Attention should also be paid to the middle group. In this case, ability group teaching may be considered. For a homogeneous group, the teacher needs to find the level of performance by examining the measures of central tendency. This will determine whether the group is high-achieving, low-achieving or moderate. In all situations, the teacher needs to adopt a teaching method to suit the class.

   In Session, 5, we learnt about the coefficient of variation. Do you remember the formula?

   It is given as: $CV = \frac{S}{\overline{X}} \times 100$

   This formula is used when the distribution is normal. However, for skewed distributions, this relative measure of variability replaces the standard deviation with the quartile deviation and the mean with the median. The formula then changes to:

$$CV = \frac{QD}{Mdn} \times 100$$

In this session, you have learnt about the quartile deviation. You have learnt how to compute the quartile deviation from both grouped and ungrouped data. In addition, you have also learnt about the strengths and weaknesses of the quartile deviation as well as the uses of the quartile deviation. I trust that you have grasped the concept of the quartile deviation well.

Self-Assessment Questions

Exercise 5.6

1. When a distribution is skewed, the most appropriate measure of variability is the quartile deviation.

    A. True
    B. False.

2. One weakness of the quartile deviation as a measure of spread is that it is……..
    A. affected by extreme scores.
    B. not dependent on all scores.
    C. too easy to interpret.
    D. useful for skewed distributions.

3. The coefficient of variation for a skewed distribution 60%. It is known that the median is 40. What is the quartile deviation of the distribution?
    A. 24
    B. 40
    C. 60
    D. 240

4. Compute the quartile deviation for the following set of scores.

    90 66 78 88 72 60 80 50 70

5. Compute the quartile deviation for the following distribution.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	35
31 – 40	27
21 – 30	12
11 – 20	11

This is a blank sheet for your short notes on:

- Difficult topics if any
- Issues that are not clear.

UNIT 6: MEASURES OF RELATIVE POSITION & NORMAL DISTRIBUTION

Unit Outline

Session 1: Percentiles and percentile ranks
Session 2: Standard scores
Session 3: Stanines
Session 4: Nature of normal distribution
Session 5: Features of normal distribution
Session 6: Applications of normal distribution

Congratulations! You have completed Unit 5. You are now welcome to Unit 6. In Unit 5 you studied the measures of variability or variation. These measures provide single values to describe the degree of variation within a set of observations or scores. The major measures you studied were the range, variance, standard deviation, coefficient of variation and quartile deviation. In this Unit, you will learn about measures of relative position and the normal distribution. The first session discusses percentiles and percentile ranks. This is followed by standard scores and stanines. The Unit will also treat the normal distribution.

Unit Objectives

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

1. Explain the purposes of the measures of relative position;
2. Define percentiles and percentile ranks and compute them;
3. Define stanines and compute them;
4. Describe the features of the normal distribution;
5. Apply the normal distribution in solving problems.

SESSION 1: PERCENTILES AND PERCENTILE RANKS

You are welcome to the first session of Unit 6 for the Educational Statistics course. As noted in Unit 5, the measures of variation tell us the degree of individual differences in a set of scores and in the classroom the teacher uses the measures to find out whether a group is homogeneous or heterogeneous. In this session, we shall learn about measures of relative position. These measures describe an individual's position within a group. We shall start with percentiles and percentile ranks.

Objectives

By the end of the session, you should be able to

(a) describe the nature of a percentile,
(b) compute percentiles from a given data,
(c) describe the nature of percentile ranks,
(d) compute percentile ranks from a given data.

Now read on...

PERCENTILES AND PERCENTILE RANKS

Percentiles and percentile ranks are measures of relative position. They indicate an individual's position on a scale of 100 with respect to the group he/she belongs. Percentiles are scores from a transformed scale of 100 and percentile ranks are positions from this scale of 100.

1.1 Nature of percentiles

Percentiles, like quartiles, divide a distribution into parts. Here they divide a distribution into 100 equal parts. There are 99 percentiles (P_1, P_2, P_3, , P_{99}) that divide a distribution into 100 equal parts. Percentiles are points in a distribution below which a given percent, P, of the cases lie. For example if the 30^{th} percentile is 48, this means that a person who scores 48, will have 30% of the scores lying below 48. Percentiles are individual scores.

The symbol, P is used to denote a percentile. For example 40^{th} percentile will be written as P_{40}, and 70^{th} percentile P_{70}. As noted above, percentiles are individual scores so percentiles are usually written with the scores attached as:

P_{40} = 60. This means 60 is the score below which 40% of the scores lie in a specific group after the scores have been arranged sequentially. This means that a student who obtains a score of 60 has done better than 40% of the members in the specific group.

P_{75} = 50. This means 50 is the score below which 75% of the scores lie in a specific group after the scores have been arranged sequentially. This means that a student who obtains a score of 50 has done better than 75% of the members in the specific group.

P_{90} = 72. This means 72 is the score below which 90% of the scores lie in a specific group after the scores have been arranged sequentially. This means that a student who obtains a score of 72 has done better than 90% of the members in the specific group.

Percentiles are group specific. A score in one group may be a different percentile in another group. For example, in Statistics Quiz 1, a student with a score of 15 out of 20 may be at P_{90} in the Arts class but the same score may put the student at P_{85} in the Science class.

Some percentiles have special names. P_{50} is the same as the median. P_{25} is the first quartile and P_{75} is the third quartile.

1.2 Computing percentiles

Percentiles can be computed from raw (ungrouped) data, ogives and grouped data.
Computing percentiles from raw (ungrouped) data is similar to using the formula method for the Quartiles.

Pause. Reflect on the quartiles. Can you remember the formula method for calculating the quartiles from raw data? Now close your book and try to write it down in your jotter. After you have written it down, open the module.

Compare what you have written with the following.

To determine Q_1, find the score at the $\frac{1}{4}(N+1)$th position after the scores are sequentially arranged.

To determine Q3, find the score at the $\frac{3}{4}(N+1)$ th position after the scores are sequentially arranged.

Now quartiles are special cases of percentiles. Remember $Q_1 = P_{25}$. $Q_2 = P_{50}$ and $Q_3 = P_{75}$
These formulas are adapted to suit the percentiles. For example:

1. To determine P_{20}, arrange the scores in a sequential order and find the score at the $\frac{20}{100}(N+1)$ th position.
2. To determine P_{45}, arrange the scores in a sequential order and find the score at the $\frac{45}{100}(N+1)$ th position.
3. To determine P_{80}, arrange the scores in a sequential order and find the score at the $\frac{80}{100}(N+1)$ th position.

Now let us try an example.

Find P_{45} in the following set of observations.

45, 60, 85, 25, 62, 58, 56, 90, 36, 40, 45, 66, 80, 30, 50, 55, 48, 32, 42

First arrange the scores in a sequential order. This gives you;

25, 30, 32, 36, 40, 42, 45, 45, 48, 50, 55, 56, 58, 60, 62, 66, 80, 85, 90

Find the $\frac{45}{100}(N+1)$ th position. This gives: $\frac{45}{100}(19+1) = \frac{45}{100}x20 = 9$
Now find the score at the 9th position which gives 48. I hope you have understood this.

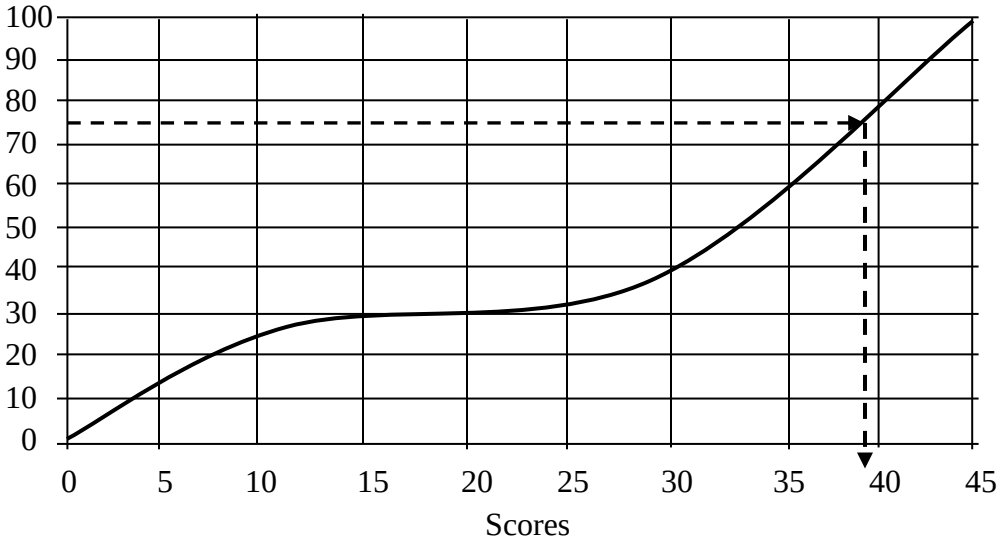
Using the same data, find P_{78}.

What was your answer? Is it 64.4? If No, then check.

P_{78} is at the $15\frac{3}{5}$th position. The score therefore lies between the 15th and 16th scores. We will therefore do an interpolation to obtain P_{78}. The 15th score is 62 and the 16th score is 66. The distance from 62 – 66 is 4. Now, $\frac{3}{5}$ of 4 is 2.4. Add 2.4 to 62 to give you 64.4.

Now let us look at the Ogive method. You remember the Ogive in Unit 2 Session 6? Ok.
To use this method, you first have to draw the ogive and read the value from the horizontal axis.

For example, given the ogive below, find P_{75}.
From the vertical axes, locate 75. Then draw a straight line to the right to touch the curve. From that point, draw a vertical line to touch the horizontal axes. Read the value. In this case it will be 39.



I hope it is easy to you. Try an example.

 Using the same data, find P_{30}.

What was your answer? Is it 15? Good. Well done. Easy, isn't it? If No, then check from the Ogive.

The third method uses the formula which is an adaptation of the median formula. Do you remember the median formula you learnt in Unit 4, Session 4?

Here it is.

$$\text{Mdn} = L_1 + \left[\frac{\frac{N}{2} - cf}{f_{mdn}}\right] i \quad \text{where}$$

L_1 is the class boundary of the median class
N is the total frequency
cf is the cumulative frequency of the class just below the median class
i is the class size/width
f_{mdn} is the frequency of the median class

Using the same formula and adapting it, P_{50} which is the median becomes:

$$\text{Median} = P_{50} = = L_{50} + \left[\frac{\frac{50N}{100} - cf_{50}}{f_{50}}\right] i \quad \text{where}$$

L_{50} is the lower class boundary of the P_{50} class
N is the total frequency
cf_{50} is the cumulative frequency of the class just below the P_{50} class
i is the class size/width
f_{50} is the frequency of the P_{50} class

So in general, if p represents the percentile, then P_p will be written as:

$$P_p = = L_p + \left[\frac{\frac{pN}{100} - cf_p}{f_p}\right] i \quad \text{where}$$

L_p is the lower class boundary of the P_p class
N is the total frequency
cf_p is the cumulative frequency of the class just below the P_p class

i is the class size/width
f_p is the frequency of the P_p class

To obtain the percentile, follow the same steps as in the median.

Step 1. Obtain cumulative frequencies for the frequency distribution.
Step 2. Identify the P_p class. It is the class that will contain the pth percentile. Find the value of $\frac{pN}{100}$. Checking from the cumulative frequency column, find the value that is equal to the position or the smallest value that is greater than the position. The class the value belongs to is the P_p class.
Step 3. Identify the lower class boundary of the P_p class and the class size.
Step 4. Apply the formula.

Now follow the steps above and do the following assignment. Bring it to FTF for discussion.

Compute P_{78} from the data below.

Classes	Freq.
46 – 50	4
41 – 45	6
36 – 40	10
31 – 35	12
26 – 30	8
21 – 25	7
16 – 20	3
Total	50

1.3 Nature of percentile ranks (PR)

Percentile ranks, denoted PR, are based on percentiles. They are the percentage of cases falling below a given point on the measurement scale. It is the position on a scale of 100 to which an individual score lies.

The symbol, PR is used to denote a percentile rank. For example, "PR of 25 = 80" means that if you score 25 in a test, your position on a scale of 100 is 80. This implies that with a score of 25, 80% of the scores are below you which, in other words, mean you have done better than 80% of the class.

PR of 60 = 40. This means that with a score of 60, you are at the 40^{th} position on a scale of 100, implying that 40% of the scores are below you. This means that a student who obtains a score of 60 has done better than 40% of the members in the specific group.

PR of 50 = 75. This means that with a score of 50, you are at the 75^{th} position on a scale of 100, implying that 75% of the scores are below you. This means that a student who obtains a score of 50 has done better than 75% of the members in the specific group.

Just as percentiles are related to specific groups, percentile ranks are also group specific. A PR of a score in one group may be a different PR in another group. For example, in Statistics Quiz 1, a student with PR of 15 being 90 may have PR of 15 to be 85 in another group that took the same test.

1.4 Computing percentile ranks

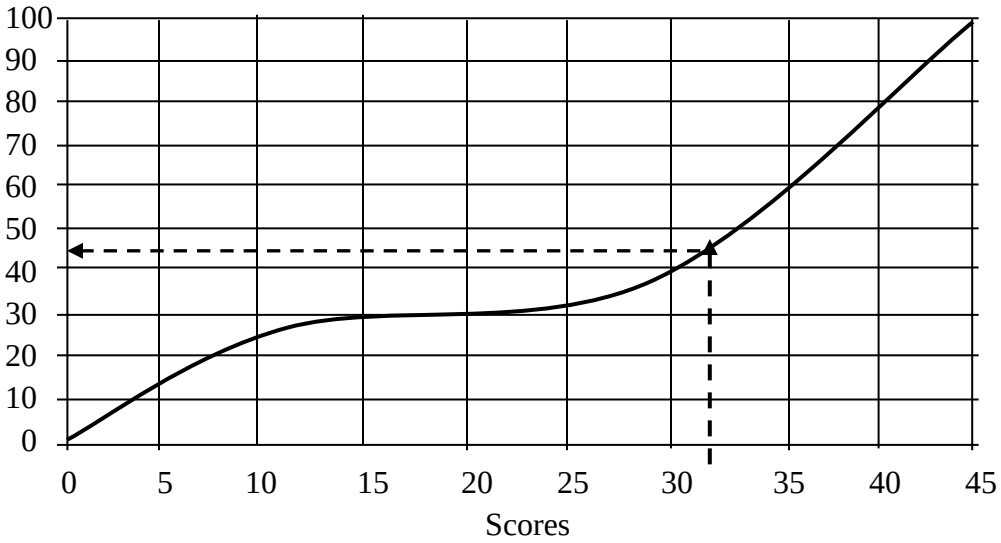
Since percentile ranks are related to percentiles, when you know how to compute percentiles, you can use the following relationship to determine the percentile rank.

1. $P_{40} = 55$ in percentile rank terms becomes, PR of 55 = 40
2. $P_{25} = 38$ in percentile rank terms becomes, PR of 38 = 25
3. $P_{95} = 72$ in percentile rank terms becomes, PR of 72 = 95
4. $P_{75} = 66$ in percentile rank terms becomes, PR of 66 = 75

However, we shall learn how to obtain percentile ranks from Ogives and from a frequency distribution.

To obtain the PR of a score of 32, for example, locate 32 on the horizontal axis. Then draw a vertical line to meet the curve. Draw a straight horizontal line from that point to meet the vertical axis. Read the value on the vertical axis.

The example is shown below.



PR of 32 is 44. This can be written also as $P_{44} = 32$

I hope it is easy to you. Try an example.

From the ogive, find PR of 20.
What did you get?

Well the answer is 30. I hope you got it right.

Now let us look at the grouped data method.

The procedure is as follows.

1. Obtain the cumulative frequencies.
2. Determine the class of the score in question.
3. Obtain the upper class boundary of the class the score belongs to.
4. Apply the following formula.

$$\text{PR of X} = \left[\frac{cf - \frac{f}{i}(X_u - X)}{N} \right] \times 100 \quad \text{where}$$

X is the score

X_u is the upper class boundary of the class X belongs to

N is the total frequency

Cf is the cumulative frequency of the class containing X

i is the class size/width

f is the frequency of the class containing X

Let us work an example:

What is the PR of the score 38 in the distribution below?

Table 5.1 Computing percentile ranks.

Classes	Class Mark X	Frequency f	cf
46-50	48	4	50
41-45	43	6	46
36-40	38	10	40
31-35	33	12	30
26-30	28	8	18
21-25	23	7	10
16-20	18	3	3
Total		50	

38 belongs to 36 – 40 class and the upper class boundary is 40.5.

Applying the formula:

$$\text{PR of } 38 = \left[\frac{cf - \frac{f}{i}(X_u - X)}{N}\right] \times 100 = \left[\frac{40 - \frac{10}{5}(40.5 - 38)}{50}\right] \times 100 = \left[\frac{40 - 2(2.5)}{50}\right] \times 100$$

$$= (40 - 5) \times 2 = 70$$

PR of 38 = 70. This means that if a student scores 38, he/she has done better than 70% of those in the class.

I believe that you have understood percentile ranks now very well.

Now obtain the percentile ranks of the scores 18, 24, and 44 from the frequency distribution table above. Bring your answers to FTF for discussion.

In this session, you have learnt about one of the measures of relative position. Percentiles and percentile ranks go together. While percentiles are individual scores, percentile ranks are positions of the individual scores on a scale of 100. Now you can compute percentiles and percentile ranks and interpret the results.

Self-Assessment Questions

Exercise 6.1

1. Emelia's score in a final examination is at the 65th percentile of the scores in the class. Emelia's score lies

   A. above the third quartile.
   B. at the median.
   C. below the first quartile.
   D. between the median and the third quartile.

2. Martin's percentile rank in an end-of-year examination was 20. His actual examination score was 60. This information means that he performed **worse** than _____ of the student's in the class.

   A. 20%
   B. 40%
   C. 60%
   D. 80%

3. What is the percentile rank of the score of 78 in the distribution below?

90 66 78 88 72 60 80 50 70

4. Compute P_{70} for the distribution below.

Classes	Frequency
61 - 70	15
51 – 60	20
41 – 50	35
31 – 40	27
21 – 30	12
11 – 20	11

5. What is the percentile rank for a score of 65 in the distribution above?

SESSION 2: STANDARD SCORES (Z, T)

You are welcome to the second session of Unit 6 for the Educational Statistics course. In Session 1 you learnt about percentiles and percentile ranks which are measures of relative position. These measures tell us how an individual's performance compares with others in the same group. You have learnt how to compute percentiles and percentile ranks and how to interpret the results. In this session, we shall learn about another set of the measures of relative position. These are known as standard scores.

Objectives

By the end of the session, you should be able to

(a) define a standard score,
(b) compute the z-score,
(c) compute the T-score,
(d) explain the importance of standard scores in educational practice.

Now read on...

2.1 Nature of standard scores

Standard scores indicate the number of standard deviation units an individual score is above or below the mean of each group. They represent an individual score that has been transformed into a common standard using the mean and the standard deviation. The most popular standard scores are Z and T.

2.2 The Z standard score

The Z standard score is often referred to as z-score. It is used to transform raw scores into a common standard using the mean and the standard deviation of a set of observations. The raw scores are transformed to a mean of 0 and a standard deviation of 1.

The z-score may either be positive or negative and ranges from −4 to +4. Positive values indicate that performance is above the mean or above average and negative scores mean performance is below the mean or below average.

2.2.1 Computing z-scores

Z-scores are obtained by using the formula

$$Z = \frac{X - \overline{X}}{s}$$

Let us do some examples.

1. A student had a Z-score of 2.5. The mean for the class was 60 with a standard deviation of 4.0. What was the student's observed score?

$$Z = \frac{X - \overline{X}}{s} \longrightarrow 2.5 = \frac{X - 60}{4} \longrightarrow 10 = X - 60 \longrightarrow X = 10 + 60 = 70$$

2 A student obtained a raw score of 70 in an examination. If the raw score gives her a Z-score of 3.5, what would be the class mean if it is known that the standard deviation is 5.0?

$$Z = \frac{X - \overline{X}}{s} \rightarrow 3.5 = \frac{70 - \overline{X}}{5} \rightarrow 17.5 = 70 - \overline{X} \rightarrow \overline{X} = 70 - 17.5 = 52.5$$

I hope you have followed the examples above.

Now I want you to try the following.

1. Scores in a Quiz had a mean of 45 and standard deviation of 14. Michael's standard score on the examination was 2.5. What was his actual examination score?

What answer did you get? Well the answer is 80. Did you get it right? Well done. If you got it wrong check your calculations again.

$$Z = \frac{X - \overline{X}}{s} \longrightarrow 2.5 = \frac{X - 45}{14} \longrightarrow 35 = X - 45 \longrightarrow X = 35+45 = 80$$

2. In a Statistics test, Alice has a score of 75. The class mean was 60 with a standard deviation of 8. How would you describe her performance relative to the rest of the class?

What is your answer? Provide support for it. Write the response in the box.

The answer will be discussed at the face to face.

2.3 The T standard score

The T standard score is based on the z-score. It is also used to transform raw scores into a common standard using the mean and the standard deviation of a set of observations. The raw scores are transformed to a mean of 50 and a standard deviation of 10. It is based on the formula: T = 50 + 10Z, where mean is 50 and standard deviation is 10.

T-scores are easier to use because they do not deal with negative values as the z-score does. In addition, it does not handle fractions in the computations. The T-score is always positive and range from 10 to 90. Values above 50 indicate that performance is above the mean or above average and values below 50 indicate that performance is below the mean or below average.

2.3.1 Computing T-scores

T-scores are obtained by using the formula

$T = 50 + 10Z$ This can be re-written as: $T = 50 + 10\left(\frac{X - \overline{X}}{s}\right)$

Let us do some examples.

1. A student had 70 in a test. The mean for the class was 60 with a standard deviation of 4.0. What was the student's T-score?

   $T = 50 + 10Z$ The Z-score is $\frac{70 - 60}{4} = 2.5$

   The T-score becomes 50 + 10(2.5) = 50 + 25 = 75. This shows that the student's performance is far above average when compared with the mean of 50.

2. A student obtained a raw score of 68 in an examination. If the raw score gives her a T-score of 90, with the class standard deviation known to be 5, what would be the class mean?

   We need to know the Z-score first.

   $T = 50 + 10Z \longrightarrow 90 = 50 + 10Z \longrightarrow 40 = 10Z \longrightarrow Z = 4$

   $4 = \left(\frac{68 - \overline{X}}{5}\right) \longrightarrow 20 = 68 - \overline{X} \longrightarrow \overline{X} = 68 - 20 = 48$

I hope you have followed the examples above.

Now I want you to try the following.

Scores in a Quiz had a mean of 45. Adam had 80 which yielded a T-score of 75. What was the class standard deviation?

What answer did you get? Well the answer is 14. Did you get it right? Well done. If you got it wrong check your calculations again.

$T = 50 + 10\left(\frac{80 - 45}{s}\right)$ $75 = 50 + 10\left(\frac{35}{s}\right)$ $25 = 10\left(\frac{35}{s}\right)$ $25s = 350$ $s = 14$

3 In a Statistics test, Joana has a T-score of 88. The class mean was 52 with a standard deviation of 8. What was her observed score?

What answer did you get? Well the answer is 82.4 which is rounded to 82. I hope you got it right.

2.4. Uses of standard scores, percentiles and percentile ranks

The classroom teacher can use the standard scores to improve teaching and learning in three ways. These are described below.

1. It helps the teacher to know an individual's position in relation to the rest of the class. A student with a Z-score of 3.2 is performing far above average. A student with a T-score of 25 is performing far below average. A student at P_{15} is performing at a low level in the group.
2. It enables the teacher to compare student's performances in different subjects to know individual strengths and weaknesses.
   For example Salome obtained the following scores in Mathematics and Social Studies.

	Mathematics	Social Studies
Class Mean	50	60
Class standard deviation	2.5	4.0
Salome's obtained score	55	55
Salome's Z score	2.0	−1.25
Salome's Percentile	P_{80}	P_{20}

   Salome has done better in Mathematics ($Z = 2.0$, P_{80}) than Social Studies ($Z = -1.25$, P_{20}), considering the class performance.
3. It helps the teacher to guide and counsel the student to choose the correct course for a future career and vocation.

For example which of the following subjects will offer the best career for George's ability?

	English	Maths	Pre-Tech
Class Mean	80	70	75
Class standard deviation	6.0	2.0	4.0
George's obtained score	85	76	80
George's Z score	0.42	3.0	1.25
George's Percentile	P_{40}	P_{90}	P_{60}

George is more likely to succeed in a Maths-related course because he had the highest Z-score and highest percentile in Mathematics.

In this session, you have learnt about standard scores which are measures of relative position. T-scores depend on Z-scores and both of them tell us about a student's performance in relation to the rest of the group. You have learnt how to compute the standard scores and how to interpret the results. You have also learnt how the classroom teacher can use these standard scores in teaching and learning.

Self-Assessment Questions

Exercise 6.2

1. Scores on an WASSCE Social Studies paper had a mean of 46. Joana obtained a score of 70, giving her a standard score of 3.0. What was the standard deviation of the scores?

    A. 3
    B. 8
    C. 24
    D. 72

2. Scores on the College of Education entrance examination had a mean of 50 and standard deviation of 5. Martin's standard score on the examination was Z = 2.0. What was Martin's actual examination score?

    A. 40
    B. 50
    C. 52
    D. 60

3. An elementary statistics course has end-of-semester scores normally distributed with mean 17. If George's Z standard score is −3.5 and his score in the Quiz is 10, what is the standard deviation of the scores?

    A. −2.0
    B. 2.0
    C. 3.5
    D. 7.0

4. Given that Marian had a T-score of 100, and that the group's mean is 60. If her actual score is 75, what is the standard deviation of the group?

    A. 3
    B. 5
    C. 50
    D. It cannot be determined.

5. Patrick's T-score in a quiz was 75. The mean class performance was 60 with a standard deviation of 12. What was Patrick's actual score in the quiz?

    A. 25

B. 30
C. 90
D. It cannot be determined.

6. Scores on the BECE Mathematics paper follow the normal distribution with mean 45 and standard deviation of 15. Martha's z standard score on the examination was 2.0. What was her actual examination score?

A. 15
B. 30
C. 65
D. 75

SESSION 3: STANINES

You are welcome to the third session of Unit 6 for the Educational Statistics course. In Session 2 you learnt about standard scores, Z-scores and T-scores which are measures of relative position. These measures tell us how an individual's performance compares with others in the same group. You have learnt how to compute Z and T scores and how to interpret the results. You have also learnt the educational importance of the Z and T scores. In this session, we shall learn about Stanines.

Objectives

By the end of the session, you should be able to

(a) define a stanine,

(b) compute a stanine,

(c) explain the importance of stanines in educational practice.

3.1 Nature of stanine

The term, Stanine, is a kind of an acronym. It comes from Standard Nine. Stanine is a method of scaling test scores on a nine-point (1, 2, 3, 4, 5, 6, 7, 8, 9) standard scale with a mean of 5 and a standard deviation of 2. In general, stanines 1 to 3 are considered below average, stanines 4 to 6 are considered average and stanines 7 to 9 are considered above average.

3.2 Computing Stanines

Stanines are generally computed from raw data. We shall learn two methods of computing stanines from raw data.

The first method involves three steps.

1. Rank the scores from the lowest to the highest.
2. Obtain the number of scores for each stanine category using the following percentages.

Result Ranking	4%	7%	12%	17%	20%	17%	12%	7%	4%
Stanine	1	2	3	4	5	6	7	8	9

3. Determine approximate cut-off points by finding the mean of adjacent scores.

Let us work through an example.

The following scores were obtained in an examination. Determine the stanine groups.

48, 36, 52, 38, 42, 50, 35, 60, 55, 47, 52, 58, 72, 54, 48, 42, 45, 56, 62, 70, 74, 75, 78, 65, 46

1. Arranging the scores in an ascending order gives:

35, 36, 38, 42, 42, 45, 46, 47, 48, 48, 50, 52, 52, 54, 55, 56, 58, 60, 62, 65, 70, 72, 74, 75, 78

2. Obtaining the number of scores for each Stanine category gives:

Result Ranking	4%	7%	12%	17%	20%	17%	12%	7%	4%
No. of scores	1	1.75	3	4.25	5	4.25	3	1.75	1
Real no. of scores	1	2	3	4	5	4	3	2	1
Stanine	1	2	3	4	5	6	7	8	9

In calculating the number of scores, 4% of 25 scores gives $\frac{4}{100} \text{ x } 25 = 1$, 7% of 25 scores gives,

$\frac{7}{100} \text{ x } 25 = 1.75$

The real number of scores are approximations where the number of scores are made whole numbers (discrete) to remove the decimal fractions. The total gives 25.

The corresponding scores, numbers and stanine are below.

35,	36, 38,	42, 42, 45,	46, 47, 48, 48,	50, 52, 52, 54, 55,	56, 58, 60, 62,	65, 70, 72,	74, 75,	78
1	2	3	4	5	4	3	2	1
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

For this group of scores, the cut-off points for the stanines are as follows:

Stanine	Score Range
1	0 – 35.5
2	35.5 – 40.5
3	40.5 – 45.5
4	45.5 – 49.5
5	49.5 – 55.5
6	55.5 – 63.5
7	63.5 – 73.5
8	73.5 – 76.5
9	Above 76.5

Note that the cut-off points are exact boundaries. These boundaries can be used because the scores are generally whole numbers (discrete).

The second method is based on the normal distribution and produces more exact cut-off points. It uses the Z-score to determine the cut-off points. The following table has been developed for use in obtaining the cut-off points.

Stanine	Z-Score Range
1	Below –1.75
2	–1.75 to –1.25

3	–1.25 to –0.75
4	–0.75 to –0.25
5	–0.25 to 0.25
6	0.25 to 0.75
7	0.75 to 1.25
8	1.25 to 1.75
9	Above 1.75

Let us work an example.

In a final district-wide examination, a mean of 52 was obtained with a standard deviation of 8. What would be the cut-off points for the stanine score ranges.

$Z = \frac{X - \overline{X}}{s}$. Solving for X gives: $X = Zs + \overline{X}$

$X = Zs + \overline{X} = \text{-}1.75(8) + 52 = -14 + 52 = 38$

$X = Zs + \overline{X} = \text{-}1.25(8) + 52 = -10 + 52 = 42$

$X = Zs + \overline{X} = \text{-}0.75(8) + 52 = -6 + 52 = 46$

$X = Zs + \overline{X} = \text{-}0.25(8) + 52 = -2 + 52 = 50$

$X = Zs + \overline{X} = 0.25(8) + 52 = 2 + 52 = 54$

$X = Zs + \overline{X} = 0.75(8) + 52 = 6 + 52 = 58$

$X = Zs + \overline{X} = 1.25(8) + 52 = 10 + 52 = 62$

$X = Zs + \overline{X} = 1.75(8) + 52 = 14 + 52 = 66$

The score ranges for the Stanines are below.

Stanine	Obtained Score Range
1	Below 38
2	38-42
3	43-46
4	47-50
5	51-54
6	55-58
7	59-62
8	63-66
9	Above 66

I do hope that you have understood the process.

Now I want you to try the following assignment and bring it to FTF for discussion.

In a certification examination, a mean of 55 was obtained with a standard deviation of 10. What are the scores ranges for the stanines. If a student obtained a score of 62, what stanine would the student get?

3.3 Uses of stanine in educational practice

Stanines are used mainly to identify an individual's level of performance in relation to a group. Generally, stanines 1, 2, 3 represent performance below average, 4, 5, 6 represent performance on the average and 7, 8, 9 represent performance above average. Using the stanine scoring system you can see, at a glance, whether a student is excelling or whether he or she may need extra help.

In this session, you have learnt about stanines which are measures of relative position. Stanines use the single digit numbers, 1 – 9 to denote levels of performance. You have learnt how to compute stanines and how to interpret the results. You have also learnt how the classroom teacher can use stanines in teaching and learning.

Self-Assessment Questions

Exercise 6.3

1. The following scores were obtained in an examination.

48, 72, 85, 45, 56, 58, 63, 68, 59, 75, 66, 64, 80, 75, 60, 68, 67, 58, 59, 54, 55, 54, 52, 51, 68
50, 49, 76, 81, 47

What stanine would the students who obtain the following scores get?

   i. 52
   ii. 68
   iii. 81

2. In a nation-wide examination in Social Studies, a mean of 56 was obtained with a standard deviation of 12. What stanine would be given to students who had the following marks?

   i. 50
   ii. 66
   iii. 76
   iv. 48
   v. 80

SESSION 4: NATURE OF NORMAL DISTRIBUTION

You are welcome to the fourth session of Unit 6 for the Educational Statistics course. In Session 3 you learnt about stanines which are measures of relative position. Stanines give us information about a person's level of performance using the single digit values of 1 – 9. Stanine values of 1, 2, and 3 show that performance is below average, stanine values of 4, 5, 6 show average performance and stanine values of 7, 8, 9 show above average performance. Stanines, like Z-scores, are based on the normal distribution and in this session, we shall learn more about the nature of the normal distribution.

Objectives

By the end of the session, you should be able to

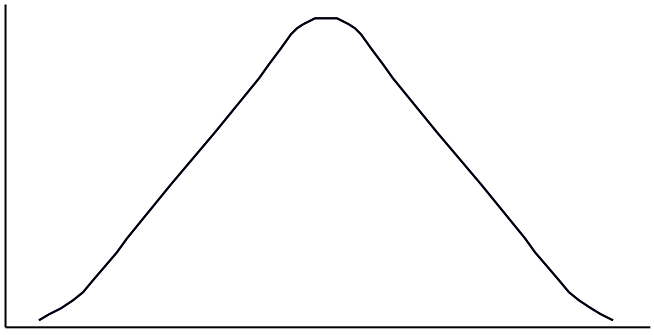
(a) describe a normal distribution,

(b) determine areas under the normal curve.

4.1 Description of the normal distribution

The normal distribution is regarded as the foundation of all statistical distributions. It can be referred to as the 'mother' of all distributions. It is often regarded as the most important of all the statistical distributions. It is often described as a symmetrical bell-shaped curve with a larger proportion of the scores located around the middle.

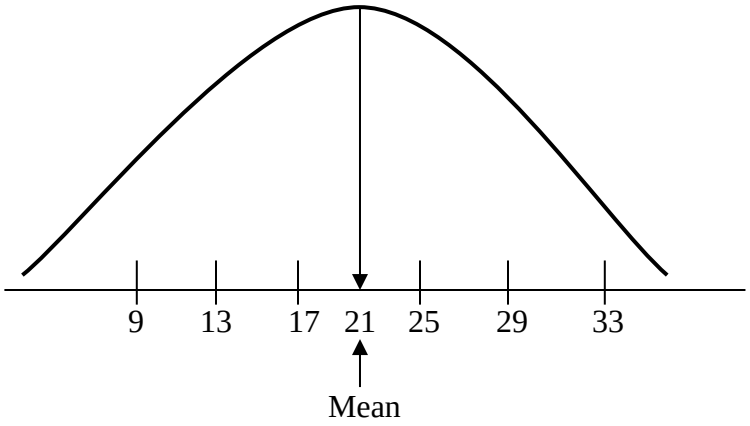
The shape of the normal distribution is shown by the normal curve below.



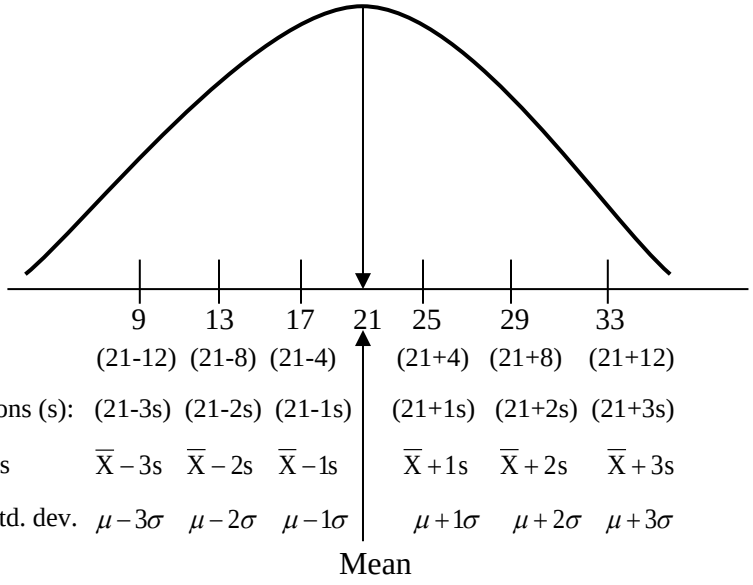
The horizontal axis is measured in terms of standard deviation units. The values decrease to the left and increase to the right from the centre, where the mean is located. The vertical axis is referred to as the **ordinate**.

This is illustrated below.

Suppose that the scores of a group of students have a normal distribution with a mean of 21 and a standard deviation of 4. The distribution takes the form below.



Note that the numbers are far from each other by the size of the standard deviation. The diagram can be shown also as:



Note the differences in the symbols. The sample mean is written as $\overline{X}$ while the population mean is written as μ. The sample standard deviation is denoted s while the population standard deviation is denoted as σ.

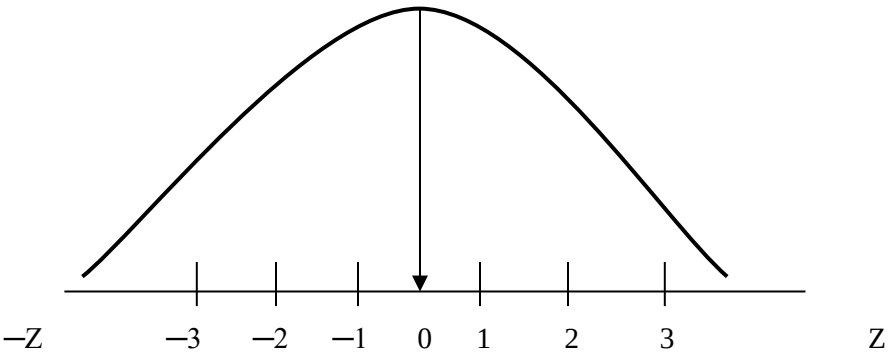
Now pause for a moment. Do you remember the features of the Z-score? Write one of them down in the space provided.

Now go back to Session 2 and check what you have written.

Well, one of the features is that it has a mean of 0. Thus if all raw scores are converted to Z scores, the mean of the Z scores is zero.

When the values of a normal distribution have been converted to standard z-scores, a standard normal curve is obtained. The standard normal curve has a mean of 0 and a standard deviation/variance of 1.

The Z scores are sometimes used to label the horizontal axis. This is illustrated below.



Note that the values −3, −2, −1 do not mean that the standard deviations are negative. What they mean is that they are to the left of the mean of 0. In theory, the values range from −4 to +4.

4.2 Using symbols

The variables like age, height, and scores in an examination which have a normal distribution are often described using a specific symbol. This is illustrated below.

A variable which is distributed normally has the symbol, $X \sim N(\mu, \sigma^2)$ where μ, is the mean and σ^2, the variance. This is read as 'the variable, X, is normally distributed with a given mean and a given variance. For example, the scores (X) in a given entrance examination have a mean of 54 and a variance of 16. This will be written in symbolic form as: X – N(54, 16).

A standard normally distributed variable is written in symbolic form as: X – N(0, 1).

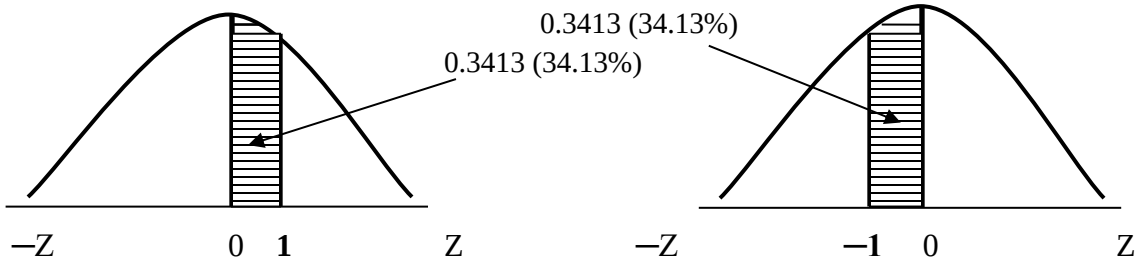
4.3 Areas under the normal curve

There is a special relationship between the standard deviation and the normal curve. This relationship relates to the areas under the normal curve. These areas are obtained from Appendix A. Please refer to the Appendix. The relationships are illustrated below.

4.3.1 Relationship 1

If one moves 1 standard deviation or 1 standard deviation unit from the mean to the right, the area covered is 34.13% (0.3413). Likewise if one moves 1 standard deviation or 1 standard deviation unit to the left, the area covered is 34.13% (0.3413).

The relationship is illustrated below.



This is written as: $\mu + 1\sigma = 0.3413$ $\mu - 1\sigma = 0.3413$

$\overline{X} + 1s = 0.3413$ $\overline{X} - 1s = 0.3413$

4.3.2 Relationship 2

If one moves 1 standard deviation or 1 standard deviation unit from the mean both to the left and right, the total area covered is 68.26% (0.6826). (i.e. 34.13% + 34.13%)

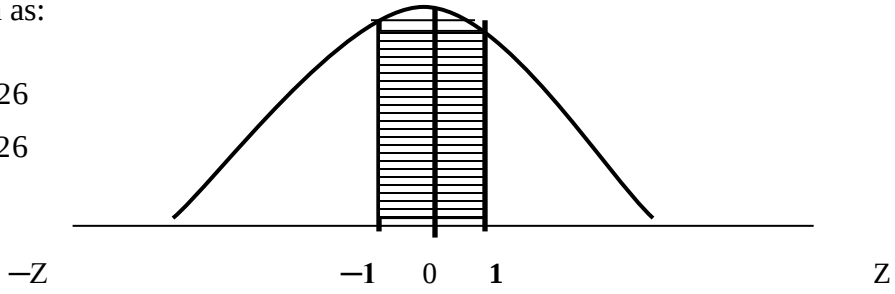
The relationship is illustrated below.

The total area of the shaded portion is 68.26% (0.6826)

This is written as:

$\mu \pm 1\sigma = 0.6826$

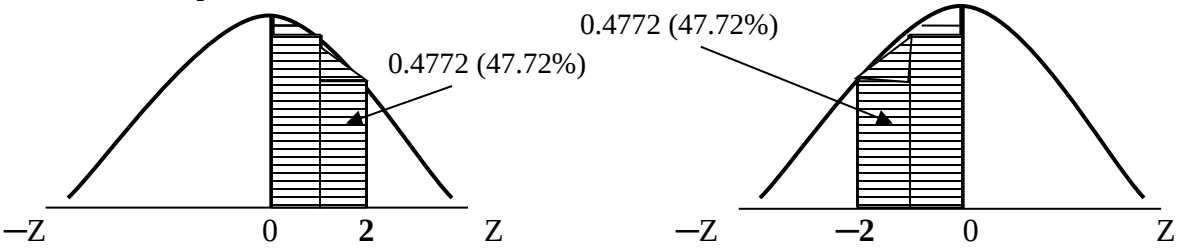
$\overline{X} \pm 1s = 0.6826$



4.3.3 Relationship 3

If one moves 2 standard deviations or 2 standard deviation units from the mean to the right, the area covered is 47.72% (0.4772). Likewise if one moves 2 standard deviations or 2 standard deviation units to the left, the area covered is 47.72% (0.4772).

The relationship is illustrated below.



This is written as: $\mu + 2\sigma = 0.4772$ $\mu - 2\sigma = 0.4772$

$\overline{X} + 2s = 0.4772$ $\overline{X} - 2s = 0.4772$

4.3.4 Relationship 4

If one moves 2 standard deviations or 2 standard deviation units from the mean both to the left and right, the total area covered is 95.44% (0.9544) (i.e. 47.72% + 47.72%)

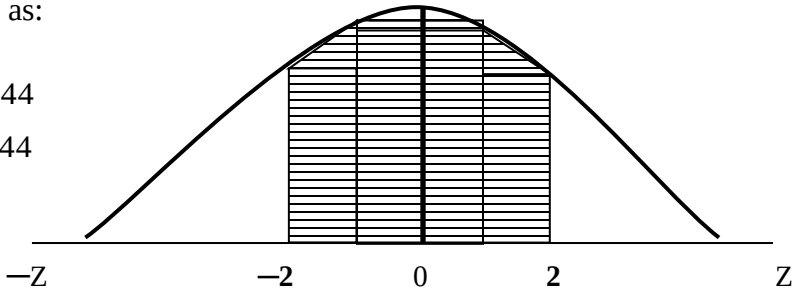
The relationship is illustrated below.

The total area of the shaded portion is 95.44% (0.9544)

This is written as:

$\mu \pm 2\sigma = 0.9544$

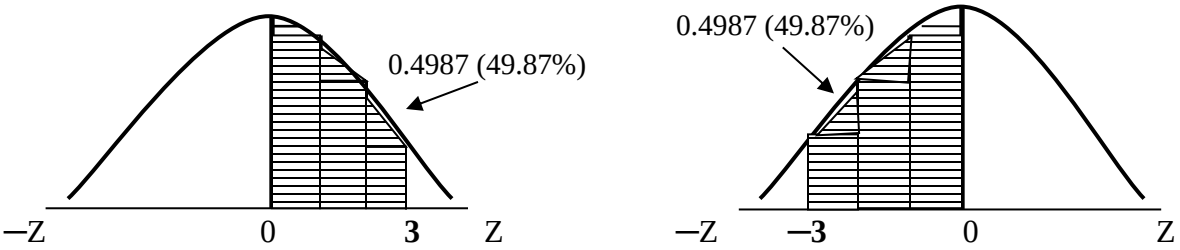
$\overline{X} \pm 2s = 0.9544$



4.3.5 Relationship 5

If one moves 3 standard deviations or 3 standard deviation units from the mean to the right, the area covered is 49.87% (0.4987). Likewise if one moves 3 standard deviations or 3 standard deviation units to the left, the area covered is 49.87% (0.4987).

The relationship is illustrated below.



This is written as: $\mu + 3\sigma = 0.4987$ $\mu - 3\sigma = 0.4987$

$\overline{X} + 3s = 0.4987$ $\overline{X} - 3s = 0.4987$

4.3.6 Relationship 6

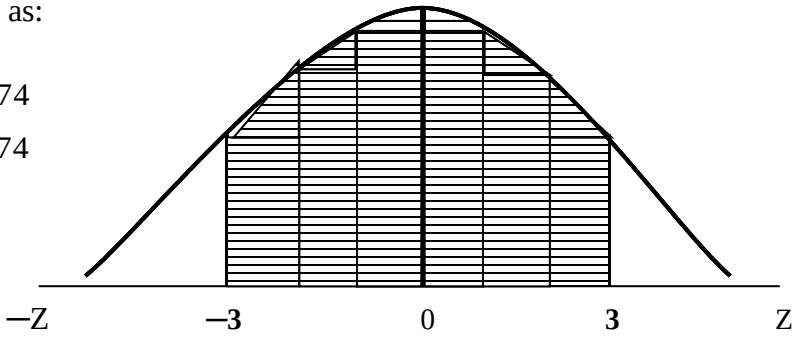
If one moves 3 standard deviations or 3 standard deviation units from the mean both to the left and right, the total area covered is 99.74% (0.9974). (i.e. 49.87% + 49.87%)

The relationship is illustrated below.

The total area of the shaded portion is 99.74% (0.9974)
This is written as:

$\mu \pm 3\sigma = 0.9974$

$\overline{X} \pm 3s = 0.9974$

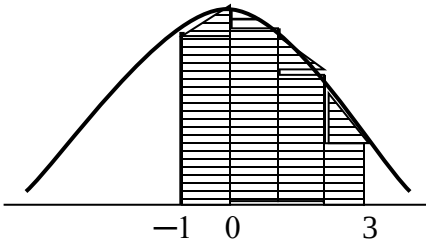


In this session, you have learnt about the nature of the normal distribution. This distribution is regarded as the 'mother' of all statistical distributions. It is represented by the normal curve which has a horizontal axis to it. The horizontal axis is marked in terms of standard deviation units. The mean is at the centre of the normal curve. Relationships exist between the standard deviation units and the areas covered under the normal curve. An area of 68.26% is covered by one standard deviation to the left and right of the mean. An area of 95.44% is covered by two standard deviations to the left and right of the mean and an area of 99.74% is covered by three standard deviations to the left and right of the mean. I trust that you have understood the nature of the normal distribution.

Self-Assessment Questions

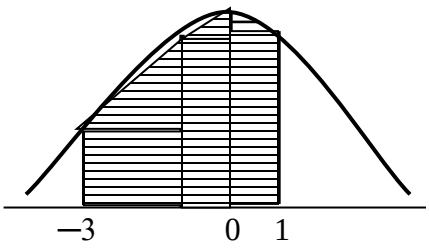
Exercise 6.4

1. What is the total area covered by the shaded portion?



2. What area of the normal curve is covered by $\overline{X} \pm 3s$?

3. In an entrance examination, a mean of 58 was obtained. If the standard deviation was 4, what area under the normal curve is covered between the mean and 62?
4. What is the total area covered by the shaded portion?



5. In a promotion examination, which is N(56, 16), what proportion of the normal curve is covered between 56 and 48?
6. What area of the normal curve is covered by $\overline{X} - 2s$?

SESSION 5: FEATURES OF NORMAL DISTRIBUTION

You are welcome to the fifth session of Unit 6 for the Educational Statistics course. In Session 4 you learnt about the nature of the normal distribution and the normal curve. You also learnt about the relationship between the standard deviation and the areas under the normal curve. In this session, we shall study the major features of the normal distribution and the normal curve.

Objectives

By the end of the session, you should be able to

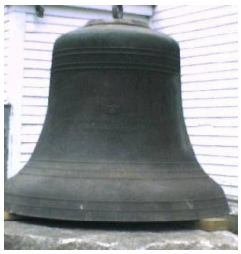
(a) describe the major features of the normal distribution and the normal curve.

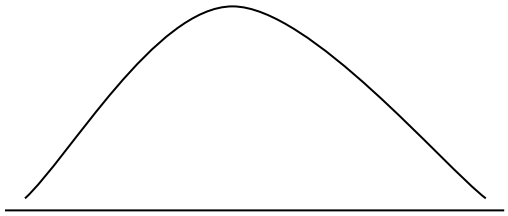
(b) determine further areas under the normal curve.

6.1 Features

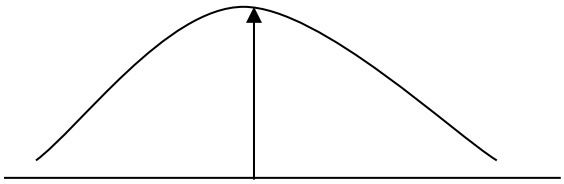
The normal distribution and the normal curve have features that distinguish them from other statistical distributions. These features are described below.

1. **It is a bell-shaped curve**. This shows that the curve is in the form of a bell. You may recall that in the basic schools, some years ago, bells were used to call pupils to school assembly and other functions and also to change lessons. Some schools still use these bells. Some churches also use bells to call members for church services.

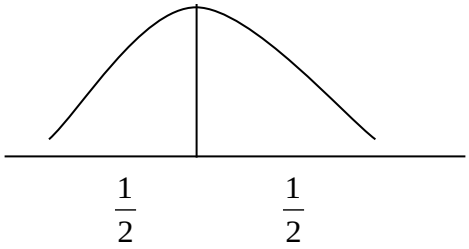




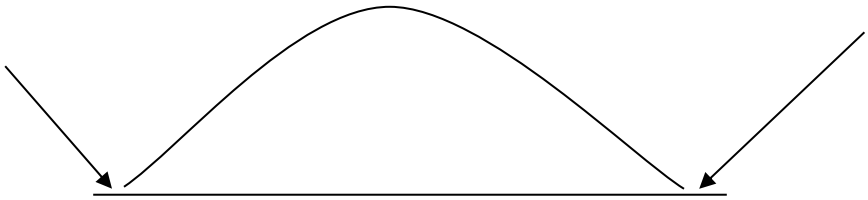
2. It is unimodal. The highest point of the curve from the horizontal axis is the mode. For the normal curve, there is only one highest point. There are other distributions which have more than one mode.



3. It is symmetrical about the mean. The curve is divided into two equal parts at the middle (mean). One part is exactly the same as the other part. If the distribution is folded at the middle, i.e. the mean, the right side will fall exactly on the left side.



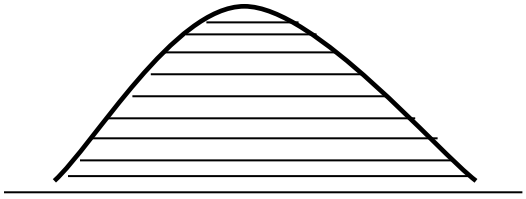
4. It is asymptotic. The normal curve gets closer and closer to the horizontal axis but does not touch it.



Do you notice that the arrows are pointing to an open space?

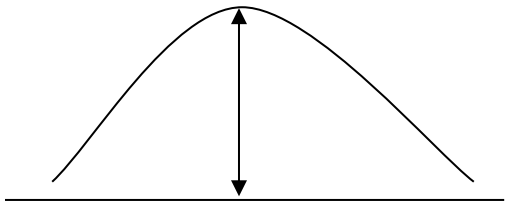
It means that no matter how wide and long the curve goes, it will never touch the horizontal axis. When you draw the normal curve, make sure that it does not touch the horizontal axis. The normal distribution extends from minus infinity ($-\infty$) to plus infinity ($+\infty$). However, for practical purposes, the range -4 to $+4$ is used.

5. The total area under the curve is 1.0. Thus the normal curve is also known as the unit normal curve.



You notice that in Session 4, we found that a relationship exists between the areas under the curve and the standard deviations. The relationships are based on the fact that the total area under the curve is 1.0

6. The mean, mode and median are all equal. For the normal distribution, the mean is equal to the median which is equal to the mode.



Mean
Median
Mode

7. There are infinite normal distribution curves. Each distribution is identified by the value of the mean and the standard deviation. However, when the values are converted to a standardized z, there is only one **standard normal curve**.

6.2 Further areas under the normal curve

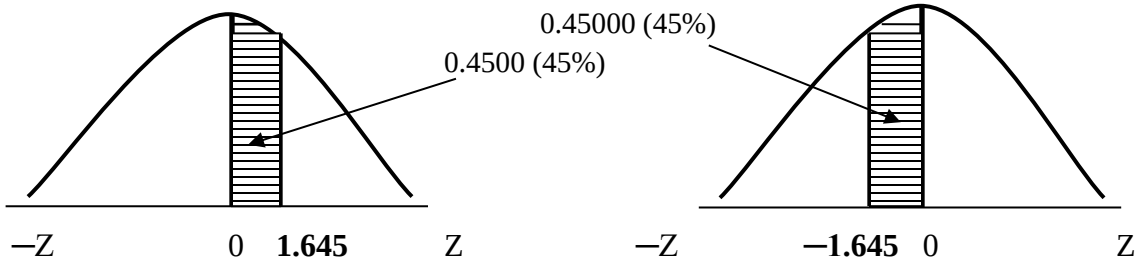
Do you remember that in Session 4 we saw a special relationship between the standard deviation and the normal curve? I hope you remember. How many standard deviation units to the right of the mean produces a covered area of 0.4772? Did you say 2? You are correct.

Now we shall look at it from the proportion of area covered. In Session 4, we looked at areas covered with decimal fractions after the percentages. In statistical analysis, interest is also in particular percentages. These are 90%, 95% and 99%. How many standard deviations away from the mean produce these popular percentages? The answers are found below. The areas are obtained from Appendix A. Please refer to the Appendix.

6.2.1 Relationship 7

If one moves 1.645 standard deviation units from the mean to the right, the area covered is 45% (0.4500). Likewise if one moves 1.645 standard deviation units to the left, the area covered is 45% (0.4500).

The relationship is illustrated below.



This is written as: $\mu + 1.645\sigma = 0.4500$ $\mu - 1.645\sigma = 0.4500$

$\overline{X} + 1.645s = 0.4500$ $\overline{X} - 1.645s = 0.4500$

6.2.2 Relationship 8

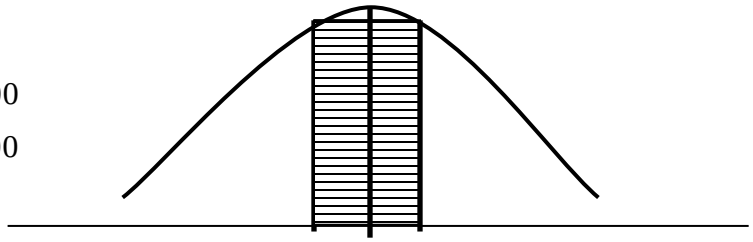
If one moves 1.645 standard deviation units from the mean both to the left and right, the total area covered is 90% (0.9000). (i.e. 45% + 45%)

The relationship is illustrated below.

The total area of the shaded portion is 90% (0.9000)
This is written as:

$\mu \pm 1.645\sigma = 0.9000$

$\overline{X} \pm 1.645s = 0.9000$

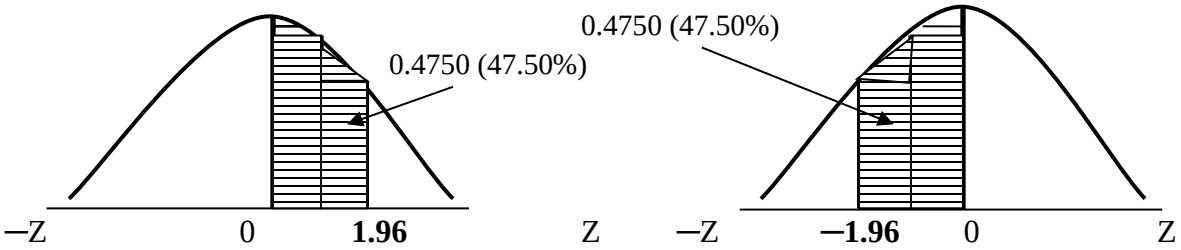


−Z **−1.645** 0 **1.645** Z

5.2.3 Relationship 9

If one moves 1.96 standard deviation units from the mean to the right, the area covered is 47.50% (0.4750). Likewise if one moves 1.96 standard deviation units to the left, the area covered is 47.50% (0.4750).

The relationship is illustrated below.



−Z 0 **1.96** Z −Z **−1.96** 0 Z

This is written as: $\mu + 1.96\sigma = 0.4750$ $\mu - 1.96\sigma = 0.4750$

$\overline{X} + 1.96s = 0.4750$ $\overline{X} - 1.96s = 0.4750$

5.2.4 Relationship 10

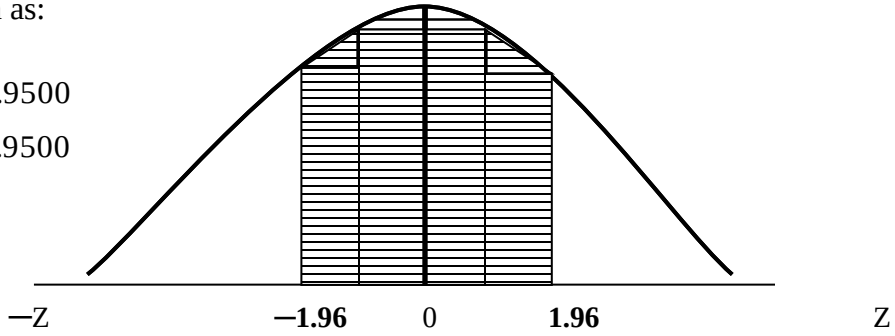
If one moves 1.96 standard deviation units from the mean both to the left and right, the total area covered is 95% (0.9500). (i.e. 47.50% +47.50%)

The relationship is illustrated below.

The total area of the shaded portion is 95% (0.9500)
This is written as:

$\mu \pm 1.96\sigma = 0.9500$

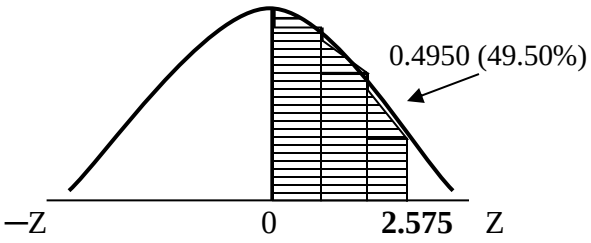
$\overline{X} \pm 1.96s = 0.9500$

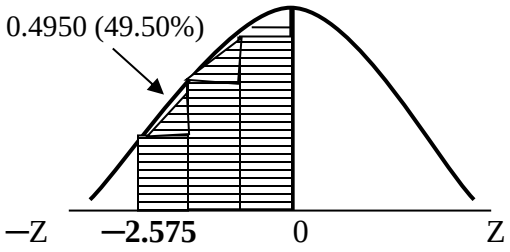


5.2.5 Relationship 11

If one moves 2.575 standard deviation units from the mean to the right, the area covered is 49.50% (0.4950). Likewise if one moves 2.575 standard deviation units to the left, the area covered is 49.50% (0.4950).

The relationship is illustrated below.





This is written as: $\mu + 2.575\sigma = 0.4950$ $\quad$ $\mu - 2.575\sigma = 0.4950$

$\overline{X} + 2.575s = 0.4950$ $\quad$ $\overline{X} - 2.575s = 0.4950$

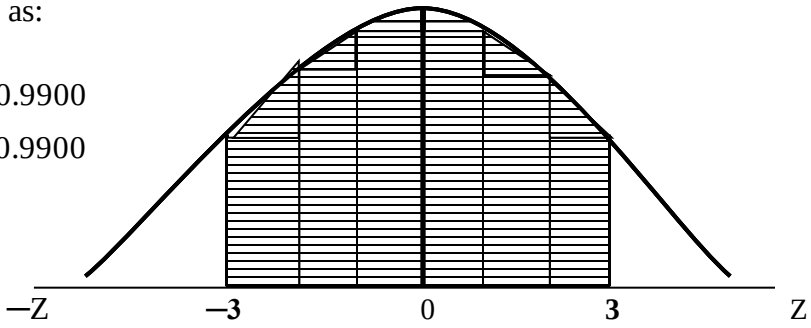
5.2.6 Relationship 12

If one moves 2.575 standard deviation units from the mean both to the left and right, the total area covered is 99% (0.9900). (i.e. 49.50% + 49.50%)

The relationship is illustrated below.
The total area of the shaded portion is 99% (0.9900)
This is written as:

$\mu \pm 2.575\sigma = 0.9900$

$\overline{X} \pm 2.575s = 0.9900$

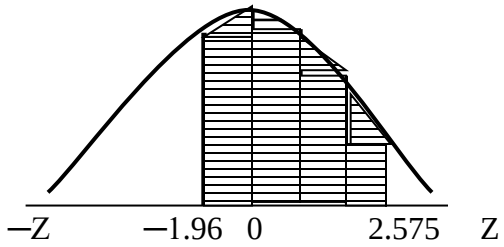


In this session, you have learnt about the features of the normal distribution and the normal curve. We have seen that the normal distribution and the normal curve are bell-shaped, symmetrical about the mean, and have the mean, median and the mode equal. An area of 90% is covered by 1.645 standard deviation units to the left and right of the mean. An area of 95% is covered by 1.96 standard deviation units to the left and right of the mean and an area of 99% is covered by 2.575 standard deviation units to the left and right of the mean. I trust that you have understood clearly the features of the normal distribution.

Self-Assessment Questions

Exercise 6.5

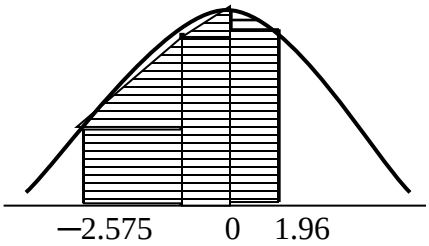
1. What is the total area covered by the shaded portion?



2. What area of the normal curve is covered by $\overline{X} \pm 1.96s$?

3. In an entrance examination, a mean of 54 was obtained. If the standard deviation was 1.96, between the mean and what value to the right of the mean would be covered by 47.50% under the normal curve?

4. What is the total area covered by the shaded portion?



5. What area of the normal curve is covered by $\overline{X} - 1.96s$?

SESSION 6: APPLICATIONS OF NORMAL DISTRIBUTION

You are welcome to the last session of Unit 6 for the Educational Statistics course. In Session 5 you learnt about the features of the normal distribution and the normal curve. You also learnt about a further relationship between the standard deviation and areas under the normal curve. In this session, we shall study the major applications of the normal distribution and the normal curve.

Objectives

By the end of the session, you should be able to

(a) compute probabilities from the normal distribution and the normal curve,
(b) compute percentages from the normal distribution and the normal curve,
(c) determine performance levels based on the normal distribution and the normal curve.

6.3 Application 1. Finding probabilities

Probabilities help us to determine the chances of an occurrence of an event. This application will help us to determine the probability of a student obtaining a score based on the normal distribution. Let us look at a few examples.

Example 1

The distribution of scores for a class of students in a Statistics examination is normal with a mean of 60 and variance of 64 (i.e. X ~ N(60, 64). A student is selected at random from the class. What is the probability that the student selected obtains a score above 68?

Let us follow the following steps to solve the problem.

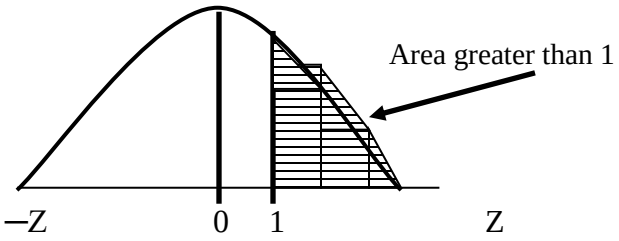
1. Write the problem in probability form P (X>68).
2. Transform the original probability form to a z-score form remembering that $Z = \frac{X - \overline{X}}{s}$

   $P(X>68) = P\left(Z > \frac{68-60}{8}\right)$ Note that the variance is 64 so the std. dev. is 8.
3. Reduce the terms.

   $P(X>68) = P\left(Z > \frac{8}{8}\right)$
4. Reduce terms further.

   $P(X>68) = P(Z > 1)$
5. Sketch the area for the normal curve.

P(Z > 1) ⟹



6. Remember that the area beyond 0 to the right is half of the area of the normal curve and this is 0.5000 and the area between 0 and 1 is 0.3413. (Session 4). Therefore to get the area shaded will be: 0.5000 ─ 0.3413
7. Write the final answer: P(X>68) = 0.5000-0.3413 = 0.1587

Let us do another similar example.

Example 2

The distribution of scores for a class of students in a Statistics examination is normal with a mean of 60 and variance of 64 (i.e. X ~ N(60, 64). A student is selected at random from the class. What is the probability that the student selected obtains a score greater than 76?

Let us follow the following steps to solve the problem.

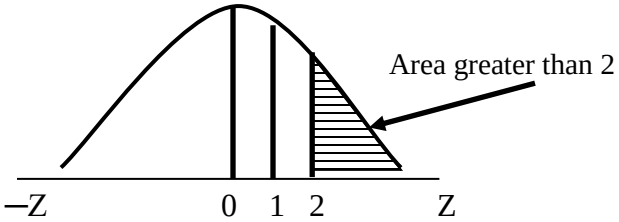
1. Write the problem in probability form P (X>76).
2. Transform the original probability form to a z-score form remembering that $Z = \frac{X - \overline{X}}{s}$

   $P(X>76) = P\left(Z > \frac{76-60}{8}\right)$
3. Reduce the terms.

   $P(X>76) = P\left(Z > \frac{16}{8}\right)$
4. Reduce terms further.

   $P(X>76) = P(Z > 2)$
5. Sketch the area for the normal curve.

P(Z > 2) ⟹



6. Remember that the area beyond 0 to the right is half of the area of the normal curve and this is 0.5000 and the area between 0 and 2 is 0.4772. (Session 4). Therefore to get the area shaded will be: 0.5000 ─ 0.4772

7. Write the final answer: P(X>76) = 0.5000-0.3413 = 0.0228

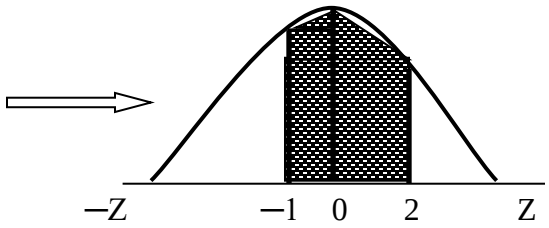
Now let us look at another example.

Example 3

The distribution for a Statistics examination is normal with a mean of 60 and variance of 64 (i.e. X ~ N(60,64). A student is selected at random from the class. What is the probability that the student selected obtains a score between 52 and 76?

We follow the same steps as above.

1. P (52<X<76)
2. $P(52<X<76) = P\left(\frac{52-60}{8} < Z < \frac{76-60}{8}\right)$
3. $P(52<X<76) = P\left(\frac{-8}{8} < Z < \frac{16}{8}\right)$
4. $P(52<X<76) = P(-1 < Z < 2)$
5. P(-1<Z < 2)



6. Remember that the area beyond 0 to the right is half of the area of the normal curve and this is 0.5000 and the area beyond 0 to the left is also half of the area of the normal curve and this is 0.5000. The area between 0 and 2 is 0.4772. (Session 4) and the area between 0 and − 1 is 0.3413 (Session4). Therefore to get the area shaded will be: 0.4772 + 0.3413
7. Write the final answer: P (52<X<76) = 0.4772 + 0.3413 = 0.8186

I want you to try the following assignments and bring them to FTF for discussion.

1. The distribution of scores for a class of students in a Statistics examination is normal with a mean of 60 and variance of 64 (i.e. X ~ N(60, 64). A student is selected at random from the class. What is the probability that the student selected obtains a score below 52?
2. The distribution for a Statistics examination is normal with a mean of 60 and variance of 64 (i.e. X ~ N(60,64). A student is selected at random from the class. What is the probability that the student selected obtains a score between 68 and 76?

6.2 Application 2. Finding percentages

This application is similar to the one for probabilities. The steps are the same except for the last step where you multiply the result with 100 to get the percentage.

Example 1

Given that the distribution of the performance of students in an examination is normal, with mean 16 and standard deviation of 2. About what percent of students obtained scores less than 12?

Let us follow the following steps to solve the problem.

1. Write the problem in probability form P (X<12)
2. Transform the original probability form to a z-score form remembering that $Z = \frac{X - \overline{X}}{s}$

   $P(X<12) = P\left(Z < \frac{12-16}{2}\right)$

3. Reduce the terms.

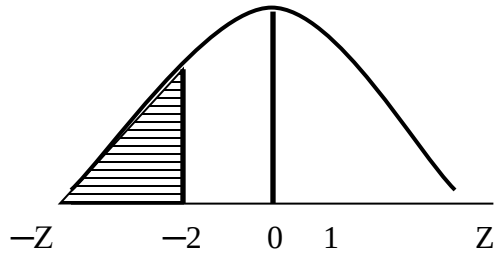
   $P(X<12) = P\left(Z < \frac{-4}{2}\right)$

4. Reduce terms further.

   $P(X<12) = P(Z < -2)$

5. Sketch the area for the normal curve.

$P(Z < -2)$ ⟹



6. Remember that the area beyond 0 to the left is half of the area of the normal curve and this is 0.5000 and the area between 0 and -2 is 0.4772. (Session 4). Therefore to get the area shaded will be: 0.5000 − 0.4772 = 0.0228
7. To find the percentage becomes: 0.0228 x 100 = 2.28%

Let us do another example.

Example 2

Given that the distribution of scores for a class of students in a Statistics examination is normal with a mean of 60 and standard deviation of 4. Approximately what percentage of students had scores between 52 and 64?

Let us follow the following steps to solve the problem.

1. Write the problem in probability form P (52<X<64).
2. Transform the original probability form to a z-score form remembering that

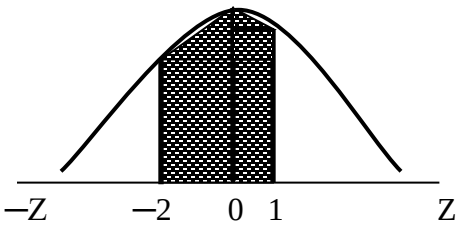
   $$Z = \frac{X - \overline{X}}{s}$$

   $$P\ (52<X<64 = P\left(\frac{52-60}{4} < Z < \frac{64-60}{4}\right)$$
3. Reduce the terms.

   $$P\ (52<X<64) = P\left(\frac{-8}{4} < Z < \frac{4}{4}\right)$$
4. Reduce terms further.

   $$P\ (52<X<64) = P(-2 < Z < 1)$$
5. Sketch the area for the normal curve.

P(-2<Z < 1) ⟹



6. Remember that the area beyond 0 to the right is half of the area of the normal curve and this is 0.5000 and the area beyond 0 to the left is also half of the area of the normal curve and this is 0.5000. The area between 0 and -2 is 0.4772. (Session 4) and the area between 0 and 1 is 0.3413 (Session4). Therefore to get the area shaded will be: 0.4772 + 0.3413 = 0.8185
7. To find the percentage becomes: 0.8185 x 100 = 81.85%

I want you to try the following assignments and bring them to FTF for discussion.

Given that a distribution is normal, with a mean of 50 and a standard deviation of 10.

1. What percentage of students had scores above 80?
2. What percentage of students had scores between 20 and 40?

6.3 Application 3. Finding number of students

This application is similar to the one for probabilities. The steps are the same except for the last step where you multiply the result with the total number of students.

Example 1

Given that the distribution of the performance of 400 students in an examination is normal, with mean 16 and standard deviation of 2. About how many students obtained scores less than 12? Let us follow the following steps to solve the problem.

1. Write the problem in probability form P (X<12)
2. Transform the original probability form to a z-score form remembering that

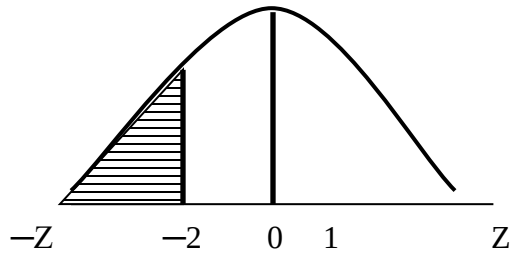
   $$Z = \frac{X - \overline{X}}{s}$$

   $$P\,(X<12) = P\left(Z < \frac{12-16}{2}\right)$$
3. Reduce the terms.

   $$P\,(X<12) = P\left(Z < \frac{-4}{2}\right)$$
4. Reduce terms further.

   $$P\,(X<12) = P(Z < -2)$$
5. Sketch the area for the normal curve.

$P(Z < -2)$ ⟹



6. Remember that the area beyond 0 to the left is half of the area of the normal curve and this is 0.5000 and the area between 0 and -2 is 0.4772. (Session 4). Therefore to get the area shaded will be: 0.5000 − 0.4772 = 0.0228
7. To find the number of students becomes: 0.0228 x 400 = 9.12
   We shall use the convention that every fraction, no matter how small, becomes a whole number since we are dealing with human beings.

   Therefore approximately 10 students had scores less than 12.

Let us do another example.

Example 2

Given that the distribution of scores for a class of 50 students in a Statistics examination is normal with a mean of 60 and standard deviation of 4. About how many students had scores between 52 and 64?

Let us follow the following steps to solve the problem.

1. Write the problem in probability form P (52<X<64).
2. Transform the original probability form to a z-score form remembering that $Z = \frac{X - \overline{X}}{s}$

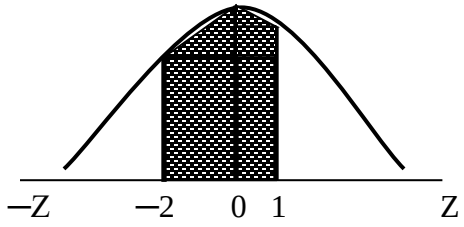
   $P\ (52<X<64 = P\left(\frac{52-60}{4} < Z < \frac{64-60}{4}\right)$
3. Reduce the terms.

   $P\ (52<X<64) = P\left(\frac{-8}{4} < Z < \frac{4}{4}\right)$
4. Reduce terms further.

   $P\ (52<X<64) = P(-2 < Z < 1)$

5. Sketch the area for the normal curve.

$P(-2<Z<1)$ ⟹



6. Remember that the area beyond 0 to the right is half of the area of the normal curve and this is 0.5000 and the area beyond 0 to the left is also half of the area of the normal curve and this is 0.5000. The area between 0 and −2 is 0.4772. (Session 4) and the area between 0 and 1 is 0.3413 (Session4). Therefore to get the area shaded will be: 0.4772 + 0.3413 = 0.8185
7. To find the number of students becomes: 0.8185 x 50 = 40.93
   Therefore the number of students is approximately 41.

In this session, you have learnt about three applications of the normal distribution and the normal curve. We have seen that the normal distribution and the normal curve can be used to compute probabilities, determine percentages and obtain the number of students who have reached a certain performance level. I trust that you have understood clearly the three applications of the normal distribution and the normal curve.

Self-Assessment Questions

Exercise 6.6

1. The distribution of an-end-of term examination scores is normal with mean 55 and standard deviation 10. In a class of 200 students, approximately how many of them obtained scores greater than 75?

2. Given that a distribution is normal, with a mean of 50 and a standard deviation of 10. From a class of 2000 students, approximately how many students obtained scores above 80?

3. In a promotion examination, a pass mark was fixed at 40. Given that the distribution is normal, with a mean of 50 and a standard deviation of 5.1, approximately how many students failed from a class of 400?

4. The distribution of a Mature Students' Examination is normal, with mean 45 and standard deviation of 15. The pass mark was 60. In a group of 400 candidates, about how many passed the examination?

5. The distribution of scores in a final psychology examination was normal with mean 55 and standard deviation of 10. What percentage of students obtained scores greater than the mean?

This is a blank sheet for your short notes on:

- Difficult topics if any
- Issues that are not clear.

UNIT 7: LINEAR CORRELATION

Unit Outline

Session 1: The concept of correlation
Session 2: Nature of the linear relationship
Session 3: Pearson product-moment correlation coefficient
Session 4: Spearman rank correlation coefficient
Session 5: Coefficients for nominal scale variables
Session 6: Uses of correlation in education

Congratulations! You have completed Unit 6. You are now welcome to the final Unit of the course, Educational Statistics. I believe that you have enjoyed the course though initially you might have had some fears. Statistics is very much alive in promoting teaching and learning in the classroom. In Unit 6 you studied the measures of relative position and the normal distribution. The normal distribution is regarded as the most important statistical distribution since many other distributions are derived from it. In this last session, we shall study correlation, which looks at the relationships that exist between variables. We shall learn how to compute various correlation coefficients and discuss how as a classroom teacher or educational practitioner, you can use the concept of correlation in your educational endeavours.

Unit Objectives

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

1. Explain the concept of correlation;
2. Describe the nature of a linear relationship between two variables;
3. Compute the Pearson product-moment correlation coefficient;
4. Compute the Spearman rank correlation coefficient;
5. Compute coefficients based on nominal variables;
6. Explain the uses of correlation in education.

SESSION 1: THE CONCEPT OF CORRELATION

You are welcome to the first session of Unit 7 for the Educational Statistics course. The concept of correlation studies relationships that exist between educational variables. These relationships, when identified, help us to take important decisions that affect the entire educational enterprise. We shall begin by understanding the concept and move on to describe the pictorial representation of the relationships. Relax and enjoy the session.

Objectives

By the end of the session, you should be able to

(a) explain the idea of correlation,
(b) draw a scatter plot for two variables,
(c) describe the nature of the linear relationship between two variables,
(d) describe the assumptions underlying the linear relationship between two variables.

Now read on...

1.1 The concept of correlation

Natural relationships exist in the world. Parents and children as well as twins have things in common. Males are normally attracted to females and rain results in good harvest.
In education, absenteeism tends to go with performance in class tests and examinations. Studies have also shown that females generally do better than men in the reading subjects like English Language, and History while males generally tend to do better than females in the science-related subjects like Physics, Chemistry and Mathematics.

The concept of correlation provides information about the extent of the relationship between two variables. Two variables are correlated if they tend to 'go together'. For example, if high scores on one variable tend to be associated with high scores on a second variable, then both variables are correlated. Correlations aim at identifying relationships between variables and also to be able to predict performances based on known results.

The statistical summary of the degree and direction of the linear relationship or association between any two variables is given by the coefficient of correlation. Correlation coefficients range between −1.0 and +1.0. Correlation coefficients are normally represented by the symbols, r (for sample) and ρ (rho) (for populations).

1.2 Scatter plots

A scatter plot or scatter diagram gives a pictorial representation of two variables and shows the nature of the relationship between the two variables. It is important that scatter plots are drawn before any analysis is done on the variables. This is because scatter plots could either be linear or curvilinear. Examples are shown below.

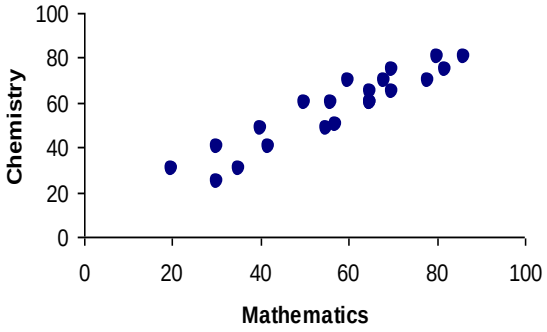


Figure 7.1 Linear relationship

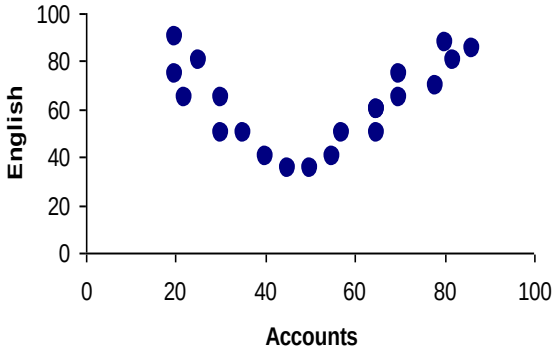


Figure 7.2 Curvilinear relationship

Figure 7.1 shows a linear relationship because a straight line can be drawn through the middle of the points. However, Figure 7.2 is such that a straight line cannot be conveniently drawn through the middle. If we draw a line through the middle, it will assume a "U" shape which is called curvilinear. For this course, we shall be concerned with only the linear relationship between any two variables.

1.2.1 Drawing a scatter plot

Follow the following steps to draw a scatter plot. These steps are used when software like Microsoft Excel are not available. It is always recommended that where software is not available, graph sheets are used.

Step 1

Draw two axes (a vertical axis and a horizontal axis). Label the horizontal axis by the first variable and the vertical axis by the second variable.

Step 2

Divide both axes by units or points, considering the lowest and highest values. Choose appropriate scales such that the graph is not too tall or too flat and must start with zero.

Step 3

For each pair of values find the meeting point in the graph space and make a mark (preferably a big dot,●).

These steps are very simple and I believe that you have followed them. I want you to try and draw a scatter plot.

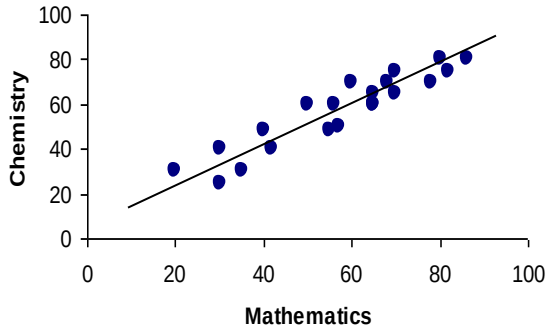
Given the data below draw a scatter plot using a graph sheet. Bring your scatter plot to FTF for discussion

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Quiz 1	14	16	15	10	9	18	18	8	8	13	15	18	10	12	16	20	13	12	20	10
Quiz 2	13	14	13	11	12	15	15	10	11	14	14	14	12	13	13	15	12	12	16	13

1.3 Assumptions

Correlational analysis is based on a number of assumptions. These assumptions are described below.

1. The variables are random. The values of the two variables are not predetermined. These values can change.
2. The relationship between the variables is linear. If a line is drawn through the middle of the points, it must be a straight line. An example is below.



3. The probability distribution of X's, given a fixed Y, is normal, i.e. the sample is drawn from a joint normal distribution. This is shown below.

   Assume the following scores in two tests.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	14	16	15	10	9	18	18	14	12	13	15	18	10	12	16	20	15	12	14	10
Y	10	12	15	10	12	15	15	12	14	14	14	10	12	15	10	12	15	15	10	14

   If Y = 10, X = 14, 10, 18, 16, 14; the distribution of X must be normal in population.
   If Y = 12, X = 16, 9, 14, 10, 20; the distribution of X must be normal in population.
   If Y = 14, X = 12, 13, 15, 10; the distribution of X must be normal in population.
   If Y = 15, X = 15, 18, 18, 12, 15, 12; the distribution of X must be normal in population.

4. The standard deviation of X's, given each value of Y is assumed to be the same, just as the standard deviation of Y's given each value of X is the same. For example:

   Assume the following scores in two tests.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	14	16	15	10	9	18	18	14	12	13	15	18	10	12	16	20	15	12	14	10
Y	10	12	15	10	12	15	15	12	14	14	14	10	12	15	10	12	15	15	10	14

If Y = 10, X = 14, 10, 18, 16, 14; same standard deviation in population.
If Y = 12, X = 16, 9, 14, 10, 20; same standard deviation in population.
If Y = 14, X = 12, 13, 15, 10; same standard deviation in population.
If Y = 15, X = 15, 18, 18, 12, 15, 12; same standard deviation in population.
The standard deviation for the X values when Y = 10 must be the same as the standard deviation for the X values when Y = 12 and must be the same for the X values when Y = 14 and likewise when Y = 15.

In this session, you have learnt about the concept of correlation. We have noted that correlation deals with the relationship between variables and for two variables, we have the bivariate situation. The correlation coefficient is used to determine the strength of the correlation or relationship. In our case, we are focusing on the linear relationship between two variables. We have also learnt about the four assumptions that must be met for the correlation analysis. I am sure you have grasped the concept very well.

Self-Assessment Questions

Exercise 7.1

Draw a scatter plot for the following set of data.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Physics	70	75	88	56	60	80	45	50	68	90	40	55	72	65	58	67	90	76	50	40
History	50	60	48	85	60	45	92	86	72	58	85	70	66	75	80	70	40	55	80	70

Is the relationship Linear or Curvilinear?

SESSION 2: NATURE OF THE LINEAR RELATIONSHIP

You are welcome to the Session 2 of Unit 7 for the Educational Statistics course. In Session 1 you learnt about the concept of correlation. You learnt that correlation deals with the relationship between variables and for two variables, we have the bivariate situation. The correlation coefficient is used to determine the strength of the correlation or relationship. You also learnt about the four assumptions that must be met for the correlation analysis. In this session, you will study the nature of the linear relationship between two variables.

Objectives

By the end of the session, you should be able to

(a) describe the directions of a linear relationsip,
(b) describe the degrees of linear relationships,
(c) state the commonly used types of correlation coefficients,
(d) explain the term, coefficient of determination.
(e) describe the relationship between causation and correlation.

Now read on...

2.1 Nature of the linear relationship

The linear relationship between two variables is described by direction and degree. The direction indicates whether there is a positive or negative linear relationship and the degree tells us about how strong the relationship is.

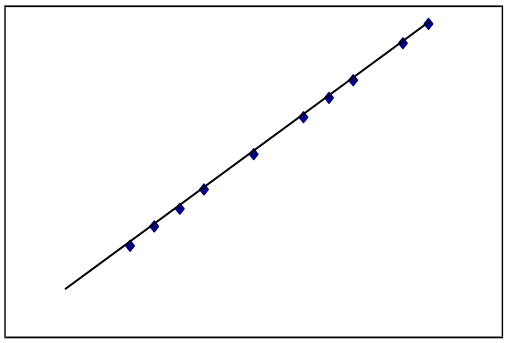
With respect to the direction, when the high values in one variable go with high values in the other variable the relationship is positive (+). Also when low values in one variable go with low values in the other variable the relationship is positive (+). On the other hand, when high values in one variable go with low values in the other variable the relationship is negative (—).

The degree of the relationship is expressed in terms of values between —1 and +1.

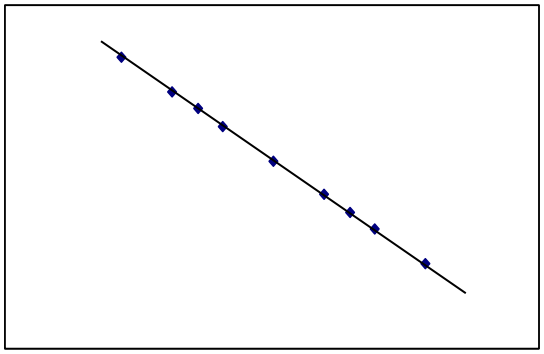
1. When the relationship is strong (high), the value of the correlation coefficient, r, is greater than 0.60 or less than −0.60, i.e. $r > 0.60$ $r < -0.60$
2. When the relationship is moderate (mild), the value of the correlation coefficient, r, lies between 0.40 and 0.60 or —0.60 and —0.40, i.e. $0.40 \leq r \leq 0.60$ or $-0.40 \geq r \geq -0.60$
3. When the relationship is weak (low), the value of the correlation coefficient, r, is less than 0.40 or greater than −0.40, i.e. $r < 0.40$ $r > -0.40$
4. When the relationship is perfect, the value of the correlation coefficient, r, is 1.0 or −1.0 i.e. $r = 1$ for perfect positive $r = -1.0$ for perfect negative.
5. When there is no linear relationship, the value of the correlation coefficient, r, is 0.0

2.2 Graphs of the linear relationships

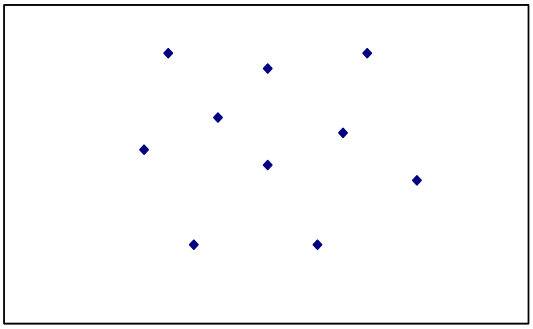
The direction and the degree of the linear relationship between two variables can be shown in graphs. Examples are shown below.



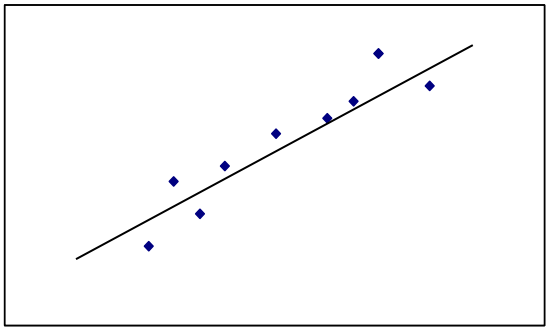
Perfect positive linear correlation
All the points lie on a straight line.
r = 1.0



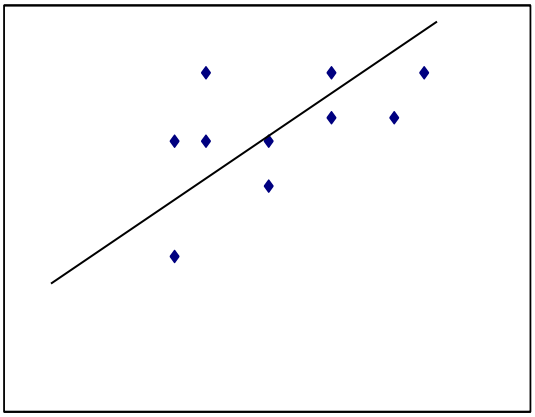
Perfect negative linear correlation
All the points lie on a straight line.
r = ─1.0



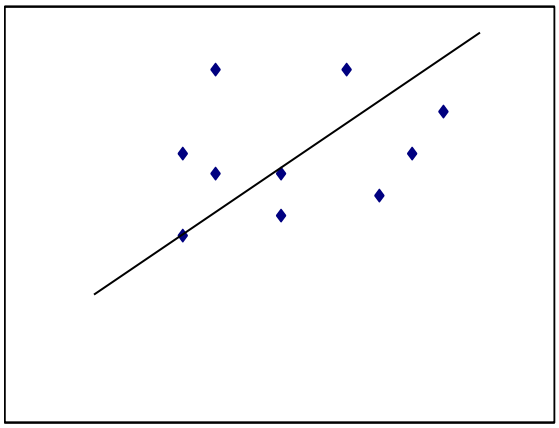
Zero linear correlation
The points surround the straight line more like an O r = 0



High positive linear correlation
The points are close to the straight line.
r > 0.60



Moderate linear correlation
The points are a bit farther away from



Low linear correlation
The points are quite far away from the

the straight line in the middle. $0.40 \leq r \leq 0.60$

middle line $r < 0.40$

2.3 Commonly used correlation coefficients

There are 4 commonly used correlation coefficients. These are described below.

1. Pearson Product Moment correlation coefficient (r). This is applicable when both variables are continuous in nature. It uses interval and ratio scale data. For example, the relationship between test scores and age of students.
2. Spearman's rank correlation coefficient (ρ - rho). This is suitable for variables that are both continuous and ranked. It uses ordinal scale data. For example, ranks in terms of school attendance and position in a test.
3. Phi coefficient (φ - phi). This is used when both variables are natural dichotomies. A natural dichotomy consists of variables that have only two natural categories, for example gender which has only male and female. The phi coefficient is applicable to nominal data. For example the relationship between gender (male, female) and political party membership (member, not member).
4. Point-biserial correlation coefficient (r_{pb}). This is applicable when one variable is continuous and the other is a natural dichotomy. It combines nominal scale data with either interval or ratio scale data. For example, the relationship between gender and test scores.

2.4 Coefficient of Determination (r^2)

The coefficient of determination is the square of the correlation coefficient. It is the proportion of the variance in Y accounted for by X. An r of 0.71 gives r^2 to be 0.50. This means that 50% of the variance in Y is associated with variability in X. For example, if the correlation between class attendance (X) and performance in Statistics (Y) is 0.8, then class attendance explains 64% (0.64) of the variation in the scores in performance in Statistics. Also if the correlation between study habit (X) and performance in Quiz 1 (Y) is 0.9, then study habit (X) explains 81% (0.81) of the variation in the scores in Quiz 1 (Y).

2.5 Causation and correlation

The presence of a correlation between two variables does not necessarily mean that there exists a causal relationship between the two variables. A very strong or high relationship between two variables does not imply that one causes the other. No cause and effect relationship is determined purely by correlation coefficients. Correlation only tells us that there is a linear relationship and how strong the relationship is and ends there.

In this session, you have learnt about the nature of the linear relationship between two variables. You have learnt that the nature is described in terms of the direction and degree of the linear relationship. You have seen examples of the linear relationships between two variables and the commonly used correlation coefficients. The concept of the coefficient of determination and the relationship between causation and correlation have also been explained to you. I believe you have had a good understanding of this session.

Self-Assessment Questions

Exercise 7.2

1. The Phi Coefficient (Φ) is the most appropriate measure of linear relationship when two variables are:

    A. Both continuous.
    B. Both natural dichotomies.
    C. Continuous and natural dichotomy.
    D. Continuous and artificial dichotomy.

2. A University professor wishes to find the relationship between the age of the students in her Statistics class and the scores in a quiz. What is the most appropriate measure of relationship to use?

    A. Pearson's product moment correlation coefficient
    B. Phi coefficient
    C. Point-biserial correlation coefficient
    D. Spearman's rank correlation coefficient

3. Which of the following correlation coefficients indicates the strongest relationship?
    A. –0.6
    B. 0.07
    C. 0.25
    D. 0.55

4. A teacher wishes to find the relationship between the spelling ability of the students in her class and their performance in the end-of-semester examination. What is the most appropriate measure of relationship to use?

    A. Pearson's product moment correlation coefficient
    B. Phi coefficient
    C. Point-biserial correlation coefficient
    D. Spearman's rank correlation coefficient

5. The correlation between study habits and achievement in Statistics has been found to be 0.92 in a study. The study implies that a student with a

    A. high score in study habits is more likely to score low in Statistics.
    B. low score in study habits is more likely to obtain a moderate score in Statistics.
    C. low score in study habits is more likely to score low in Statistics.
    D. moderate score in study habits is more likely to score high in Statistics.

6. A teacher wishes to find the relationship between the gender of students in his class and their performance in the end-of-semester examination. What is the most appropriate measure of relationship to use?
    A. Phi coefficient
    B. Spearman's rank correlation coefficient.
    C. Point-biserial correlation coefficient.
    D. Pearson's product moment correlation coefficient.

SESSION 3: PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT

Hello! You are welcome to Session 3 of the last Unit for the Educational Statistics course. In Session 2, you learnt about the nature of the linear relationship between two variables and how this nature is described in terms of the direction and degree of the linear relationship. You have also seen examples of the linear relationships between two variables and the commonly used correlation coefficients. You have studied the concept of the coefficient of determination and the relationship between causation and correlation. In this session, we shall learn how to compute the correlation coefficient using the Pearson method.

Objectives

By the end of the session, you should be able to

(a) describe the conditions necessary for the use of the product moment correlation method,
(b) compute the correlation coefficient using the covariance method,
(c) compute the correlation coefficient using the product method,
(d) interpret the correlation coefficient.

Now read on...

3.1 Obtaining the Pearson product moment correlation coefficient

The product moment correlation coefficient was derived by a statistician called Karl Pearson. Two conditions are necessary for obtaining the correlation coefficient, r, using the product moment method. These conditions are explained below.

1. The two variables must be continuous in nature and must come from interval or ratio scales. For example, examination scores and ages of students are variables from interval and ratio scales.
2. Evidence must be shown that the relationship between the variables is linear. To get this evidence a scatter plot is drawn and observed.

3.2 Calculating the correlation coefficient.

There are two methods for calculating the correlation coefficient using the Pearson product moment. These are the covariance method and the product method.

3.2.1 Using the covariance method

This method uses the covariance between the two variables. It involves calculating the covariance and dividing the result by the product of the standard deviations of the two variables. The formula is given below.

$$r = \frac{\text{Covariance}(X, Y)}{S_X . S_Y}$$

The formula can be rewritten as $r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{nS_x S_Y}$

However, to make the computation easier the following formula is used.

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 . \sum (Y - \overline{Y})^2}}$$

Here are the steps to follow when using the covariance method.

1. Obtain the mean for the X values ($\overline{X}$).
2. Obtain the deviations from the mean ($X - \overline{X}$).
3. Obtain the mean for the Y values ($\overline{Y}$).
4. Obtain the deviations from the mean ($Y - \overline{Y}$).
5. Multiply Step 2 with Step 4 ($X - \overline{X}$)($Y - \overline{Y}$)
6. Sum the values in Step 5 $\sum(X - \overline{X})(Y - \overline{Y})$. This gives the numerator.
7. Square the values in Step 2 $(X - \overline{X})^2$
8. Sum the values in Step 7 $\Sigma(X - \overline{X})^2$
9. Square the values in Step 4 $(Y - \overline{Y})^2$
10. Sum the values in Step 9 $\Sigma(Y - \overline{Y})^2$
11. Multiply the values in Step 8 with the values in Step 10. $\Sigma(X - \overline{X})^2 \ \Sigma(Y - \overline{Y})^2$
12. Find the square root of the value obtained in Step 11. $\sqrt{\left[\left(\sum X - \overline{X}\right)^2 . \left(\sum Y - \overline{Y}\right)^2\right]}$
13. Divide the result in Step 6 by the result in Step 12.

These steps appear long but they are easy to follow. An example is done for you. The steps are highlighted.

Step 2 Step 7 Step 4 Step 9 Step 5

Student No.	Quiz 1 X	Quiz 2 Y	$X-\overline{X}$	$(X-\overline{X})^2$	$Y-\overline{Y}$	$(Y-\overline{Y})^2$	$(X-\overline{X})(Y-\overline{Y})$
1	4	6	-2	4	-1	1	2
2	8	8	2	4	1	1	2
3	10	9	4	16	2	4	8
4	7	7	1	1	0	0	0
5	6	8	0	0	1	1	0
6	3	2	-3	9	-5	25	15
7	8	9	2	4	2	4	4
8	5	10	-1	1	3	9	-3
9	5	6	-1	1	-1	1	1
10	4	5	-2	4	-2	2	4
Total	60	70		44		50	33

Step 1: $\overline{X} = 6$

Step 3: $\overline{Y} = 7$

Step 8 Step 10 Step 6

Using Formula 1:

$$r = \frac{33}{\sqrt{(44)(50)}} = \frac{33}{46.9} = 0.7$$

Step 12 Step 11 Step 13

You will notice that the steps are simple and easy to follow.

So what does r = 0.7 mean in terms of direction and degree? Do you have any ideas?

Well, the result shows that there is a strong positive relationship between Quiz 1 and Quiz 2. This implies that if you get high scores in Quiz 1, you are likely to get high scores in Quiz 2.

Now I want you to follow the steps and calculate the correlation coefficient for the data below.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Quiz 1	14	16	10	10	9	18	18	8	8	13	10	16	10	12	13	20	13	12	20	10
Quiz 2	13	14	13	11	12	15	15	10	11	14	14	14	11	12	13	15	12	12	16	13

What answer did you get? Did you get an answer close to 0.85?

If your answer is 0.85 or close, then congratulations, you have done well. If your answer is very different, then check your steps and your calculations again.
Now let us try the product method.

3.2.2 Using the product method

This method uses the product of the two variables and squares of each variable for the computations. The formula is given below.

$$r = \frac{n\sum XY - \sum X \sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

Here are the steps to follow when using the product method.

1. Identify the sample size, n.
2. Obtain the product of the two variables, XY.
3. Obtain the sum of the products in Step 2 $\sum XY$.
4. Multiply the result in Step 3 by n, $n\sum XY$
5. Find the sum of the X values, $\sum X$
6. Find the sum of the Y values $\sum Y$
7. Find the product of $\sum X$ and $\sum Y$ $(\sum X)(\sum Y)$
8. Subtract Step 7 from Step 4. $(n\sum XY) - (\sum X)(\sum Y)$. This gives the numerator.
9. Square the X values and find the sum. $\sum X^2$
10. Multiply the result in Step 9 with the sample size, n. $n\sum X^2$
11. Square the result in Step 5 and subtract from result in Step 10 · $n\sum X^2 - (\sum X)^2$
12. Square the Y values and find the sum, $\sum Y^2$
13. Multiply the result in Step 12 with the sample size, n. $n\sum Y^2$
14. Square the result in Step 6 and subtract from result in Step 13 · $n\sum Y^2 - (\sum Y)^2$
15. Multiply the results in Steps 11 and 14 and find the square root. This gives the denominator.
16. Divide the result in Step 8 by the result in Step 15 to get the answer.

These steps appear long but they are easy to follow. An example is done for you. The steps are highlighted.

Student No.	Quiz 1 X	Quiz 2 Y	XY (Step 2)	X^2 (Step 9)	Y^2 (Step 12)
1	4	6	24	16	36
2	8	8	64	64	64
3	10	9	90	100	81
4	7	7	49	49	49
5	6	8	48	36	64
6	3	2	6	9	4
7	8	9	72	64	81
8	5	10	50	25	100
9	5	6	30	25	36
10	4	5	20	16	25
Total	60	70	453	404	540

Step 1 Step 5 Step 6 Step 3 Step 9 Step 12

Using the formula and substituting the values gives:

$$r = \frac{n\sum XY - \sum X \sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} = \frac{4530 - 4200}{\sqrt{(440)(500)}} = \frac{330}{469.04} = 0.7$$

Step 4 Step 7 Step 8 Step 10 Step 13 Step 15 Step 11 Step 14

You will notice that the steps are simple and easy to follow. You will notice that the answer we got here is the same as the answer we got with the covariance method. It does not matter therefore which method is used. The answers will always be exact or very close.

Now I want you to follow the steps and calculate the correlation coefficient for the data below.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Quiz 1	11	17	12	15	8	15	16	10	17	12	17	15	12	14	13	15	20	20	12	9
Quiz 2	10	14	15	16	12	16	15	15	18	16	18	18	15	16	10	12	20	19	14	11

What answer did you get? Did you get an answer close to 0.7?

If your answer is 0.708 or close, then congratulations, you have done well. If your answer is very different, then check your steps and your calculations again.

In this session, you have learnt about the conditions necessary for the use of the Pearson product moment correlation coefficient (r). You have also learnt how to compute the correlation coefficient using both the covariance method and the product method. I believe you got the examples you did right. Congratulations for completing this session.

Self-Assessment Questions

Exercise 7.3

Given the data below compute the Pearson product moment correlation coefficient using:

1. The covariance method
2. The product method

Interpret your result.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Physics	70	75	88	56	60	80	45	50	68	90	40	55	74	64	58	64	80	65	90	88
History	50	60	48	85	60	70	92	86	72	58	85	45	66	75	80	70	46	60	42	50

SESSION 4: SPEARMAN RANK CORRELATION COEFFICIENT

Hello! You are welcome to the Session 4 of the last Unit for the Educational Statistics course. In Session 3, you learnt about the conditions necessary for the use of the Pearson product moment correlation coefficient (r). You also learnt how to compute the correlation coefficient using both the covariance method and the product method. I believe you followed the steps closely and got the examples right. Congratulations. In this session, we shall learn how to compute the correlation coefficient using the Spearman rank order correlation method.

Objectives

By the end of the session, you should be able to

(a). describe the conditions necessary for the use of the Spearman rank order correlation method,

(b). compute the correlation coefficient using the rank order correlation method,

(c). interpret the Spearman correlation coefficient (ρ, rho).

Now read on.....

4.1 Obtaining the Spearman rank order correlation coefficient (ρ, rho).

The rank order correlation coefficient was derived by a statistician called Charles Spearman. It generally provides an estimate of the linear relationship between two variables. It is not as accurate as the product moment correlation coefficient. This is because it does not deal with the actual quantities from the variables but rather the ranks of the variables.

Two conditions are necessary for obtaining the correlation coefficient, ρ,(rho) using the rank order method. These conditions are explained below.

1. The two variables must be ranked in nature and must come from ordinal scales. For example, ranks in national examinations and ranks in terms of school size.
2. Evidence must be shown that the relationship between the variables is linear. To get this evidence a scatter plot is drawn and observed.

4.2 Calculating the rank order correlation coefficient

This method involves calculating the differences between the ranks. The formula is given below.

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

Here are the steps to follow when using the rank order correlation method.

1. Obtain the sample size, N.
2. Obtain d, the differences between the ranks. $R_1 - R_2 = d$
3. Square the differences and sum up the differences. $\sum d^2$
4. Multiply the result in Step 3 by 6. $6\sum d^2$
5. Obtain the values of $N(N^2 - 1)$.
6. Divide the values in Step 4 by the result in Step 5.
7. Subtract the result from 1 and obtain ρ (rho)

These steps are few and are easy to follow. An example is done for you. The steps are highlighted.

Given the following scores:

Student No.	Quiz 1 Ranks	Quiz 2 Ranks	d R_1-R_2	d^2
1	8.5	7.5	1	1.0
2	2.5	4.5	-2.0	4.0
3	1	2.5	-1.5	2.25
4	4	6	-2	4.0
5	5	4.5	0.5	0.25
6	10	10	0	0.0
7	2.5	2.5	0	0.0
8	6.5	1	5.5	30.25
9	6.5	7.5	-1.0	1.0
10	8.5	9	-0.5	0.25
				43.00

Step 1 Step 2 Step 3

Now compute the correlation coefficient by substituting the values in the formula.

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)} = 1 - \frac{6(43)}{10(100 - 1)} = 1 - \frac{258}{990} = 1 - 0.26 = 0.74$$

Step 4 Step 5 Step 6 Step 7

The result, ρ = 0.74 shows that there is a strong positive relationship between Quiz 1 and Quiz 2.

You will notice that the steps are simple and easy to follow.

Now I want you to follow the steps and calculate the rank order correlation coefficient for the data below.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Quiz 1 rank	7	6	14	14	18	3.5	3.5	19	20	8.5	17	5	16	12	8.5	1.5	11	10	1.5	14
Quiz 2 rank	6	18	12	14	7	3	20	4	9	19	18	5	17	8	13	2	10.5	10.5	1	16

What answer did you get? Did you get an answer close to 0.7?

If your answer is 0.87 or close, then congratulations, you have done well. If your answer is very different, then check your steps and your calculations again.

In this session, you have learnt about the conditions necessary for the use of the Spearman rank order correlation coefficient (ρ). You have also learnt how to compute the correlation coefficient (ρ) and to interpret the results. I trust that you got the example right. Congratulations for completing this session.

Self-Assessment Questions

Exercise 7.4

Given the data below compute the Spearman rank order correlation coefficient and interpret the results.

1.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Maths rank	10	4	2	10	8	3	14	10	6.5	1	15	12	5	6.5	13
Statistics rank	7	3	1	12.5	9	2	12.5	10	5	4	14	11	15	7	7

2.

School	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Enrolment rank	9	4	14	7.5	2	3	15	10	1	11	12.5	7.5	5	6	12.5
Examination rank	15	13	1	12	14	11	2	6	5	3	4	9.5	9.5	8	7

SESSION 5: COEFFICIENTS FOR NOMINAL SCALE VARIABLES

Hello! You are welcome to the Session 5 of the last Unit for the Educational Statistics course. In Session 4, you learnt about the conditions necessary for the use of the Spearman rank order correlation coefficient (ρ). You also learnt how to compute the correlation coefficient (ρ) and to interpret the results. I trust that you have understood the procedure. Congratulations for completing this session. In Session 4 it was emphasised that the Spearman rank order correlation coefficient is used for variables that are from the ordinal scale. In this session we shall look at how to compute the correlation coefficients for variables from the nominal scale.

Objectives

By the end of the session, you should be able to

(a) describe the conditions necessary for computing correlation coefficients from nominal variables,
(b) compute the phi correlation coefficient,
(c) compute the contingency correlation coefficient,
(d) interpret the correlation coefficients computed.

Now read on...

5.1 Obtaining the correlation coefficient for nominal variables

We shall study two correlation coefficients that are used for nominal variables. These are the Phi coefficient and Contingency coefficient. Nominal variables do not generally follow any statistical distribution and are therefore termed, non-parametric. The major condition necessary for the use of the Phi and Contingency coefficient is that the variables must nominal.

Do you remember the features of nominal variables? Write two features in the space provided below.

Now go to Unit 1 Session 2 and read over the features of nominal variables.

I believe you got the two features correct. Well done.

We shall start with the phi coefficient.

5.2 Calculating the phi (φ) correlation coefficient

The phi (φ) is computed for nominal variables that have natural dichotomies or have only two naturally occurring categories, for example, gender which produces male and female, performance in examination which produces pass and fail. It is most appropriate for nominal variables that have only two categories or are dichotomous.

The formula is given below.

$$\text{Phi } (\varphi) = \sqrt{\frac{\chi^2}{n}}$$

The formula can be defined as the square root of the chi-square (χ^2) value divided by the sample size, n. Before the phi coefficient is calculated, the chi-square (χ^2) value has to be obtained. The formula is given below with an example.

The formula for calculating the χ^2 value is as follows.

$$\chi^2 = \sum_{j=1}^{r}\sum_{k=1}^{c}\frac{(f_o - f_e)^2}{f_e} \quad \text{where}$$

f_o is the observed count in each cell.

f_e is the expected count in each cell.

To obtain f_e for each cell, multiply the row total for the cell by the column total for the cell and divide the result by the grand total.

$$f_e = \frac{\text{row total for the cell} \cdot \text{column total for the cell}}{\text{grand total}}$$

Example: Association between Halls of Residence and Region of Birth

Region of Birth	Hall of Residence		Total
	Hall 1	Hall 2	
Region 1	40 (50)	60 (50)	100
Region 2	100 (75)	50 (75)	150
Region 3	60 (75)	90 (75)	150
Total	200	200	400

The numbers without brackets are the observed counts, f_o while the numbers in bracket are the expected counts, f_e

The expected counts are as follows:

Region 1 Hall 1 $\frac{100 \text{ x } 200}{400} = 50$ Region 1 Hall 2 $\frac{100 \text{ x } 200}{400} = 50$

Region 2 Hall 1 $\frac{150 \text{ x } 200}{400} = 75$ Region 2 Hall 2 $\frac{150 \text{ x } 200}{400} = 75$

Region 3 Hall 1 $\frac{150 \text{ x } 200}{400} = 75$ Region 3 Hall 2 $\frac{150 \text{ x } 200}{400} = 75$

Now let us apply the formula:

$$\chi^2 = \sum_{j=1}^{r}\sum_{k=1}^{c}\frac{(f_o - f_e)^2}{f_e} = \frac{(40-50)^2}{50} + \frac{(60-50)^2}{50} + \frac{(100-75)^2}{75} + \frac{(50-75)^2}{75} + \frac{(60-75)^2}{75} + \frac{(90-75)^2}{75}$$

$$\chi^2 = \frac{(-10)^2}{50} + \frac{(10)^2}{50} + \frac{25^2}{75} + \frac{(-25)^2}{75} + \frac{(-15)^2}{75} + \frac{(15)^2}{75}$$

$$\chi^2 = \frac{(-10)^2}{50} + \frac{(10)^2}{50} + \frac{25^2}{75} + \frac{(-25)^2}{75} + \frac{(-15)^2}{75} + \frac{(15)^2}{75}$$

$$\chi^2 = \frac{100}{50} + \frac{100}{50} + \frac{625}{75} + \frac{625}{75} + \frac{225}{75} + \frac{225}{75}$$

$$\chi^2 = 2.0+2.0+8.3+8.3+3.0+3.0$$

$$\chi^2 = 26.6$$

Now let us try an example to obtain the phi (φ) coefficient.

What is phi (φ) value for the data below? What is the strength of the relationship between job category and educational qualification?

The data below gives the number of students in each job and the educational qualification.

Qualification	Job Category	
	Clerical	Managerial
Diploma	60	40
1st degree	20	80

1. First we compute the expected counts.

   Diploma Clerical $\frac{100 \text{ x } 80}{200} = 40$ Diploma Managerial $\frac{100 \text{ x } 120}{200} = 60$

   1st Degree Clerical $\frac{80 \text{ x } 100}{200} = 40$ 1st Degree Managerial $\frac{100 \text{ x } 120}{200} = 60$

2. Calculate the χ^2

   $$\chi^2 = \frac{(60-40)^2}{40} + \frac{(40-60)^2}{60} + \frac{(20-40)^2}{40} + \frac{(80-60)^2}{60}$$

   $$\chi^2 = \frac{(20)^2}{40} + \frac{(-20)^2}{60} + \frac{(-20)^2}{40} + \frac{(20)^2}{60}$$

   $$\chi^2 = \frac{400}{40} + \frac{400}{60} + \frac{400}{40} + \frac{400}{60}$$

   $$\chi^2 = 10.0+6.7+10.0++6.7$$

   $$\chi^2 = 33.4$$

$$\text{Phi } (\varphi) = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{33.4}{200}} = \sqrt{\frac{33.4}{200}} = \sqrt{0.167} = 0.4$$

The result shows that there is a moderate positive relationship between job and educational qualification.

I trust that you have clearly understood the procedure for calculating the phi value.

Now I want you to try the assignment below and bring it to FTF for discussion.

What is phi (φ) value for the data below? What is the strength of the relationship between gender and passing Statistics?

The data below gives the gender of students and their result in a Statistics examination.

Gender	Exam Result	
	Pass	Fail
Female	180	20
Male	220	80

5.3 Calculating the contingency coefficient (C)

The contingency coefficient (C) is computed for nominal variables. It is most appropriate when one of the variables has more than two categories.

The formula is given below.

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

The formula is based on the sample size and the value of the chi-square (χ^2).

Using the value for the example on region of birth and hall of residence where we have 3 rows by 2 columns (3 x 2), the contingency coefficient gives us:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{26.6}{400 + 26.6}} = \sqrt{\frac{26.6}{426.6}} = \sqrt{0.062} = 0.25$$

The result shows that there is a weak positive relationship between region of birth and hall of residence.

I trust that you have clearly understood the procedure for calculating the contingency coefficient.

Now I want you to try the assignment below and bring it to FTF for discussion.

What is the contingency coefficient (C) for the data below? What is the strength of the relationship between gender and the subjects taken?

The data below gives the gender of students and the courses registered for in a College.

Courses	Gender	
	Female	Male
BA	650	380
BEd	350	320
BSc	100	400
BCom	400	400

Congratulations for completing this session. In this session, you learnt about the conditions necessary for the use of the phi and the contingency coefficients. You also learnt how to compute the phi and the contingency coefficients and to interpret the results. I trust that you have done the assignments for FTF discussion.

Self-Assessment Questions

Exercise 7.5

1. Given the data below compute phi and interpret the result.

   The data below gives the gender of students and their final examination results.

Gender	First Class	
	No	Yes
Female	1180	320
Male	2020	480

2. Given the data below compute contingency coefficient and interpret the result.

   The data below gives the final examination classification and the employment.

Classification	Employment		
	Commerce	Industry	Teaching
1st Class	750	150	100
2nd Class (Upper)	350	1500	150
2nd Class (Lower)	100	350	1550

SESSION 6: USES OF CORRELATION IN EDUCATION

Wow! You have reached the last session of the last Unit of the Educational Statistics course. Congratulations! In Session 5, you learnt about the conditions necessary for the use of the phi and contingency coefficients. You learnt how to compute these coefficients which are based on the Chi-square. (χ^2). You also learnt how to interpret the results. I trust that you have understood the procedures. In this session we shall look at the uses of correlation in education. I wish you a good understanding of this last session.

Objectives

By the end of the session, you should be able to explain

(a) how correlation is used for selection and placement,
(b) the importance of correlation in reliability of test scores,
(c) the importance of correlation in the validity of assessment instruments,
(d) how correlation aids in prediction,
(e) how correlation aids research in education.

Now read on...

6.1 It is useful for selection and placement

Selection involves getting the right calibre of people for admission and promotion. Those not acceptable are rejected. Placement involves placing students in courses and classes where they are likely to succeed in the future. Knowledge of correlation between two variables does help in the selection and placement process. For example, if mathematics scores relate well with scores in chemistry, then mathematics scores can be used for selection into a chemistry class without conducting a chemistry selection examination. This action reduces or eliminates waste of resources.

In the 1990s, university entrance examinations (UEE) were conducted in Ghana for selection into the Universities of Ghana to supplement GCE "A" Level results. The entrance examinations used a lot of resources and placed financial burdens on some applicants. However, studies conducted showed that a strong correlation existed between performance in the GCE "A" Level results and the scores on the university entrance examinations. Since the GCE "A" Level and the UEE were performing the same tasks (as evidenced by the strong correlation), it was necessary to terminate the entrance examination and concentrate on the GCE "A" Levels for admission.

6.2 It is used to determine the reliability of standardized and classroom tests

In doing this programme, I believe that you have learnt about the principles in writing classroom tests in one of your courses. Good. Achievement tests have two major properties or characteristics when the results are being interpreted and used.

What are the two major properties or characteristics of achievement tests? Have you remembered them?

Well the two characteristics or properties are reliability and validity.
Now let us turn to reliability.

What is meant by the term, reliability, when applied to achievement test results? Reliability refers to the consistency of achievement test results over time. It is the degree to which achievement test results are the same when (1) the same tasks are completed on two different occasions by the same group of people, (2) different but equivalent tasks are completed on the same or different occasions by the same group of people, and (3) two or more raters mark performance on the same tasks.

What are the methods used in estimating the reliability of achievement test results? Do you remember them? What are the methods?

Those who took a course in Introduction to Measurement and Evaluation, will remember that there are 5 methods. These methods are:

1. Test-retest reliability
2. Alternate or Equivalent forms reliability
3. Split-Half reliability
4. Kuder-Richardson reliability
5. Inter-rater reliability

Of the five methods, four of them use correlation coefficients.

In test-retest, a test is first administered to a group of students. After a while, the same test is administered to the same group of students. The correlation coefficient is obtained for scores on the first administration of the test and the second administration of the test to give an estimate of the reliability of the test.

In alternate or equivalent forms, a test is first administered to a group of students. After a while, an alternate or equivalent form of the first test is administered to the same group of students. The correlation coefficient is obtained for the scores on the administration of the tests on the two different occasions. This gives an estimate of the reliability of the test.

For the split-half method, one test is administered and split into two parts. The correlation coefficient is obtained for the scores from each half of the test. The correlation coefficient obtained is adjusted by using the Spearman-Brown split-half method to obtain the reliability of the full length test.

In inter-rater reliability, two persons score or rate each student's paper. The scores for the two raters are correlated to obtain the scorer reliability.

6.3 It aids in the provision of evidences for the validity of assessment instruments

We have just noted that there are two important characteristics or properties of a good achievement test. These are reliability and validity. We saw that correlation coefficients are needed for determining the reliability of test results.

Now let us look at validity.

What is meant by the term validity, when applied to achievement test results?

Validity refers to the soundness and appropriateness of the interpretation and use of achievement test results. To determine the validity of achievement test results, evidences must be gathered.

What are the evidences that must be gathered to aid the interpretation and use of achievement test results? Do you remember them?

There are three main evidences. These evidences are (1) content-related evidence, (2) construct-related evidence, and (3) criterion-related evidence. Two of these evidences, construct-related evidence and criterion-related evidence rely on correlation coefficients.

Construct-related evidence is based on constructs which are traits and attributes which are not physically seen but observed through behaviours. In education, constructs include reading comprehension, mathematical reasoning, creativity, and motivation.

How are these constructs measured? There are several methods of measuring these constructs but one of them uses correlation. With the correlation method, the scores on the instrument measuring the construct of interest are correlated with scores on instruments which are **known** to show such constructs. If both of them yield a strong or high correlation coefficient, then the former instrument is also measuring the construct of interest.

For example, Mr. Mensah has produced an instrument to measure achievement motivation. His instrument is called Mensah Motivation Inventory (MMI). To know whether MMI really measures motivation, we have a group of students to respond to items on MMI and an internationally known and well-established instrument that measures motivation like achievement motivation inventory (LMI) (Schuler and Prochaska, 2001). The scores on both instruments are then correlated. If a strong or high correlation is obtained, then it implies that MMI is also measuring achievement motivation.

Criterion-related evidence produces information to determine whether two measures are related. This evidence identifies a predictor and a criterion, for example, period of hours studying (predictor) and a score on a quiz (criterion). Another example is the use of calculators (predictor) and performance in a mathematics examination (criterion). You may have noticed that one set of scores are called predictors and another called criterion. In criterion-related evidence, the

predictors are used to predict performance on the criterion. The use of the correlation coefficient is important here. To obtain the evidence, the correlation coefficient is obtained between the scores on the predictor and the criterion.

6.4 It is useful for prediction

Correlation puts the teacher in a position to predict the future performance of a student. An established relationship between two subjects is often used as the basis for predicting performance, **but not with 100% certainty**. For example, if those with aggregate 6, from WASSCE have been found in the University of Cape Coast to be obtaining First Class degrees, then it can be predicted that any one with WASSCE aggregate 6, would do well in the University.

In a study, Steward and Al-abdula (1989) reported relationships between critical thinking and academic performances for 237 undergraduates in the United States of America. These researchers indicated that, in general, students who scored high on the Watson-Glaser Critical Thinking Appraisal, WGCTA also had high Grade Point Averages (GPAs). In this case, WGCTA was used to predict Grade Point Averages (GPAs).

In another study in the USA, Gadzella, Baloglu, and Stephens (2002) found that Educational Psychology grades predicted GPAs better than the WGCTA. This information was very valuable to those who teach Educational Psychology because Educational Psychology courses provided the foundation for other Education courses.

6.5 It is useful for research purposes

Research plays very important roles in education and as teachers you are also expected to undertake research.

What role does research play in education? Close your module and write down two important roles research plays in education. When you are done, open your module.

Now compare what you have written with the roles listed below.

1. Researches in education provide information that adds to knowledge about educational issues.
2. Researches in education suggest improvements that must take place in educational practice.
3. Educational research provides information that is used in the formulation of educational policies.

These roles are very important and imply that research in education should be carried out as a regular activity. There are several types of research and one of them is correlational research.

Correlational research involves collecting data in order to determine whether a relationship exists between variables. Correlational research has two purposes. It serves to explore relationships between variables and uses these relationships to make predictions. Thus correlational studies may either be exploratory/relationship studies or prediction studies. The major tool that is used in correlational studies is the correlation coefficient.

Now I want you to try an assignment. Spend some time to think about school and educational variables. List two topics you might want to research into that will use the correlation coefficient as the statistical tool for data analysis. Bring the topics to FTF for discussion.

Congratulations for completing this session and the Unit. In this session, you learnt about the uses of correlation in education. You learnt that correlation can be used for selection and placement, it used to compute the reliability estimate of test scores and the validity of assessment results. It also aids in prediction as well as research in education.

Self-Assessment Questions

Exercise 7.6

1. One advantage of correlation studies is that it permits the
    A. change of scales of measurement.
    B. development of continuous variables.
    C. prediction of performance.
    D. study of cause and effect relationships.

2. In Steward and Al-abdula (1989) study of relationships between critical thinking and academic performances, a student who scores low in critical thinking will score
    A. high in academic performance.
    B. low in academic performance.
    C. the same marks in academic performance.
    D. zero in academic performance.

3. Gadzella, Baloglu, and Stephens (2002) did a study in the USA and found that Educational Psychology grades predicted GPAs better than the WGCTA. Which variable is the criterion?
    A. Educational Psychology
    B. GPA
    C. USA
    D. WGCTA

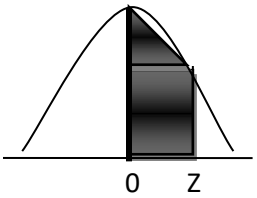
4. An estimate of the split-half reliability coefficient is obtained by correlating the scores on the first administration of the test and the second administration of the test.
    A. True
    B. False

5. Content-related evidences use correlation coefficients to estimate the validity of assessment results.
    A. True
    B. False

6. The relationship between the continuous assessment scores and final examination scores in a College of Education is 0.95. This implies that continuous assessment scores can be used to replace final examination scores.
    A. True
    B. False

APPENDIX A

Standard Normal Distribution Table



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2157	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4215	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4492	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998.	0.4998	0.4998.	0.4998.	0.4998	0.4998
3.6	0.4998	0.4999	0.4999	0.4999	0.4999.	0.4999	0.4999.	0.4999.	0.4999	0.4999

ANSWERS TO SELF ASSESSMENT ITEMS

UNIT 1

Exercise 1.1

1. B
2. C
3. D
4. A
5. B
6. B

Exercise 1.2

1. C
2. C
3. D
4. B
5. D

Exercise 1.3

1. B
2. B
3. B
4. A
5. D
6. B
7. B
8. B
9. A

Exercise 1.4

1. B
2. B
3. A
4. B
5. A
6. A

Exercise 1.5

1. C
2. B
3. D
4. A
5. D
6. B

Exercise 1.6

1. A
2. A
3. D
4. B
5. A
6. A

UNIT 2

Exercise 2.1

1. B
2. A
3. A
4. A
5. A
6. B

Exercise 2.3

1. B
2. A
3. C
4. A
5. A
6. B

Exercise 2.4

1. B
2. A
3. A
4. D
5. A
6. A

Exercise 2.2

1. (a) Categorical
   (b) 122
2. B
3. B

.

Exercise 2.5

1. C
2. C
3. A
4. B
5. A

Exercise 2.6

1. B
2. C
3. B
4. B
5. A
6. A

UNIT 3

Exercise 3.1

1. C
2. B
3. A
4. A
5. (i) Not applicable educational background is not a numerical data
   (ii) 122

Exercise 3.2

1. C
2. A
3. B
4. D
5. A
6. B

Exercise 3.3

1. (i) 25%
   (ii) 40 minutes
   (iii) Cannot be determined
   (iv) 25%
2. (i) 75%
   (ii) 75%
   (iii) There was an overall improvement
3. A
4. B

Exercise 3.4

1. B
2. C
3. A
4. D
5. B
6. C

Exercise 3.5

1. C
2. A
3. A
4. C
5. A
6. A

UNIT 4

Exercise 4.1

1. C
2. B
3. B
4. B
5. A
6. A

Exercise 4.2

1. 1136
2. 863
3. 650
4. 139808
5. 66.9
6. 3.18

Exercise 4.3

1. B
2. B
3. D
4. *C*
5. B

.

Exercise 4.4

1. C
2. C
3. D
4. D
5. D
6. 37

Exercise 4.5

1. B
2. C
3. D
4. C
5. 36.6

.

Exercise 4.6

1. A
2. D
3. A
4. D
5. 33.2
6. 41

UNIT 5

Exercise 5.1

1. B
2. A
3. B
4. B
5. A
6. B

Exercise 5.2

1. 78
2. 88
3. 75
4. 60
5. A

Exercise 5.3

1. A
2. 95
3. 238.24

Exercise 5.4

1. A
2. A
3. B
4. 11.9
5. 14.28

Exercise 5.5

1. 50.2
2. 109.4
3. 35.7
4. Achievement. CV is higher.
5. B
6. B

Exercise 5.6

1. A
2. B
3. A
4. 10.5
5. 12.46

UNIT 6

Exercise 6.1

1. D
2. D
3. 86.4
4. 55
5. 93

Exercise 6.2

1. B
2. D
3. B
4. A
5. C
6. D

Exercise 6.3

1. (i) Stanine 3
   (ii) Stanine 6
   (iii) Stanine 8
2. (i) Stanine 4
   (ii) Stanine 7
   (iii) Stanine 8
   (iv) Stanine 4
   (v) Stanine 9

Exercise 6.4

1. 84%
2. 99.74%
3. 34.13%
4. 84%
5. 47.72%
6. 47.72%

Exercise 6.5

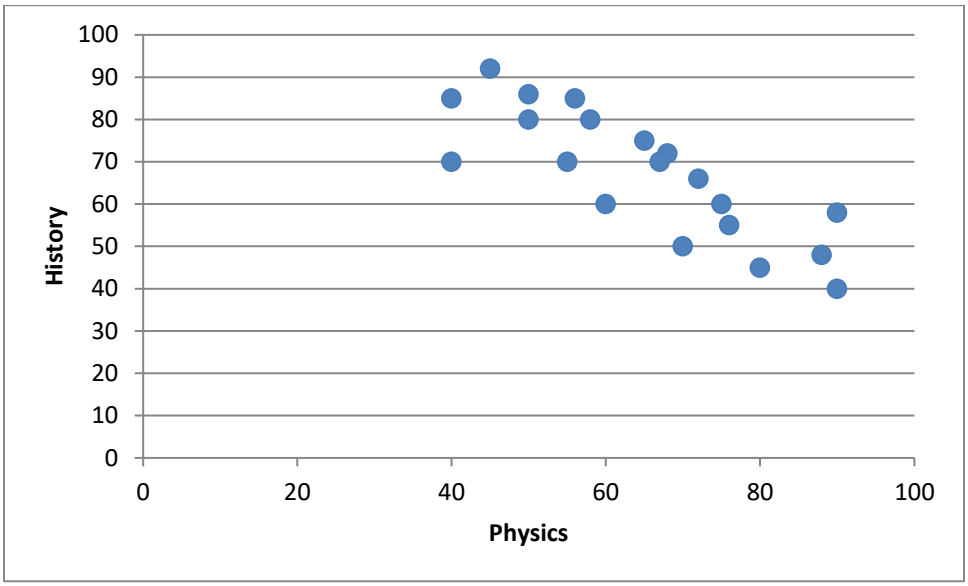
1. 97%
2. 95%
3. 55.96%
4. 97%
5. 47.5%

Exercise 6.6

1. 5
2. 3
3. 10
4. 64
5. 50%

UNIT 7

Exercise 7.1



Exercise 7.2

1. B
2. A
3. −0.6
4. A
5. C
6. C

Exercise 7.3

1. −0.728

There is a strong negative linear correlation between Physics scores and History scores. A high performance in Physics will likely go with a low performance in History.

Exercise 7.4

1. 0.69

High positive linear relationship

2. −0.625

Moderate negative relationship

Exercise 7.5

1. 0.025

Very weak positive linear relationship

2. 0.278

Weak positive linear relationship

Exercise 7.6

1. C
2. B
3. B
4. B
5. B
6. A

GLOSSARY

Arithmetic Mean (or the Mean) is the sum of the observations in a set of data divided by the total number of observations.
Class boundaries. These are the exact or real limits of a class interval.
Class interval. The range within which a group of scores lie.
Class limits. These are the end points of a class interval.
Class mark: The midpoint for each class interval. They are obtained by adding the two class limits and dividing the result by 2.
Class size/class width. The number of distinct or discrete scores within a class interval.
Coefficient of determination. It is the square of the correlation coefficient and the proportion of the variance in Y accounted for by X.
Coefficient of variation. The ratio of the standard deviation to the mean. It is often expressed as a percentage, so that the value is multiplied by 100.
Continuous variables. Variables that have values which in theory; assume any value on a number line between two points.
Correlation. A study of the extent of the relationship between two variables.
Cumulative frequency. This is the successive sum of the frequencies in a frequency distribution table starting from the frequency of the bottom class.
Descriptive statistics. The use of a single number or summary data to describe the group.
Discrete variables. Variables that have values, which in theory, assume only certain distinct values or whole numbers on a number line.
Frequency distribution. Any arrangement of data that shows the frequency of occurrence of different values of a variable.
Industrial statistics. Numbers or values collected on variables in an industry.
Inferential statistics. The use of information or data from a small group called a sample to make conclusions or generalizations about a much larger group called a population from which the sample is taken.
Interval scale. The classification of the elements that constitute a variable into equal units such that elements with the same characteristics are in the same category and differences between intervals are equal with a relative zero.
Median. A score for a set of observations such that approximately one-half (50%) of the scores are above it and one-half (50%) are below it when the scores are arranged sequentially
Mode. The number in a distribution that occurs most frequently.
Nominal scale. The classification of the elements that constitute a variable into two or more categories and each category or group is assigned a number for the purpose of identification.
Open-ended classes. These are classes with a value at one end, either at the beginning or the end and a description at the other end. These intervals are put either at the top or bottom of a frequency distribution table.

Ordered variables. Variables where the attributes differ in magnitude along a quantitative dimension.

Ordinal scale. The classification of the elements that constitute a variable into ranks in terms of the degree to which they possess the characteristic or attribute of interest.

Percentile Ranks , denoted PR, are based on percentiles. They are the percentage of cases falling below a given point on the measurement scale. It is the position on a scale of 100 to which an individual score lies.

Percentile ranks. Percentage of cases falling below a given point on a measurement scale of 1 - 100.

Percentiles. Points in a distribution below which a given percent, P, of the cases lie and they divide a distribution into 100 equal parts.

Quartile Deviation. Also known as the semi-inter quartile range. It is half the distance between the first quartile (Q_1) and the third quartile (Q_3).

Quartiles. Individual scores of location that divide a distribution into 4 equal parts such that each part contains 25% of the data.

Range. The difference between the highest (largest) and the lowest (smallest) values in a set of data.

Ratio Scale. The classification of the elements that constitute a variable into equal units such that elements with the same characteristics are in the same category and differences between intervals are equal with an absolute zero.

Relative frequency. It is the ratio of the frequency of a class to the total frequency..

Reliability refers to the consistency of achievement test results over time. It is the degree to which achievement test results are the same when (1) the same tasks are completed on two different occasions by the same group of people, (2) different but equivalent tasks are completed on the same or different occasions by the same group of people, and (3) two or more raters mark performance on the same tasks.

Scales of measurement. The categorization of variables or numbers according to specific properties.

Standard deviation. The square root of the mean of the squares of the deviations of the scores from the mean of the distribution.

Standard scores. indicate the number of standard deviation units an individual score is above or below the mean of each group. They represent an individual score that has been transformed into a common standard using the mean and the standard deviation

Stanine. It comes from Standard Nine. It is a method of scaling test scores on a nine-point (1, 2, 3, 4, 5, 6, 7, 8, 9) standard scale with a mean of 5 and a standard deviation of 2.

Statistics. The body of numbers or data collected in any field of endeavour. It is also the collection of two or more values that provide summary information on a set of data. One value is a ‘statistic’ (singular), with two or more values, statistics (plural. It is also defined as the study of methods and procedures used in collecting, organizing, analyzing, and interpreting a body of numbers for information and decision making.

T standard score (T-score). Raw scores that are transformed to a mean of 50 and a standard deviation of 10 and based on the z-score.

Unordered variables. Variables where the attributes are classified into two or more mutually exclusive categories that are qualitatively different.

Validity. It is the soundness and appropriateness of the interpretation and use of achievement test results.

Value. An assigned number or label representing the attribute of a given individual or object or a group of individuals or objects.

Variable. Any characteristic or attribute of an individual or object that can take on different values.

Variance. The mean of the squares of the deviations of the scores from the mean of the distribution.

Vital statistics. Measurements of the bust, waist and hip of a person (but mostly women). It also refers to government records on number of births, deaths, marriages and divorces in a country.

Z standard score (Z-score). The number of standard deviation units an individual score is above or below the mean of a group. It is used to transform raw scores into a common standard using the mean and the standard deviation of a set of observations.

REFERENCES

Borg, W. R., Gall, J. P. & Gall, M. D. (2006). *Educational research. An introduction.* White Plains, NY: Longman.

Etsey, Y. K. A. (2010). *Educational statistics*: Mimeograph. University of Cape Coast, Cape Coast.

Ferguson, G. & Takane, Y. (1989). *Statistical analysis in psychology and education.* New York: McGraw Hill Co.

Fraenkel, J. R. & Wallen, N. E. (2000). *How to design and evaluate research in education.* Boston, MA: McGraw Hill Co.

Gadzella, B. M., Baloglu, M., & Stephens, R. (2002). *Prediction of GPA with educational psychology grades and critical thinking scores*. Education, Spring, 2002

Glass, G. V. & Hopkins, K. D. (2008). *Statistical methods in education and psychology.* Boston, MA: Allyn and Bacon.

Gordor, B. K. & Howard, N. K. (2000). *Elements of statistical analysis.* Accra: City Printers.

Hays, W. L. (1994). *Statistics*. New York, NY: Holt, Rinehart and Winston.

Nsowah-Nuamah, N. N. N. (2005). *Basic Statistics.* Accra, Gh.: Acadec Press.

Pagano, R. R. (2013). *Understanding statistics in the behavioural sciences* (10th Ed.). Belmont, CA: Wardsworth Co.

Schuler, H. & Prochaska, M. (2001). *Leistungsmotivations-Inventar (LMI).* Göttingen: Hogrefe.

Steward, R., & Al-abdulla, Y. (1989). *An examination of the relationships between critical thinking and academic success on a university campus*. ERIC Document Reproduction Service No ED 318936.

Tamakloe, E. K., Atta, E. T., & Amedahe, F. K. (1996). *Principles & methods of teaching*. Accra: Black Mask Ltd.