

Applied Statistics

Statistical Methods I



National Open University of Nigeria

2006

CONTENTS

DESCRIPTIVE STATISTICS

- Unit I: The nature of data/scientific observations
- Meaning of data/observations
 - Meaning of statistics
 - Purposes of statistics
 - Types of statistics: descriptive and inferential
 - The nature of statistical methods
- Unit II: Basic concepts in statistics
- Population
 - Sample
 - Parameter
 - Estimates of statistics
 - Measurements
 - Errors of measurement
- Unit III: Statistical notations/shorthand
- Unit IV: Measurement scales:
- Nominal, ordinal, interval and ratio
 - Types of variables: discrete and continuous
 - Parametric versus non-parametric
- Unit V: Organisation and presentation of data:
- i. Organisation of data
 - ii. Frequency table
 - iii. Composite frequency table
 - iv. Pictograms or ideographs
- Unit VI: Graphical representation of data:
- Bar graph
 - Pie chart
 - Histogram
 - The frequency polygon
 - Cumulative frequency curve
 - Cumulative percentage curve
- Unit VII: Measures of Central Tendency
- The Mean
 - The Median

- The Mode
- Unit VIII: Measures of variability/dispersion I
- The range
 - The quartile deviation
 - The deciles
 - The percentiles
- Unit IX: Measures of variability/dispersion II
- Mean deviation
 - Variance
 - Standard deviation
- Unit X: Shapes of curves:
- The normal curve
 - Properties of a normal curve
 - Skewness
 - Kurtosis
- Unit XI: Some Measure of Association and Agreement I
The concept of correlation
The scatter grams:
Bivariate frequency distribution
The Pearson product moment correlation coefficient
- Unit XII: Some measures of Association and Agreement II.
- Types of values and correlation methods
 - Spearman-Brown Rank-order correlation coefficient
 - Point-biserial correlation coefficient.
- Unit XIII: Standard scores:
- The Z-scores
 - The T-scores
 - The stamine scores

INTRODUCTION

Edu 701 statistical methods I is a one semester course for all post graduate students pursuing masters' degree in education at the National Open University of Nigeria (NOUN). It can serve as a reference material for students in other schools or doing research in other fields. It is a three credit course which is compulsory for all education students at the masters' level.

The course will consist of 13 units which include the nature of data/scientific observations, basic concepts in statistics, statistical notations/shorthand, measurement scales, organisation and presentation of data, graphical representation of data, measures of central tendency, measures of variability/dispersion, shapes of curves, some measures of association and agreement and the standard scores. The material has been developed to suit learners in Nigeria by using examples from the local environment.

The course is designed for people who have earned a professional qualification in education. Most of the teachers would have been teaching for some time or would have been aspiring for leadership positions, in their various places of work. Others would have been in management/leadership positions as HODs, principals, supervisors etc where they will be expected to:

- i. Read, comprehend and interpret technical papers, reports or records,
- ii. Go beyond records of raw scores assigned to students scripts or answer papers in test situations
- iii. Present, interpret and discuss the general characteristics of students from records of their performances given in numerical figures,
- iv. Compare two or more groups of students
- v. Predict outcomes and interpret/draw inference sometimes from large amount of data.

WHAT YOU WILL LEARN IN THIS COURSE

The overall aim of Edu 701, statistical methods I is to introduce you to descriptive statistics. During the course you will learn the meaning and types of statistics; types of data, scales and variables; organisation. Presentation and representation of data using tables, graphs charts etc. you will also learn how to describe the data using different methods or measures such as the central tendency, variability, association etc. You will in addition learn the types of curves and their properties; and how to transform raw scores into standard scores.

Statistics as a course is very necessary for you because there is nothing you will do in education which does not require your knowledge of it. Indeed there is no human

activity which is devoid completely of the knowledge of statistics. It is important in the home, in the farm, in the office in the market and in all aspects of life. This is why you must take it with every seriousness due to it.

COURSE AIMS

The main aim of the course is to introduce you to descriptive statistics and give you the understanding of how to present and represent your data and to describe your observations in data form.

This will be achieved by aiming to:

- Introduce you to the nature and types of data
- Outline the importance of data and statistics
- Compare and describe large groups of data in a concise statistical language
- Predict statistical out-come from a number of statistical observations.

COURSE OBJECTIVES

Each unit of the course has specific objectives which are included at the beginning of the unit. You are required to read them before you start working through the unit. You should always refer to them as you work through and at the end of the unit to check your progress. However the objectives of the course are as follows:

On successful completion of the course you should be able to:

1. Describe data and how to handle them.
2. Explain the concept of statistics and its importance.
3. Explain the types of statistics and their applications
4. Organise and represent data using various means
5. Present data in tables, graphs, charts etc
6. Compare data using the measures of central tendency
7. Describe data using the measures of dispersal
8. Describe data using the measures of association
9. Explain the properties of curves
10. transform raw scores to standard scores
11. Demonstrate how to compute using different formulae in descriptive statistics.

COURSE MATERIALS

These include: course guide, course material, text books, assignment files etc. in addition you should have a calculator, a mathematical set, graph sheets and statistical tables.

Study units

There are 13 study units in this course they are:

Unit 1	The nature of data/scientific observations
Unit 2	Basic concepts in statistics
Unit 3	Statistical notations
Unit 4	measurement scales
Unit 5	organisation and presentation of data
Unit 7	measures of central tendency
Unit 8	measures of variability/dispersion I
Unit 9	measures of variability/dispersion I
Unit 10	shapes of curves
Unit 11	some measures of association and agreement I
Unit 12	some measures of association and agreement II
Unit 13	standard scores.

Some of them are deliberately long for easy flow.

The course is designed to last for 15 weeks or a semester. It implies that each unit should be studied in one week. The reference books are listed after each unit. Statistics books are available in the markets and bookshops.

Assessment

Assessment in this course shall be made up of two parts. These are the

1. Tutor marked assignments TMAs which have a total of 40%. At least six TMA as should be submitted out of which the best four will be used for assessment and grading.
2. Examination: the written examination which shall last for three hours at the end of the course will have 60%

Both the TMAs and the examination must be passed at a minimum percentage before you can be successful in the course. The examination will consist of questions which reflect the types of self-assessment exercises and the TMA questions

You should use the time between finishing the last unit and sitting for the examinations for your revisions. Information from all parts of the course will be examined.

COURSE MARKING SCHEME

ASSESSMENT	MARKS
Assignment 1-6	At least six assignments to be submitted, out of which four will be used. At 10% each = 40% of course marks
Final examination	60% of overall course marks
Total	100% of course marks

How to get the most from this course:

Open and distance learning is not the same thing as face to face learning. Therefore, there is no lecturer in ODL. The self-learning material has replaced the lecture. This means that you can study the materials at your own time and place. The self-learning material can do every thing the lecturer can do for you, if you follow it carefully.

Each of the units follows the same pattern. The format ranges from:

- i. Introduction
- ii. Objectives
- iii. The main body
- iv. Conclusion
- v. Summary
- vi. TMA
- vii. References

To work through without any hitch, follow the under-listed practical strategies.

- a. Read the course guide thoroughly
- b. Organise a study schedule
- c. Stick to your study schedule
- d. Assemble the study materials before you start reading.
- e. Go through the introduction and objectives before any unit
- f. Work through the units in sequence as provided
- g. Keep in touch with the study centre and your facilitators
- h. Do your assignments and submit as scheduled
- i. Review the objectives at the end of each unit to confirm that you have achieved them.

- j. On completing the last unit, review the course and prepare yourself for the final examination.

If you run into trouble contact your tutorial facilitator at the study centre or contact the course co-ordinator of the course at the Headquarters of National Open University of Nigeria, Victoria Island, Lagos. Note that both the facilitator and the co-ordinator are there to help you. Do not hesitate to call and ask them to help.

TUTORS AND TUTORIALS

Tutorials are provided in support of this course. You will be notified of the dates, time and locations of these tutorials, together with the names and phone numbers of your tutorial facilitator and the course-co-ordinator as soon as you are registered with National Open University of Nigeria at your study centre. Do not hesitate to contact your tutor or the course co-ordinator if you do not understand any part of the study units or the assigned readings; or you have difficulty with the self-tests or exercises; or if you have a question or problem with an assignment, tutor's comments on an assignment or with the grading of an assignment.

You should try to attend the tutorials regularly. This is the only way you can have face to face contact or interaction with the facilitator who is there to answer your questions.

SUMMARY

Edu 701: Statistical Method I is one of the two courses you will work through in your programme.

Edu 701 Is designed to teach you descriptive statistics upon the completion of this course you will be able to answer such questions as:

- What is statistics
- What are the types of data dealt with in statistics
- What are the purposes of statistics
- What are the types of statistics
- Why do we use samples instead of population
- What are the measurement scales
- What are the variables in statistics
- How do you organise data in statistics
- How do you present data
- How do you re-present data
- How does bar chart differ from histogram
- What are the measures of central tendency
- What are the measures of dispersion

- How does a normal curve differ from skewed curve
- What are the measures of agreement
- Why do you convert raw scores to standard scores

If you have completed the course successfully you would have been equipped with the basic knowledge of descriptive statistics. This means that you can answer even more questions that are given above. You are also equipped to do some arithmetic which you can do easily with your calculator.

We wish you success with the course, we hope you will find it very interesting. Enjoy your programme at National Open University Nigeria. We wish you every success in your future.

Unit 1

NATURE OF DATA/SCIENTIFIC OBSERVATION

1.0 INTRODUCTION

Welcome to, perhaps, your first course in statistics and statistical methods. This first unit introduces you to raw material you would be working with, data, how they are derived and how they are used in our daily lives. The relationship between data and statistics and the various types of statistics are then introduced. The approach to the statistical methods course is then presented.

2.0 OBJECTIVES

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

1. define statistical data and statistics;
2. conceptualise how data are derived from narrative description/information;
3. state the purposes of statistics and statistical methods
4. identify various types of statistics
5. distinguish between statistical methods and educational statistics
6. appreciate the importance of data in our everyday life

3.1 THE CONCEPT OF DATA/SCIENTIFIC OBSERVATION

The raw material of all statistical works is data whether you are looking at the enrolment figures in our school system, the number of participants from various states in a national conference; the number of successful candidates in a public examination or the salaries of teachers, you are presented with a large amount of information, often in the form of numerical figures.

These figures are referred to as data (from singular-datum). In a world that is increasingly demanding for evidence to support decision making, the provision of numerical data, or what is sometimes referred to as empirical data/evidence, help to strengthen your case.

In the sciences, all observations are presumed describable either quantitatively or qualitatively. Qualitative observations, such as gender classification, are sometimes represented by figures (such as 0 for female, 1 for male) to make the record easy for processing. Thus, all scientific observations are also referred to as data.

Let us quickly look at an example of how data is used in our daily lives.

A restaurant has prepared three different menu for lunch for a group of people attending a two day conference.

Day I There was jollof rice eba and fried yam and plantain. As part of the preparation of the organisers for the group lunch, participants are required to write their names and indicate in one of three columns their preferred lunch. By the time full attendance was taken, the following observation was made:

31 participants indicated preference for jollof rice
 7 participants indicated preference for eba
 18 participants indicated preference for fried yam/plantain

This observation can be translated into a table, as in table 1

Table 1 participant preference for lunch (day 1)

Number indicating preference	Jolof Rice	Eba	Fried Yam/Plantain
	31	7	18

Table 1 represents data of observations on this occasion

Exercise 1.1

Identify two instances in which you have come across data today.

So the use of data is pervasive, we use data every day in our life, although we do not always describe them as data, note that for brevity, we can use A, B and C to represent the variety that was available.

A	B	C
31	7	18

If we want to know how many participants would be in the restaurant, we add up $31 + 7 + 18 = 56$, thus 56 participants must be catered for.

The supervisor can now make further decisions, on the basis of this information, such as how many attendants must be engaged to provide quick services, and how many chairs must be engaged to be provided, the manner in which we have handled this information so far is referred to as statistics.

3.2 MEANING OF STATISTICS

Statistics is the science of handling information, particularly quantitative information (data), it is the science of organising describing summarising and interpreting data to provide concise manageable information for decision making.

Now let us consider further our example in table 1, suppose the following day, the same menu of three varieties A, B and C are offered and the choices this time around are as follows:

Table 2 Participant preference for launch (day 2)

No of persons	A	B	C	Total
	21	17	16	54

Now the supervisor has an opportunity to compare the two sets of data, Day 1 and Day 2. He can do so by again looking at the columns of totals. He can also compare the numbers eating the same food in the two days and make a number of inferences. For instance, he might infer that the two participants who failed to turn up for lunch are dissatisfied with the services. He may want to verify this by seeking further evidence. He could also infer that eba has gained popularity and jollof rice has lost patronage. Again, the data provides him a number of options.

So, in addition to its descriptive function, statistics enables you to examine patterns and compare groups.

3.3 PURPOSES OF STATISTICS

From the foregoing, it may be concluded that statistics helps us to:

- present a large amount of quantitative information
- in an organised way, go beyond a meaningless record of school test results to a meaningful interpretation
- predict how likely an event will occur
- make inferences from observations

- save us time and energy by condensing large amount of information concisely and conveniently in a table.

In addition, there is hardly any discipline today even in education, the arts and social sciences that does not require some level of statistics for its understanding. Research reports in most disciplines are enriched by statistics.

Lastly, one requirement for higher degree in most disciplines is that you carry out and report your own independent research.

School managers are constantly faced with situations in which they have to make inferences from observations, for instance, the proprietor of a private school took a week's attendance and found the following:

Table 3: A week's school attendance in school A.

Class	Week days						No Class	in
	Monday	Tuesday	Wednesday	Thursday	Friday			
Primary one	38	39	38	36	31	40		
Primary two	35	34	33	32	30	35		
Primary three	45	41	40	43	40	45		
Primary four	36	35	38	39	18	40		
Primary five	41	42	42	40	36	42		
Primary six	28	26	29	21	27	31		

If she makes similar observations over several weeks, the proprietor can deduce from these that there's hardly full attendance in a school day, she may also deduce that attendance is least toward the week end.

You do not always know what type of research you will carry out, whether it is a documentary analysis, a historical analysis or an experimental study.

You will however find the knowledge of statistics most valuable in your readings, conception design and analysis of your work.

Exercise 1.2

Locate a copy of the latest scholarly journal in your subject of specialisation. Read the first published articles in the journal identify and write down every statistical jargon you come across in that article.

You will find some of them esoteric and extremely difficult to comprehend. You should not worry, for we shall be coming across some of them and indeed learn to handle them in this course.

3.4 **TYPES OF STATISTICS**

The purposes listed in section 1.3 give rise to three types of statistics, namely:

- i. descriptive statistics which refers to the manner of presentation of large amount of data,
- ii. correlation statistics which allows you to examine patterns, compare groups and predict future events from extrapolation,
- iii. inferential statistics which as the name implies, allows you to draw out facts that are not immediately available/visible from the data presented.

3.5 **THE NATURE OF STATISTICAL METHOD**

You will notice that this course is described as statistical methods as opposed to statistics which we have been discussing indeed, there is a complementary statistical methods II course which you may be required to take during the second semester. They are so called in order to draw attention to the fact that the emphasis in this course is not to make you statisticians. Rather, you are expected to grasp the method statisticians use to process information, the circumstances in which one approach is used in preference to another and the limitation accompanying any so-called hard/statistical facts. You will be expected to pay special attention to the procedure, rather than the theory of statistics.

A distinction should also be made between statistical methods and perhaps, the more familiar educational statistics. Whereas statistical methods is concerned with the manner in which statistics are derived, the latter is concerned with school related statistics such as enrolment, teacher to pupil ratio, learning event, female/male participation urban/rural participation and

so on. This information is important to educational planning and management.

4.0 CONCLUSION

In this unit you should have learnt the concept of data, as the foundation of all statistical analyses and how data are derived from repeated events, information or scientific observation. You should also have learnt to define statistics, the purpose of statistics and types of statistics.

The statistical methods course is concerned with how to handle data derived in educational contexts clearly, several of such contexts exist in education and other social sectors. The statistical methods course is applicable in a wide variety of disciplines. These contexts and related concepts shall be examined in the following unit.

5.0 SUMMARY

In this unit, we defined data as the reduction to numerical figures, a large body of information and that data are the basis of all scientific investigation and analysis. The decision maker who has no data to back his/her claim is unlikely to convince many listeners. We also identified three types of statistics, namely: descriptive, co-relational and inferential. We are aware that not all users of statistics want to be statisticians, just like not all drivers of motor vehicles would want to be repairers to know how the motor works so that when you are faced with a problem, you would know what to report to the repairer and how to go about it. Statistical methods are a “how-to” course.

6.0 TUTOR MARKED ASSIGNMENT

1. What do you understand by the word statistics?
2. What are the purposes for your study of statistics?
3. What are the types of statistics?
4. Briefly explain the word data in relation to statistics

Unit 2

BASIC CONCEPTS IN STATISTICS

1.0 INTRODUCTION

In unit 1, we defined statistics as the science of handling a large body of numerical information (data). You also learnt that data may be derived from several contexts and would be handled according to the context.

2.0 OBJECTIVES

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

- i. Distinguish between population and sample
- ii. Distinguish between parameter and estimates of statistics
- iii. Distinguish between the concept of observation and measurement
- iv. Relate these concepts in statistical methods.

3.0 BASIC CONCEPTS OF POPULATION AND SAMPLE

We have already asserted that data are the fundamental raw material of statistics and data are usually numerical information about objects, events, etc. We cited the example of a school proprietor who monitored and took a week's attendance in the private school, clearly the proprietor can come to some conclusion only for that week, from the observations or data generated in table 3. If she wanted to know the attendances throughout the school term or school year, she would have to take attendance every week during the whole term of school year, as the case may be. Thus, the one week attendance is only part of a large body of possible information. When a statistician deals with whole rather than part information about an object, events or phenomenon he is dealing with a population, population is all possible objects, beings, events incidences with the same characteristics that are the focus of the observer.

We are all conversant with the national census. What is the principal focus or objective of a national census exercise? The population of a country, the population of Nigeria in 1991 was, according to the National Population Commission (). This means that as of the time the national headcount was done, there were human beings (males & females, foreigners, rural and urban dwellers) in the political entity called Nigeria.

Note that in this case we were looking only at the number of human beings, not animals or houses. Note also that human beings can still be categorized

as males and females, nationals and expatriates, rural and urban dwellers etc. native language speakers – Hausa, Igbo, Efik and other language speakers.

When a population is sub-divided into well defined classes or categories such as gender (male & female) or habit (rural and urban) we say that it is desegregated by the characteristic which defined the classes or categories.

Now, think if we are interested in taking decisions about objects (living or non-living) events (such as school attendance) etc, it is not always very important to first count all the possible objects or events that is our focus. For relatively small populations, it is possible to do a head count. But for large populations, it is impossible to do a head count every time we want to know the population or some characteristics about it.

In such situations we can make do with a sample, a sample is a selected group from the population. Any subset of the population may be called a sample. For instance, the Igbo's are Nigerians, meaning that the Igbo language native speakers come from the population of Nigeria. However, not all Nigerians are native speakers of Igbo. There are other native languages. Igbo speakers constitute a sample of Nigerians similarly Yoruba and Nupe speakers constitute two separate samples of Nigerians. Suppose a point is to be made about the Nigerian culture to someone who is unfamiliar with the situation on ground in Nigeria, if the point is made only on the basis of the knowledge of the Nupes in Niger State of Nigeria, the point may be in error. Why?

The point would be in error because although the Nupe culture is Nigerian, not all Nigerian cultures are to be found with the Nupes. If a point is to be made about the Nigerian culture, then we must select or choose a sample that would be representative of Nigeria i.e. one in which all the characteristics of Nigerian cultures are present.

Statisticians make inferences from representative samples. It is the task of the statistician to find an appropriate method to draw its sample from a population.

Exercise 2.1

1. State the populations which the following samples represent.
 - i. Monday to Friday in the week
 - ii. All students in JS II purple house in FGGC Zaria
 - iii. The salary bill of teachers in a rural private school

- iv. Scores in WAEC School Certificate examination in King's College, Lagos
 - v. The height of 1,500 Birom men.
2. How would you improve on the representativeness of the populations which the samples in 1 above represent?
 3. What population does all SS1 students in Government Day Secondary, Kiyiye represent?
 4. Why is it not technically correct to say, the "population of my school" is 4,400 when you are referring to the total students' intake? What would you suggest as a better alternative of describing all the students in a school?

3.2 PARAMETER AND STATISTICS

The statistician is interested in samples preferably representative samples, when she cannot reach the population. Note however that whenever census or voters registration exercises are being conducted, names of individuals which are descriptions that identify the individuals are written down. But when reports about census or voters are given, they are given as numbers (numerical information). You have already learnt that numerical information or data are the raw material with which statisticians work. So, whether the statistician has access to the population or a sample from the population he/she would always reduce the information that is his/her focus to numerical information. The statisticians would talk about total number, number of males allowed, maximum height of residential houses, body weights of a herd of cattles, the waist circumference of 15 year old girls, the scores in a test etc, in all such cases, the interest is in the numerical properties of the object or phenomenon that is the focus of the statistician or researcher and statisticians and data-based researchers usually focus on those characteristics of their populations or samples that are measurable.

If the numerical property which is descriptive of the population is available or measurable, then the statistician or researcher would say she has a parameter. Strictly speaking therefore, census figures such as the total number of households in a ward, are parameters.

We have already stated that we cannot always reach a population. If a decision is to be made from a representative sample, such a sample based decision can only approximate the reality of population. A sample is not the population. It approximates a population. An approximate is as good as the manner in which a sample is drawn. The more the various characteristics of a

population are represented in a sample, the more closely decision made from the sample approximate the true position of the population.

For instance, suppose you visit a wall-fenced boarding girls' secondary school for the first time and at the entrance you are confronted with 20 girls all wearing white blouse top over green skirt. If you have not seen any other, you might conclude that "white blouse over green skirt" is the school uniform. In other words, you can make a decision/conclusion about the school uniform from a sample of 20 girls.

Suppose that soon after you saw the girls, the school bell rings for a break and all the other girls are wearing white blouse over brown shirt, you will have to admit your error of judgement and change your conclusion.

It is extremely useful to ensure that samples are drawn in such ways that would represent the characteristics associated with the population of interest.

The numerical properties of a representative sample would give you a good estimate of the numerical properties of the population. Thus if we get the average of take-home pay of principals of five Federal Unity Schools well chosen across the nation you will get an amount which is approximately the average take-home pay of all principals of Federal Unity Schools. If you want to make a statement about average income of principals of Federal Unity Schools, you do not have to know the income of all of them. If you take the average of the income of a representative sample of such principals, the figure you get will be very near correct or approximate average income of all such principals. Note that the emphasis is "it approximates the information you are seeking". So, the numerical properties of a representative sample are approximately the numerical properties of its parent population (i.e. the population from which the representative sample is drawn).

When a statistician or researcher knows or can determine a numerical property of a sample, he/she says he/she has an estimate of statistic. When you have several such estimates from a sample, you say you have estimates of statistics or simply statistics. So statistics are to a sample what parameters are to a population.

Let us quickly distinguish between statistics with capital letter S, which defines the subject a science discipline dealing with handling numerical data and statistics with small letters, which are numerical estimates of a population derived from its representative sample.

3.3 MEASUREMENT

Measurement is a basic process in science, all observations whether it is qualitative or quantitative must be measured to some degree of accuracy. If an observation is qualitative, such as would be observed by colour change at the end point of a titration experiment in a chemistry laboratory, the chemist would like to know how much of the acid is titrated to what quantity of base before the change of colour. Note that the change of colour is an observation and that for our observation to be scientifically useful it must be transformed into measurement. Similarly tailors would measure the circumference of your waist line to make your trouser waist fits. However, experienced tailors can often estimate the trouser length of their clients by observation. After sewing clothes for a long time, you can observe a client and simply “guess” his height and trouser length.

Note that different objects, events or phenomena require different measuring instruments or measuring scales. The tape or ruler is useful in measuring lengths. The weighing balance is used to measure weights. The semester is used to measure an academic year. Intensity is used to measure the colour of light. The degree of agreement is used to measure perception or attitude.

Thus any object, event or phenomenon with a scientific study must have some measurable property/quality. It is these measurements that give rise to the data which the statistician uses for his/her analysis. Thus, performance in a test is measured (scored) against the criteria defined in the marking scheme.

ERRORS OF MEASUREMENT

Tailors who are regularly involved in measurements would tell you that no two persons who measure an individual’s trouser length using the same tape would give exactly the same values, if they really want to be accurate. Similarly, two teachers marking an essay are unlikely to give the same score to the writer even when the teachers use the same marking scheme.

Besides, if an individual measures the same thing over and over again using the same instrument, he/she would begin to notice slight differences if you take the record of your blood pressure under normal situation for many days you are bound to find slight variation. Yet you could still be described as normal.

The slight variations are described as errors of measurement. Errors in measurement can arise due to:

- i. Inaccuracy in the measurement instrument

- ii. Changing circumstance of what is measured
- iii. The limitations in human judgement

Whatever the source(s) of error, scientists have observed that every measurement is fraught with some degree of error. In other words, scientific data are only as accurate as the degree of error which is associated with how they were measured. We say there is a degree of uncertainty associated with every measurement. This concept would become very important later in this course, as we work towards using statistics for decision making.

Exercise 2.2

- i. Using a measuring tape or ruler, measure the length and breadth of your room and record them. Let the other members of your class measure the same room and record them.

What did you observe?

- ii. List all possible sources of the variation which you may have observed.

5.0 SUMMARY

In this unit, you should have learnt the concepts of population and sample. While population refers to totality of members of a well defined group (objects, beings things, events, measurement, phenomena) a sample is a part there from. A sample is said to be representative of a population if the sample has all the characteristics of its parent population. When numerical properties of a population are known or derived, we say we have its parameter. When numerical properties of a sample are known or derived, we say we have statistics, as opposed to the discipline of statistics spelt with a capital letter S.

Parameters and statistics are derived from some kind of measurement, which in turn gives rise to data. The process of measuring depends on what's being measured. We also learnt that measurement has inbuilt errors. Although errors can be minimized, they cannot be entirely eliminated from measurement. Errors are to be recognised.

UNIT 3

STATISTICAL NOTATIONS/SHORTHAND

1.0 INTRODUCTION

In the two earlier units, you learnt a number of related concepts. You learnt about how scientific observations are transformed into data for the purpose of statistical analysis. You also learnt the concepts of population and sample and the indication that are used to describe the measurements made therefrom.

Since you will be dealing with a lot of numerical data, some of them may have to be handled the way mathematicians handle them. In others, you will quite often have to add, subtract, multiply and divide large number with notations/symbols or shorthand which statisticians use to describe the operations that you may need to carry out with the numbers. We shall restrict ourselves to operation you may have to carry out with samples and their estimates of statistics.

2.0 OBJECTIVES

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

- i. Recognise the symbols used most often for statistical operations;
- ii. Apply them in data handling;
- iii. Interpret combined statistical operations
- iv. Define the arithmetical mean of a set of scores

3.1 In statistics, you unavoidably have to handle large quantities of figures coming from different sources. For example, as a class teacher you would have handled examination results of your class of say 50 students in eight or more subjects, namely, English language, literature, history, geography, mathematics, biology, chemistry, physics, agric science, similarly an educational researcher may have to test his samples in more than one area of focus, such as in achievement, practical works, attitude, aptitude and personality type. Each of these areas would produce a set of scores. In each set, scores would of course vary. Each set of scores is described as a variable. A variable is any characteristic of an object or event which can be measured. Thus, length of an object is a variable, height, weight, time colours and scores in examination are all variables. Two different students may not score the same mark in an English language test. Scores in a test constitute a variable. If you want to discuss or analyse the scores in an English test, it is cumbersome to refer to “the scores in the English test” all

the time. Statisticians devise a shorthand way of making this reference by describing it as X . The letter X is therefore used to describe a set of scores. If you have another set of scores, it is important to ensure that the second set of scores is not confused with the first. The letter Y is used to represent a second set of scores or variable.

If you have a third set of scores, it will be designated with letter Z and so on. Indeed any Roman capital letter may be used to denote a set of scores from one variable.

If capital letters X, Y, Z are used to denote 3 variables, then small letters x, y, z are often used to denote subsets (a small group chosen) from X, Y, Z , (large group of scores).

For instance, if X represents the scores of 50 students in your English class test x represents the scores of 5 10 on any number of students chosen from among the 50 students.

So in summary,

X	denotes the set of scores in one variable
Y	denotes the set of scores in another
Z	denotes the set of scores in a third variable and so on

Likewise,

x	is a subset of X
y	is a subset of Y
z	is a subset of Z

If you have more than 3 groups you can use capital letters, P, Q, R, S, T, U, V , etc. In effect, you are simply assigning names to the variables.

In many statistics textbooks you would find different letters or symbols used to denote a set of scores in a variable. Once you have learnt one system that is convenient to you, the best practice is to stick to one system consistently.

- 3.2 You would find however that certain letters or Roman alphabets are consistently used by many authors to denote particular types of variable. For instance, the letter N is often used to represent the total number of members

So if we have 50 students in a class, we can write $N = 50$ meaning the number of students in the class = 50.

If, as we had learnt earlier, we want to look at the scores of 7 students only out of the 50 in the class, we can write $n=7$ meaning we are concerned with $n(7)$, a smaller number taken out of the larger number (50) in the class. Statisticians often have to deal with samples and populations. If N represents the number in the population, then n is the sample size.

Another letter that is commonly used is the letter f . F is used to denote frequency that is the number of times a particular number has appeared.

If in a class of 50, six students earned the same score, say 57, the frequency (f) of the score of 57 is 6 ($f=6$). As you can see, we are already beginning to learn the language of handling numbers. We are learning to describe large sets of numbers with which we may be confronted from time to time.

3.3 Statisticians often carry out a lot of mathematics operations:

- addition (+) that is add two sets of scores together
- subtraction (-) meaning take away one set of scores from another
- multiplication (X) meaning take multiples of a set of scores and
- division (\div) meaning take a set of scores and break it into parts.

You are no doubt familiar with these operations. You may not be so familiar with the square root operation ($\sqrt{\quad}$): when you are asked to take the square root of a number, say 25, it should be interpreted to mean “what number do you multiply by itself” to get 25? The answer of course is 5, for $5 \times 5 = 25$, therefore the root of $25 = 5$.

It is written as $\sqrt{25} = 5$. Similarly the root of

4	is	2
$\sqrt{4}$	=	2
$\sqrt{9}$	=	3
$\sqrt{16}$	=	4
$\sqrt{36}$	=	6
$\sqrt{49}$	=	7
$\sqrt{64}$	=	8

This process would prove helpful in this course. The reverse operation is described as “take the square of “...” and it is written with the figure 2 as an upper script to the figure you want to square, thus:

7^2	=	49
6^2	=	36
9^2	=	81
10^2	=	100

Exercise 3.1

- (a) What are the square roots of: 121, 100, 400, 265, 39
- (b) What are the squares of 8, 11, 13, 19 and 57?

- 3.4 Let us now try to combine these concepts. Quite often the statistician wants to add one set of scores to another, as would be the case when the continuous assessment (CA) scores are added to the end of year scores in English language, in order to describe the overall year-round performance of a student in one subject. If the set of scores in the CA is denoted as X , the set of scores representing end-of-year examination may be denoted as Y . The operation to add these sets of scores together can be denoted simply as $X + Y$, $X + Y$ means add the set of X scores (CA) to the set of Y scores (end of year) scores.

Clearly, what needs to be done is brief and concise. Everybody who handles statistics would immediately understand what operation needs to be done. If three sets of scores X , Y , Z , are to be added, it is denoted by $X + Y + Z$.

Another very common operation is to add up the set of scores in one variable, if we want to know how much money in our hands in a class, you would have to add up what each and everyone has with him/her. The procedure would be to write down the names of everybody in the class and against each name state the amount of money with him/her. Similarly, you obviously have to add up the scores in each item of an examination to get a total score for a candidate in that exam, when the scores in a group of scores (a variable) is to be added.

Symbol Σ , represents summation.

, Thus, ΣX reads as sum, means add up all the scores represented by X , ΣY means add up all the scores represented by Y .

Let us say X represents the set of the odd numbers between 1 and 10 inclusive, that is 1, 3, 5, 7, 9 and Y represents the set of even numbers (that is numbers divisible by 2) between 1 and 10 inclusive, thus, Y represents 2, 4, 6, 8, 10

For convenience, you may re-write these numbers in columns rather than rows to make them conform with our customary or more familiar approach to adding numbers,

Thus,

X	Y
1	2
3	4
5	6
7	8
9	10
<u>25</u>	<u>30</u>

ΣX means add up all the scores under the column X

ΣY means add up all the scores under the column named Y

$$\therefore \Sigma X = 25$$

$$\Sigma Y = 30$$

You can quickly see how the statistician has reduced and can reduce a whole page of figures concisely into a four or five symbol summary. Note the introduction of the symbol, which simply means "therefore" it should be read as,

If X represents 1,3,5,7 and 9;

Therefore or then, $\Sigma X = 1+3+5+7+9=25$

Similarly,

If Y represents 2,4,6,8 and 10

$$\therefore \Sigma Y = 30.$$

Note also, that between 1 and 10 inclusive, there are five number of numbers and no number has occurred more than once, similarly, between 1 and 10 inclusive, there are five number of numbers and again no number was repeated,

We can state that n of x is 5

n of Y is also 5

In this case, the value of n is the same in both instances. They need not be the same.

Suppose, X represents 0, 1, 3, 5, 7, 9

That is one set of scores starts and includes 0, as would meaningfully be the case if you are adding up cash-in- hand in your class, where you are likely to come across, at least, one person who has no money with him or her.

Note that ΣX still remains 25

But n =6, that is there are now six cases

Note also that the frequency of occurrence of any particular score in the group of scores (X) is 1

Thus, frequency (f) of 9 = 1

Similarly frequency (f) of 7 = 1

Exercise 3.2

Take the example $X = 0, 1, 3, 5, 7, 9$ write down the frequency of occurrence of each scores (f) now take ΣF . What did you observe?

Frequently, a set of scores (Y) may have to be subtracted from a set of scores (X).

It is written as $X - Y$, meaning take the set of scores in X and subtract corresponding scores in Y.

Two sets of scores can be multiplied with each other. It is written as $X \cdot Y$. Here the dot (.) is used to represent the more familiar multiplication sign, in order to avoid confusing it (X) with the X- variable or set of scores. A very special case of multiplication is when you multiply a score by itself as in $4 \cdot 4 = 16 = 4^2$. or a set of scores in which each score is multiplied by itself, as in $X \cdot X = X^2$ $Y \cdot Y = Y^2$. This form of representation of the process is particularly useful, when you want to multiply X by itself and by itself, namely $X \cdot X \cdot X = X^3$. You must learn to distinguish between X^3 as in $3^3 = 3 \times 3 \times 3 = 27$ and $3x = 3 \times x$ which is x multiplied by a factor of 3, as in $3 \times 3 = 9$, $3 \times 4 = 12$, $4^3 = 4 \cdot 4 \cdot 4 = 64$. Similarly, X may be divided by Y and it is written as X/Y .

Exercise 3.3

(i) How would you interpret $E(X - Y)$?

Note that according to the rules of mathematical operations, whenever you see brackets in a combination of mathematical operation, you must first complete the task required in the bracket. In the same manner, the "of", "dot" square or multiplication sign takes precedence over other operations, other than brackets. Division comes next in the order of precedence. The addition and subtraction are done in that order of precedence. Note also that the signs which accompany a numerical figure or its representation are determined by what sign is the product with another.

Another very special case is when you take ΣX and divide by ΣF .

You would have observed that $\Sigma F = N$, total number of scores.

The combined operations $\Sigma X/N$, means add up all the scores in X and divide by the total number of scores. $\Sigma X/N$

The product of this process is described as the Arithmetical Mean or simply, the mean of the scores. The mean has a special significance in statistics. It is the figure which is most often used to represent a large set of figures. You would often learn, "what is the mean performance of your class in a Maths examination?" The mean gives a representative idea of how each person has performed. What is the mean temperature? The response gives an idea of how hot or cold the environment is over the year.

What is the mean salary grade level of primary school teachers? (GL 05).

The mean of a set of scores(X) is represented by \bar{X} , pronounced as \bar{x} .

$$\therefore \frac{\sum X}{N} = \bar{X}$$

Exercise 3.6

- (i) Take $X = 0, 1, 3, 5, 7, 9$ what is \bar{X} ?
- (ii) Take $Y = 2, 4, 6, 8, 10$. What is \bar{Y} ?
- (iii) if $X = 1, 3, 5, 7, 9$ }
and $Y = 2, 4, 6, 8, 10$ } what is $\Sigma(Y-X)/N$?

4.0 CONCLUSION

Remember that our task is to learn to handle large amount of numerical data in concise meaningful ways, which all would understand, statisticians have invented all kinds of notations/shorthand /symbols that would enable you describe data and what mathematical operations can be carried out on them.

Next we can now take a deeper look at the nature of data types, the way they are derived from measurement and the various measurement scales that are used. The nature of the object to be measured, dictates the type of measurement that is done and the scale that is used. The type of data obtained determines the kind of statistical analysis that is appropriate.

5.0 SUMMARY

In this unit, you have learnt how to represent a set of numbers with letter X and another set with Y . you have learnt that technically multiple sets of numbers can be represented by convenient Roman alphabets, P, Q, R, S, T , Or A, B, C, D , etc.

However, some particular alphabets are commonly used in textbooks to denote particular concepts or processes. In order to remain intelligible it is important that you learn to adopt one style and to use that style consistently. Regardless whatever style you adopt, the statistical methods involved, the actual procedures for handling the data remain the same. The mean at all times is the sum of all scores divided by the number of scores. The sum of the frequency of each score is the same as the total number of cases or scores.

$$\Sigma F = N \quad \text{and} \quad \frac{\Sigma X}{N_x} = \bar{X}, \quad \frac{\Sigma Y}{N_y} = \bar{Y}$$

6.0 TEACHERS MARKED ASSIGNMENT

1. What do these symbols used in statistics stand for?

- i. Σ .
- ii. \bar{X} .
- iii. f .
- iv. N .
- v. $\sqrt{\quad}$
- vi. X, Y .
- vii. \div
- viii. $=$
- ix. $+$
- x. Σf
- xi. X^2
- xii. \hat{X}

7.0 REFERENCES, FURTHER READING AND OTHER SOURCES

Indira Gandhi, National Open University (1999)
Statistical Techniques of Analysis. ES 333

Unit 4

MEASUREMENT SCALES.

1.0 INTRODUCTION

All data are derived from some kind of measurement. The scores in a football tournament are measures of how many goals are scored by competing teams. At the end of the tournament, all that you see is a record of performances

Such as: Lion 3 -2 Super Eagles
Enyinba 1 -2 Jets
Shooting stars 0-0 Abiola Babes
Katsina United 0 – 3, 1. Nationale

Note that these scores are really counts of goals scored and that there is no half - goal! Note also that we can talk about total goals scored in the tournament and goal differences but we cannot multiply or divide goals scored.

Compare the scores in a tournament situation with the scores of students in your Physics class. Typically, students are scored over a total possible score (say 20) as determined in the marking scheme. A student who scored 3 out of 20 may know the correct answers but has not responded according to the pre-determined marking scheme.

This type of information, score, is obviously different from the type we got in the football tournament. It must mean that each object, event or phenomenon must have its own peculiar way of scoring it. This unit is concerned with the various scales that are used to measure different variables.

2.0 OBJECTIVES

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

- (i) Distinguish between the four measurement scales,
- (ii) Classify data/scores according to their measuring scales,
- (iii) Define a variable and classify a variable as either discrete or continuous
- (iv) Conceptualize what kinds of analyses are possible with what types of data/scores.

3.1 SCALES OF MEASUREMENT

Measurement refers to the assignment of numbers to object and events according to logical acceptable rules. The numbers have many properties, such as identity, order and additivity. If we can legitimately assign numbers in describing objects and events, then the properties of numbers should be applicable to the objects and events. It is essential to know the different kinds of measurement scales, as the number of properties applicable depends upon the measurement scale applied to the objects or events.

Let us take four different situations for a class of 30 students:

- Assigning them roll nos. from 1 to 30 in no particular, manner or on random basis so long as no student has more than one number;
- asking the students to stand in a queue as per their heights and assigning them position numbers in the queue from 1 to 30;
- administering a test of 50 marks to all students and awarding marks from 0 to 50, as per their performance;
- measuring the height and weight of students and making student-wise record.

In the first situation, the numbers have been assigned purely on arbitrary basis. Any student could be assigned No. 1 while any one could be assigned No. 30. No two students can be compared on the basis of allotment of numbers, in any respect. The students have been labeled from 1 to 30 in order to give each an identity. This type of scale is described as a nominal scale. Here the property of identity is applicable but the properties of order and additivity are not applicable.

In the second situation, the students have been assigned their position numbers in a queue from 1 to 30. Here the numbering is not arbitrary. The numbers have been assigned according to the height of the students. So the students are comparable on the basis of their heights, as there is a sequence in this regard. Every subsequent child is taller than the previous one, and so on. This type of scale is described as an ordinal scale. Here the object or event has got its identity, as well as order. As then difference in height of any two students is not known so the property of addition of numbers is not applicable to the ordinal scale.

In the third situation, the students have been awarded marks from 0 to 50 on the basis of their performance in the test administered on them. Consider the

marks obtained by 3 students, which are 30, 20 and 40 respectively. Here, it may be interpreted that the difference between the performance of the 1st and 2nd student is the same, as between the performance of the 1st and 3rd student. A student getting 0 marks cannot be described as having zero achievement level. No student who has attempted the test would score 0 marks, since the scores are relative to a marking scheme. Similarly, the 2nd student cannot be said to have half the intelligence of the 3rd student, simply because the 2nd has 20 and the 3rd has 40. This type of scale is described as interval scale. Here the properties of identity, order and additivity are applicable. Note also that the scores can have fractional values.

In the fourth situation, the exact physical values pertaining to the heights and weights of all students have been obtained. Here the values are comparable in all respect. If two students have heights of 120 cm and 140 cm, then the difference in their heights is 20 cm and the heights are in the ratio 6:7. This scale refers to ratio scale.

Exercise 4.1

Which of the four measurement scales is being used in the following examples:

- a. Age of students of your school.
- b. The roll numbers assigned to the students of a class.
- c. The rank of students' of SS III of your school in the NECO Examination.
- d. The marks obtained by the students of JSS in Intro Tech.
- e. Year group of SSS students
- f. Letter grades (A, B, C, D, E) in on essay Examination
- g. Amount of money with students in the class.
- h. Nationality of participants in an International Conference.
- i. Courses in which students are registered.
- j. Shoe sizes of JSS 1 students in a Federal Government Girls College.
- k. Number of books on the shelves.
- l. Enrolment of children of school age in a LG Area.
- m. The size of a family.
- n. Height of school children.

3.2 TYPES OF VARIABLES

A variable is any characteristic or property of an object event or phenomenon that can take on different set of values or qualities. We have used this concept so far to describe anything to which numerical figures (such as

scores in a test) or qualitative descriptions can be ascribed (such as Low, middle and high male/female).

Variables may be classified as qualitative or quantitative.

A qualitative variable is one which can be described by assigning names, such as male or female. Class I, II, III, IV, V & VI. Qualitative variables usually take on finite sets of values, that is, you cannot associate fractions with qualitative variables. The measurements are discrete. Thus, discrete data are associated with either nominal or ordinal scales of measurement.

A quantitative variable is one which can be measured, using an acceptable scale and procedure, to assign numerical values. Quantitative variables usually take on continuous sets of value. The volume of water can be measured using a measuring glass. Grains can be measured using "the mudu". Body temperature can be measured using thermometers.

In all these cases, the values can have fractions and the fractions are meaningful. Thus, continuous data are usually associated with interval and ratio scales of measurement.

3.3. **PARAMETRIC VERSUS NON PARAMETRIC STATISTICS.**

You have now learnt that the type of data you have depends on the type of measurement you have made. The type of measurement you make is itself dependent on the characteristic or property of an object or event that is being measured. If your "measurement" is about an object or event that is described by counting, classification, categorization, "order of" degree of "" or "extend to" we say you are doing qualitative analysis and the data are qualitative. A critical analysis of a historical event is a qualitative analysis.

Qualitative data can be verbal, as in gender classification male/female and in the measurement of attitude to specification of something; strongly disagree, disagree, agree, strongly agree. Note that these qualities, male and female, strongly disagree...strongly agree, can be converted to numerical figures for instance, we can say male =1, female = 2.

So that, in an analysis where this conversion has been done for the variable of gender, each time you see gender=1, we know we are talking about males and when gender=2, we know we are talking about female. These conversions do not make them any less qualitative, because in this case, you cannot add 1+2!

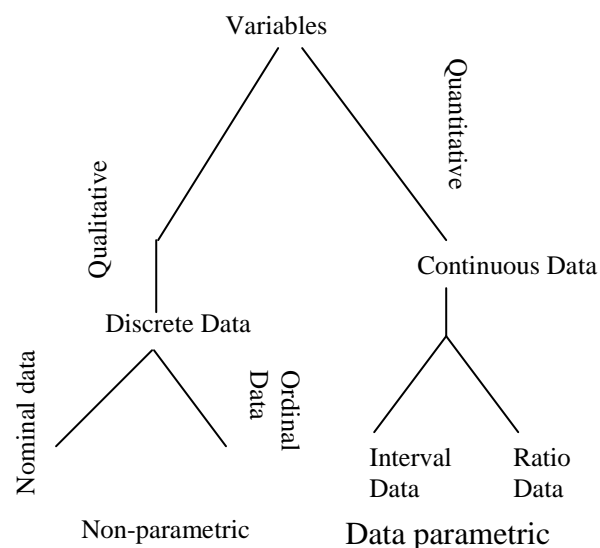
It is simply meaningless in this context. Furthermore, between 1 and 2 is a $1\frac{1}{2}$ gap. There is no $1\frac{1}{2}$ or 1.75. It is either 1 or 2. There's no number before 1 and there are no numbers after 2.

In short, we cannot do much mathematical operations with the figures derived from qualitative data.

The statistical analysis which are associated with qualitative data are described as non-parametric.

If, on the other hand, your measurement is about an object or event that is continuous, such as time, weights, heights, scores in a test, temperature and intelligence, then the data is continuous and the statistical analysis is quantitative. Most scientific experiments involve measurements of attributes which are continuous in nature; scientific data are therefore almost always numerical and can be treated mathematically. They can be added or subtracted, multiplied or divided. The statistical analyses which are associated with quantitative data are described as parametric analyses.

These concepts can be represented diagrammatically as follows.



Conclusion

In this unit, you have learnt four types of measurement scales that give rise to four types of statistical data, namely:

- i. Nominal data recognized by name (from the French which is nom) e.g. names of staff.
- ii. Ordinal data recognized by order/rank but no magnitude such as academic titles and ranks (Professor, Associate Professor, Senior Lecturers, Lecturers etc).
- iii. Interval data recognized by the meaningfulness of part scores between units and equality of units but no absolute zero (e.g.) the meter ruler.
- iv. Ratio data recognized by the characteristics or attitudes associated with interval data in addition to a nominal and ordinal data,

UNIT 5

ORGANIZATION AND PRESENTATION OF DATA

1.0 INTRODUCTION

So far you have been exposed to the various types of data, you have learnt the scales of measurement or the levels of measurements used in generating data, we shall move forward in the handling of data by arranging and presenting the data using different methods. This is a step towards the analysis of data. Although we have always emphasized that statistics is about large amount of data. We shall be using relatively small amount of data in this course, in order to enable you learn the methods very well and to enable you do the calculations conveniently by hand.

2.0 OBJECTIVES

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

- i. Arrange a given set of scores in a given order.
- ii. Prepare a composite frequency distribution table for both ungrouped and grouped data.
- iii. Represent a given data using ideograph/pictograms.
- iv. Deduce some elementary characteristics of a composite table containing grouped data such as class interval, real limits etc.

3.1 ORGANIZATION OF DATA

When a statistician is confronted with data, the first thing he does is some kind of organization. Organization of data involves the arrangement or grouping of data in order of magnitude or according to the species/type/kind/form/ or pattern. Naturally data or scores collected from observations are disorderly and not arranged in any manner. There is therefore a need to organize the scores in an order.

When the number of data collected is small, such as in a class of 5 students. If they are given a test, the teacher will list the names of the students and against each name he will write down their scores. The teacher can use the scores to identify the highest, the lowest, the middle etc.

However, when the number of students involved is large, the simple listing of names and scores will not make it easier for proper analysis. Sequencing can be used.

Example 5.1 in a mathematics test given to 30 students in a class, the scores obtained out of a maximum of 10 were as follows:

7, 5, 3, 9, 5, 8, 4, 7, 4, 2

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

5, 7, 6, 8, 5, 4, 7, 3, 5, 6

These scores are not arranged in any order. They can be arranged in ascending or descending order, as follows.

2, 2, 3, 3, 3, 4, 4, 4, 4, 4

5, 5, 5, 5, 5, 5, 6, 6, 6, 6

6, 7, 7, 7, 7, 7, 8, 8, 8, 9

This is called sequencing; it simply means organizing the data in a logical order, either in ascending order, from the lowest to the highest scores or in descending order, from highest to lowest scores.

Activity 5.1

Given that a class of 40 students was given a test in technical drawing, out of a maximum of 15 marks, the following scores were obtained.

5, 10, 8, 7, 5, 3, 12, 14, 9, 8

7, 2, 0, 5, 3, 10, 11, 12, 13, 7

13, 12, 9, 7, 8, 5, 12, 13, 9, 4

15, 14, 10, 2, 5, 6, 8, 11, 4, 7

- i. Arrange the scores in descending order
- ii. What is the lowest score? How many people earned the highest score?
- iii. What is the most frequent score?
- iv. If the pass mark is fixed at 7. How many people passed? How many people failed?

3.2 FREQUENCY TABLE

In the last activity, i.e. activity 5.1, where you arranged the scores of 40 students in order, you noted that this method of organising data is very clumsy, especially when the number of scores in the distribution is very large.

A better approach therefore is tallying and presentation of the scores in a frequency table. A frequency distribution is a tabular presentation of data showing how often each score or group of scores has occurred. Now let us form a frequency table using the scores given above.

Example 5.2. Given the same set of scores as in example 5.1 above, we are required to form a frequency table with it.

7, 5, 3, 9, 5, 8, 4, 7, 4, 2
 3, 2, 5, 6, 4, 8, 6, 4, 7, 6
 5, 7, 6, 8, 5, 4, 7, 3, 5, 6

To present the scores in a frequency table:

- i. Organise the scores in order and put in a table
- ii. Every score that is lifted, is neatly cancelled and tallied
- iii. Each complete bundle of tallies has 5 tallies in it
- iv. The tallies are counted for every score and written on the frequency column
- v. The table is drawn

Score	Tally	F.
2		2
3		3
4		5
5		6
6		5
7		5
8		3
9		1

Scores	Frequency
2	2
3	3
4	5
5	6
6	5
7	5
8	3
9	1

Activity: 5.2

Present the data below in a frequency table

5, 10, 8, 7, 5, 3, 12, 14, 9, 8,

7, 2, 0, 5, 3, 10, 11, 12, 13, 7

13, 12, 9, 7, 8, 5, 12, 13, 9, 4

15, 14, 10, 2, 5, 6, 8, 11, 14, 7

The type of data presented above is called ungrouped data. You will note that most of the times data collected and recorded especially from the social sciences and education may be so many in numbers. In such a situation we need to group the scores so as to organize and present them. This is because it is cumbersome to study or interpret large data without grouping it, even if it is arranged sequentially. The data are therefore organized into groups called classes and presented in a table which gives the frequency of each group or class. Such a frequency table gives us a better overall view of the distribution of data and enables us to rapidly understand important characteristics of the data.

Example 5.3

The scores of a group of students from an examination were recorded as follows:

82, 56, 68, 74, 86, 80, 83, 91, 70, 67

76, 92, 86, 65, 81, 61, 63, 65, 62, 73

68, 66, 78, 66, 81, 82, 63, 55, 93, 71

62, 84, 78, 72, 71, 70, 76, 80, 61, 59

93, 87, 71, 73, 77, 88, 58, 70, 79, 55

70, 69, 68, 56, 87, 82, 67, 58, 87, 71

78, 68, 72, 72, 77, 86, 77, 80, 90, 69

71, 75, 76, 81, 81, 48, 72, 76, 78, 75

These scores can be grouped and presented in a frequency table using the following tips:

- i. Find the range: the highest score minus the lowest score. In this case, it is $93 - 48 = 45$
- ii. Decide on how many classes or groups you want. This depends on the size of data, but between 5 and 20 groups are recommended. For very large data, 10 to 20 groups are recommended but for relatively small size use between 5 and 10 groups.
- iii. The width of the intervals or class interval (C.I) is got by dividing the range by the number of classes.
 e.g. range = 45
 number of classes = 10
 \therefore Class interval = $\frac{45}{10} = 4.5 = \underline{5}$
 The length of class interval preferred is 2,3,5,10 or 20
- iv. Group the scores and tally as usual.

S/No	Group or C. I.	Tally	F
1	93 – 97		2
2	88 – 92		4
3	83 – 87	 	8
4	78 – 82	 	15
5	73 – 77	 	12
6	68 – 72	 	20
7	63 – 67	 	8
8	58 – 62		6
9	53 – 57		4
10	48 - 52		1
		N =	80

Activity 5.3

Present the data below in a frequency table, using class interval of 5.

15, 8, 12, 18, 44, 30, 15, 18, 23, 6
 23, 16, 20, 17, 21, 12, 12, 23, 25, 13
 19, 17, 17, 28, 13, 17, 17, 28, 18, 16
 20, 7, 14, 8, 15, 27, 10, 19, 13, 15
 18, 10, 8, 11, 16, 40, 18, 21, 14, 27
 15, 32, 28, 22, 10, 9, 18, 12, 25, 25
 18, 20, 21, 18, 18, 16, 9, 8, 21, 17
 29, 23, 14, 14, 25, 15, 12, 10, 20, 16
 24, 19, 15, 11, 21, 12, 15, 8, 17, 19

3.3 COMPOSITE FREQUENCY TABLE

In both example 5.3 and activity 5.3, you have noted the frequency tables have only two columns when you remove the tally column. But from frequency distribution we may have other columns added for relative frequency, cumulative percentage distribution. When these are added the table becomes a composite table.

Before we move to the composite table, let us explain these terms.

- i. Relative frequency: This tells us the proportion of the total interval of each score or group.
- ii. Cumulative frequency: This shows the number of scores which are below the upper limit of each class interval.
- iii. Cumulative percentage distribution: This tells us the percentage of scores which are below the upper limit of each class interval. Now, let us construct a composite frequency distribution table using the data in example 5.3

Example 5.4

S/No	Class Int.	F	R. F.	C. F.	C %
1	93 – 97	2	$\frac{2}{80} = 0.025$	$2+4+8+15+15+12+20+8+6+4+1=80$	$\frac{80}{80} \times \frac{100}{1} = 100\%$
2	88 – 92	4	$\frac{4}{80} = 0.050$	$4+8+15+12+20+8+6+4+1 = 78$	$\frac{78}{80} \times \frac{100}{1} = 97.5\%$
3	83 – 87	8	$\frac{8}{80} = 0.100$	$8+15+12+20+8+6+4+1 = 74$	$\frac{74}{80} \times \frac{100}{1} = 92.5\%$
4	78 – 82	15	$\frac{15}{80} = 0.180$	$15+12+20+8+6+4+1 = 66$	$\frac{66}{80} \times \frac{100}{1} = 82.5\%$
5	73 – 77	12	$\frac{12}{80} = 0.150$	$12+20+8+6+4+1 = 51$	$\frac{51}{80} \times \frac{100}{1} = 63.75\%$
6	68 – 72	20	$\frac{20}{80} = 0.250$	$20+8+6+4+1 = 39$	$\frac{39}{80} \times \frac{100}{1} = 48.75\%$
7	63 – 67	8	$\frac{8}{80} = 0.100$	$8+6+4+1 = 19$	$\frac{19}{80} \times \frac{100}{1} = 23.75\%$

8	58 - 62	6	$\frac{6}{80} = 0.180$	6+4+1	=11	$\frac{11}{80} \times \frac{100}{1} = 13.75\%$
9	53 - 57	4	$\frac{4}{80} = 0.050$	4+1	=5	$\frac{5}{80} \times \frac{100}{1} = 6.25\%$
10	48 - 52	1	$\frac{1}{80} = 0.125$	1+0	=1	$\frac{1}{80} \times \frac{100}{1} = 1.25\%$

N = 80

Activity 5.4

Present the data in activity 5.3 in a composite table having columns:

- i. class interval, ii. Frequency iii. Relative frequency iv. Cumulative frequency and v. cumulative percentage frequency.


3.4 PICTOGRAMS OR IDEOGRAPHS

So far you have learnt how to present data using tables. Now, you are going to see another method of presenting data. This method is called Pictograms or Ideographs. It involves the representation of groups of numerical data by the use of some pictures or diagrams. This is done such that a single picture or diagram represents a specified number of scores, items or objects.

Example 5.5

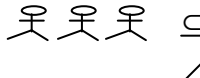
Given that the population of some towns in Abia State are as follows (this is hypothetical):

Town	Aba	Umuahia	Ohafia	Bende	Ukwa	Obingwa
Population	4,500,000	4,000,000	3,500,000	3,000,000	2,500,000	5,000,000

This table can be represented with ideographs this way. First, we choose a convenient and representative picture which will represent a specified number. For instance, you can say let  represent 1,000,000 people.

Therefore, Aba has

Aba =  Umuahia = 

Ohafia = 


Bendel =  Ukwa = 

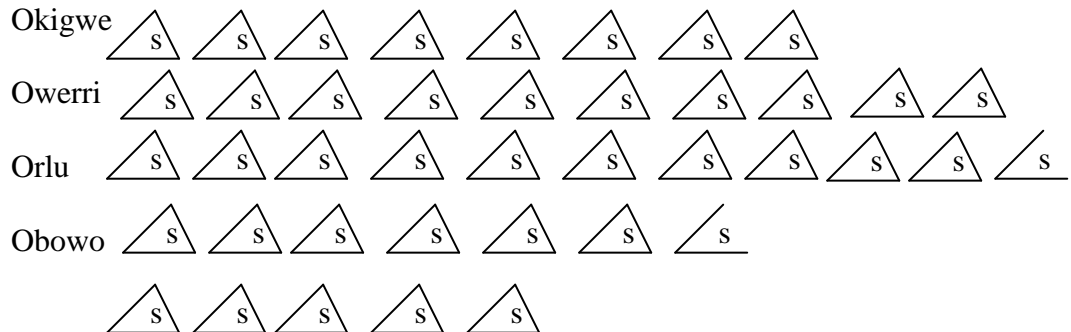
Obingwa = 

Example 5.6


Given the number of schools in some towns in Imo state as follows

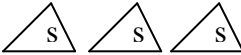
Town	Okigwe	Owerri	Orlu	Obowo	Oru	Onuimo	Orsu
No. of Sch.	80	100	105	65	50	45	30

We can decide to say let  represent 10 schools. Then:



Oru

Onuimo 

Orsu 

Activity 5.5

Present the following data in ideographs

a. No of houses in the following towns

- i. Enugu 850,000.
- ii. Kaduna 700,000
- iii. Kano 650,000
- iv. Onitsha 750,000
- v. Ibadan 920,000
- vi. Benin 680,00
- vii. Port Harcourt 930,000
- viii. Umuahia 800,000
- ix. Minna 540,000

b. No of palm trees in the plantations:

- | | |
|----------------------|------------------------|
| i. Erei = 1,000,000 | v. Ngbei = 450,000 |
| ii. Ulonna = 950,000 | vi. Ihe = 200,000 |
| iii. Ada = 800,000 | vii. Nkwo = 325,000 |
| iv. Zaki = 500,520 | viii. Ihitte = 600,000 |

c. List the demerits of this method

4.0 CONCLUSION

In this Unit you have been exposed to the various ways of organizing and presenting your data. Since in education you will be involved in generating a lot of numerical data from examinations, tests and research results, it is expected that you will not be in a tight corner as far as handling the data is concerned. You can decide to take any method to present your data for analysis.

5.0 SUMMARY

In this Unit you have seen how to arrange data in order either ascending or descending, this is referred to as sequencing and it is called organization of data. You have also learnt how to present data in a frequency table, using tallies in both ungrouped and grouped data. Frequency implies the number of occurrences in a particular group or set. In this unit too, you have learnt how to group data using class interval which is got by dividing the range by the number of groups. Remember that the number of groups is dependent upon the size of the data. For not too large size, 5 to 10 is recommended, but for large number of data, the recommended number of groups is from 10 to 20. You have seen how to construct a composite frequency distribution table to include the class intervals, frequencies, relative frequencies, cumulative frequencies and cumulative percentage distribution.

Apart from sequencing and tabular presentations you have learnt that pictures or diagrams can be used. Other methods of presenting data called graphical methods will be seen in the next unit.

6.0 TUTOR MARKED ASSIGNMENT

The scores of 80 students in a technical drawing examination were given as follows:

68, 84, 75, 82, 68, 90, 62, 88, 76, 93
 73, 99, 88, 73, 60, 93, 71, 59, 85, 75
 61, 65, 75, 87, 74, 62, 95, 78, 63, 72
 66, 78, 82, 75, 94, 77, 69, 74, 68, 60
 96, 78, 89, 61, 75, 95, 60, 79, 83, 71
 79, 62, 67, 97, 78, 85, 78, 65, 71, 75
 63, 80, 73, 57, 88, 78, 62, 76, 53, 74
 86, 67, 73, 81, 72, 62, 76, 75, 85, 77

1. Group the scores with a class interval of 5 and use the groups to form a composite table.
2. Arrange the scores above in ascending order of magnitude.
3. What is an ideograph?

Given that the numbers of books in a local Library are as follows:-

- Section A 5,000
- Section B 7,000
- Section C 15,000
- Section D 3,000
- Section E 1,500
- Section F 6,300
- Section G 2,800
- Section H 3,500

Represent them using ideographs.

7.0 REFERENCE

Indira Gandhi National Open University, School of Education, Statistical Techniques of Analysis (1999) ES-333. Educational Evaluation, New Delhi.

Ogomaka, P. M. C. (1990) Descriptive Educational Statistics:- A guide to Research. Owerri. Top books.

Ughamadu, K. A., Onwuegbu, O. C. and Osunde, A. U., (1990) Measurement and Evaluation in Education, Onitsha, Enba.

7UNIT 6

GRAPHICAL REPRESENTATION OF DATA

1.0 INTRODUCTION

Data which are shown in tabular form as you have seen in the last unit, can also be displayed using graphs. You will note that a well constructed graphical presentation is the easiest way to show a given set of data. In this unit the graphical methods of representing data in statistics, such as bar chart, pie chart, histogram, frequency polygon. Cumulative frequency curve and cumulative percentage curves will be explained and illustrated.

2.0 OBJECTIVES.

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

- i. represent statistical data using the bar graph
- ii. construct a pie chart to represent statistical data
- iii. construct a histogram when given a set of data
- iv. represent a given set of data in a frequency polygon
- v. produce a composite table and construct a cumulative frequency curve
- vi. explain cumulative percentage curve

3.1 BAR GRAPH

You are familiar with graphs and their constructions. In your mathematical lessons or science lessons in your secondary school days, you were used to plot graphs. The bar chart or bar diagram or bar graph involves using orthogonal reflections or shadows of bars or rectangular bars of equal breadth and different heights or lengths to represent the table. Like every other graph which you are used to, the bar chart has two axes, which are the frequency axis or the vertical axis and the item axis or the horizontal axis. The height of each bar is proportional to the frequency.

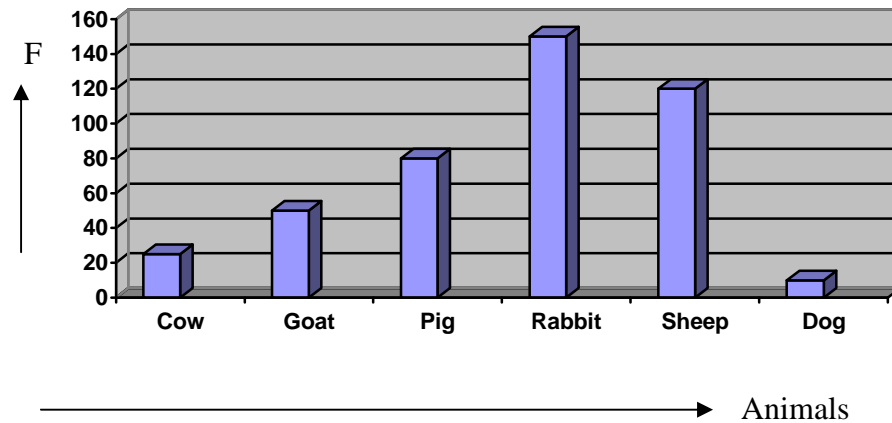
Example 6.1

Given that Mr. Adewale's farm has the following animals:

Animals	Cow	Goat	Pig	Rabbit	Sheep	Dog
Frequency	25	50	80	150	120	10

Represent the data in a bar graph

You remember that the frequencies are plotted on the vertical axis. Therefore note that the highest number on the frequency is 150. Since it is difficult to divide the frequency axis into 150 units, we choose a suitable scale. Let us take 1 cm to represent 10 units along the frequency axis. So we set up the axes as shown.



In constructing a bar chart there are some very important notes to take. These are:-

- i. Bars in a bar chart do not touch each other. This is because items or data represented here are discontinues and have no natural boundaries or common units of measurement between one another. In other word the variables are discrete and the classes are not comparable in terms of magnitude. Bar graphs are used for presentation of discreet data.
- ii. No bar can touch the lines of the frequency axis.
- iii. The unit of measurement on the horizontal axis is not important. So bars are equally spaced and are of equal width on the horizontal axis.

Activity 6.1

Represent the following data in a bar chart.

A

Crops	Yam	Cassava	Maize	Potatoes	Fruits	Vegetables
Frequency	500 kg	2000 kg	900 kg	1500 kg	800 kg	300 kg

B

Colours	Red	White	Yellow	Black	Blue	Brown	Pink
Frequency	360	120	500	80	450	200	300

3.2 THE PIE CHART:

The word pie can be traced to the British circular pie dish or the Mathematical pie. While the pie chart involves using a circular construction to represent the statistical data, such that the data are placed in sectors got using proportions that represent the frequencies.

Example 6.2

Mr Obi's livestock farm has the following animals:

Animals	Cow	Dog	Goat	Pig	Rabbit	Turkey	Sheep	Total
Frequency	15	30	90	45	150	108	102	540

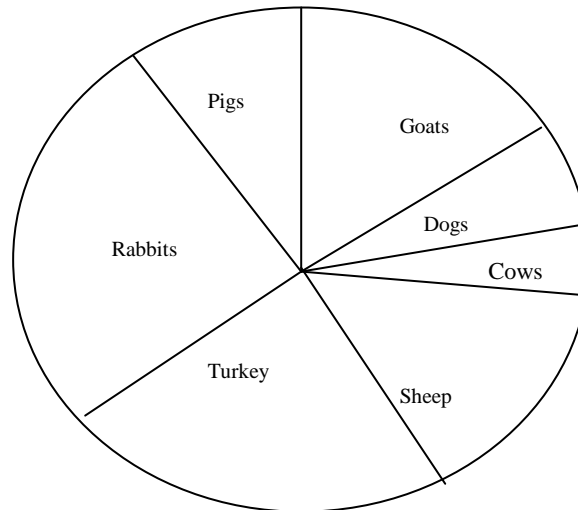
Represent the data in a pie chart.

To do this, you will follow the steps below:

- i. Find the total number of cases or items or frequencies i.e. $15 + 30 + 90 + 45 + 150 + 108 + 102 = 540$
- ii. Divide each frequency by the total frequencies, i.e. $15/540, 30/540, 90/540, 45/540, 150/540, 108/540, 102/540$
- iii. Multiply the results above by 360° . You know that a circle is made up of 360° . i.e. $15/540 \times 360/1, 30/540 \times 360/1, 90/540 \times 360/1, 45/540 \times 360/1, 150/540 \times 360/1$ etc.
- iv. Complete the table as shown below:

Animal	Cows	Dogs	Goats	Pigs	Rabbits	Turkey	Sheep	Total
Frequencies	15	30	90	45	150	108	102	540
Degrees	10°	20°	60°	30°	100°	72°	68°	360°

- v. Use a pair of compasses to draw a circle
- vi. Use a protractor to measure and draw in the angles according to the degrees obtained in (iii) above.
- vii. Write in the groups or items into the sectors that represent them. This is the pie chart.



Activity 6.2

The table below shows the frequency distribution of students offering some subjects in SSCE in 2001/2002 session in a secondary school in a state in Nigeria.

Complete the table and use it to construct a pie chart.

Subject	Maths	English	Physics	Chem.	Geography	T. D.	Agric	Total
Frequency	180	200	80	50	90	45	60	
Degree								

3.3 HISTOGRAM

You will recall from example 6.1 that bar charts are constructed with rectangular bars of equal width and different heights which correspond to the frequencies. The histogram is also represented with rectangular bars. But the bars are not separated as is the case with the bar chart. Can you guess why? This is because the data here are continuous data; therefore, the continuity is shown at the base of the rectangles. The histogram is therefore a bar graph frequency distribution in which the bars are not separated.

The histogram is constructed by plotting the frequencies against the scores or class boundaries of corresponding class intervals.

You have noted the introduction of class boundaries here. Class boundaries mark the limits of a class interval or group. In other words a class interval or group has two class boundaries:- the upper class boundary and the lower class boundary. Let us take the group 62-67 for instance, the upper class boundary is 67 and the lower class boundary is 62. But in the histogram we make use of the Real limits. So we have the real lower limit and the real upper limit. In this case the real lower limit is 61.5 and the real upper limit is 67.5. For you to understand it more, take the number I, for instance. The Real lower limit or boundary is 0.5 and the Real upper boundary or limit is 1.5. So I can be described with its real limits as from 0.5 – 1.5.

To plot the graph;

- i. Draw the vertical and horizontal axes (it is easier to use graph papers where the lines are already drawn).
- ii. Mark the real class boundaries along the horizontal axis or the score axis.
- iii. Mark the frequencies along the vertical axis.
- iv. Construct the bars for each class boundary with a height corresponding to the frequencies.

Example 6.3

Scores of 100 students in an introductory technology test were grouped as shown in the table below. Use the table to construct a histogram.

Steps to follow:

- i. Complete the composite table as shown
- ii. Plot the points marking the bars and using the

frequencies against the real class boundaries using suitable scales, lets use 2 cm to represent 3 units on the frequency axis.

S/No	Class Interval	Class Boundaries		M. Pt or Class M	F
		Lower	Upper		
1	60 – 64	59.5	– 64.5	62	2
2	55 – 59	54.5	– 59.5	57	2
3	50 – 54	49.5	– 54.5	52	6
4	45 – 49	44.5	– 49.5	47	8
5	40 – 44	39.5	– 44.5	42	12
6	35 – 39	34.5	– 39.5	37	14
7	30 – 34	29.5	– 34.5	32	24
8	25 – 29	24.5	– 29.5	27	12
9	20 – 24	19.5	– 24.5	22	16
10	15 – 19	14.5	– 19.5	17	4
N	=				100

—————→ Scores

Points to note:-

- i. The horizontal axis otherwise called abscissa represents the score possibilities, which may be single scores or class intervals. In the case of grouped data, the abscissa is usually marked off by the mid points or the real limits of the intervals.
- ii. Start on the left with the lowest values and then proceed to the right with as many intervals as are necessary to include all the scores. Do not extend this axis to zero, unless scores of zero or near zero have been observed. By convention, leave an empty interval at both ends to show zero frequency in those intervals.
- iii. The vertical axis otherwise called the ordinate represents the frequencies. This axis is marked off with zero at the bottom and moves upwards to the greatest frequency.
- iv. The selection of the distance or scale along either axis is arbitrary, but it is a convention among statisticians to follow the “ $\frac{3}{4}$ high rule”. This states that the vertical axis should be laid out such that the height of the maximum point or highest frequency is approximately $\frac{3}{4}$ of the

- length of the horizontal axis. (Take note of this rule when ever you want to plot a graph in statistics).
- v. A bar or rectangle is drawn above each score interval on the horizontal axis. The width extends from the lower real limit to the upper real limit of the intervals. Bars are adjacent to and touch each other to show the continuity of the scores in a continuous data. Where there is zero frequency, leave an empty space or interval.
 - vi. The vertical axis should be labelled for frequency while the horizontal axis is also labelled to show what is being measured (e.g. scores, height in, weight in gms, time in seconds, temperature in degree c/f etc). Always put the descriptive title indicating what the graph is showing.

Activity 6.3

Given below are the frequency distributions of grouped scores for a set of students in a Geography test. Prepare a composite table and use it to construct a Histogram.

S/No	Class Interval	F
1	42 –	44
2	39 –	41
3	36 –	38
4	33 –	25
5	30 –	32
6	27 –	29
7	24 –	26
8	21 –	23
9	18 –	20
10	15 –	17
11	12 –	14
12	9 –	11

3.4 THE FREQUENCY POLYGON

You have learnt how to construct graphs but the graphs you have so far constructed in this unit were bar graphs. The frequency polygon considers the frequencies against the scores as in the Histogram, but this time the polygon is a line graph which uses the frequencies against the class marks or the mid points of the class intervals. It is a graph joining the points of interception between the two points marked or shown by Xs or dots. These points are joined with straight lines which are made to rest on the horizontal axis. It can also be plotted on a histogram, but this takes time.

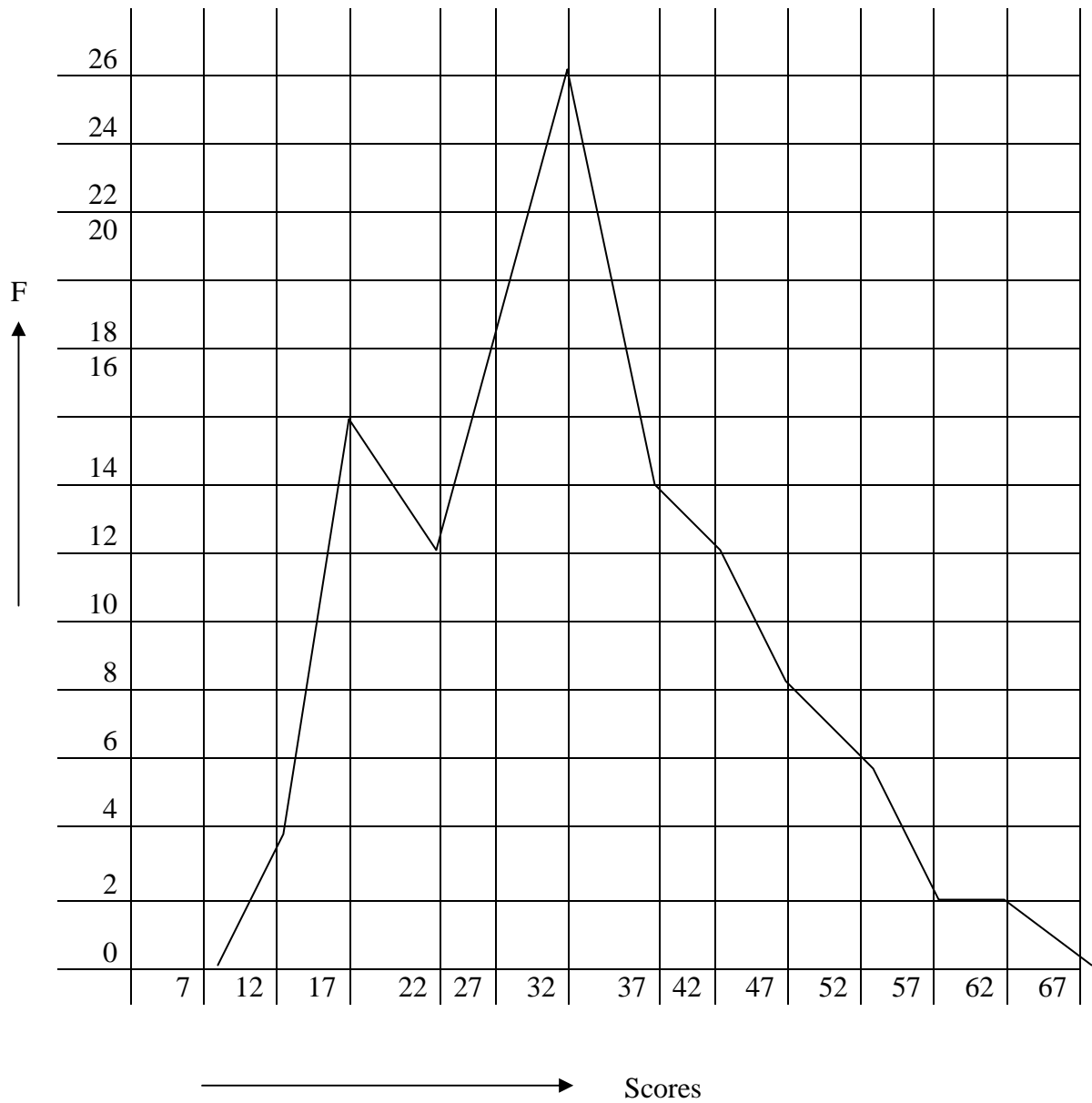
Example 6.4

Consider the group scores obtained from a computer science test by a group of students in a school.

S/No	Class Interval			Mid pt X	F
*	60	–	64	62	0
1	55	–	59	57	2
2	50	–	54	52	2
3	45	–	49	47	6
4	40	–	44	42	8
5	35	–	39	37	12
6	30	–	34	32	14
7	25	–	29	27	24
8	20	–	24	22	12
9	15	–	19	17	16
10	10	–	14	12	4
*	5	–	9	7	0

Steps to follow:

- i. Complete the composite table with the following columns as shown above – class interval, class mark or mid point and frequency.
- ii. Set up the vertical axis for the frequencies and the horizontal axis for the mid points, with suitable scales.
- iii. Add one interval with zero frequency each side (left & right)
- iv. Mark the points of interception and join with straight lines
 On the vertical axis, let 1 cm represent 2 units
 On the horizontal axis, let 1 cm represent 5 units



Activity 6.4

Marks obtained in statistics by a group of students are grouped in a table below. Represent the scores in a frequency polygon.

S/No	Marks			F
1	85	–	89	1
2	80	–	84	2
3	75	–	79	6
4	70	–	74	10
5	65	–	69	15
6	60	–	64	25
7	55	–	59	38
8	50	–	54	20
9	45	–	49	13
10	40	–	44	5

3.5 CUMULATIVE FREQUENCY CURVE

You have learnt that cumulative frequency shows the number of scores which are below the upper limit of each class interval. The cumulative frequency curve which is otherwise called Ogive is a graph of the cumulative frequencies against the scores or the real exact limits of the class interval. The points of contact represent the cumulative frequencies at the exact upper limits of the intervals.

You will have to take note that the general trend of the Ogive is progressively rising, there are no inversions or set backs. The upward rise is not a straight line. It usually takes the shape of a shallow S. While the upper branch approaches its limit N gradually, the lower branch approaches its limit of Zero but not as gradually as the upper branch.

Example 6.5

Represent the grouped data below in a cumulative frequency curve or Ogive.

Frequency curve or Ogive

S/No	Class interval			F
1	55	–	59	1
2	50	–	54	1
3	45	–	49	3
4	40	–	44	4
5	35	–	39	6
6	30	–	34	7
7	25	–	29	12
8	20	–	24	6
9	15	–	19	8
10	10	–	14	2

S/No	Class Interval	F. U. L	F.	C. F
1	55 - 59	59.5	1	50
2	50 - 54	54.5	1	49
3	45 - 49	49.5	3	48
4	40 - 44	44.5	4	45
5	35 - 39	39.5	6	41
6	30 - 34	34.5	7	35
7	25 - 29	29.5	12	28
8	20 - 24	24.5	6	16
9	15 - 19	19.5	8	10
10	10 - 14	14.5	2	2

$$N = 50$$

STEPS TO FOLLOW

- i. Complete the composite table to include, the real upper limit (RUL) and the cumulative frequencies (C. F.)
- ii. With suitable scales draw the vertical and horizontal axes, with the vertical axis having the cumulative frequencies, while the horizontal axis has the real upper limits
- iii. Match the cumulative frequencies against their corresponding real upper limits and join with a smooth curve.

S/No	C. I.	F.
1	95 - 99	2
2	90 - 94	4
3	85 - 89	6
4	80 - 84	10
5	75 - 79	14
6	70 - 74	14
7	65 - 69	50
8	60 - 64	40
9	55 - 59	33
10	50 - 54	8
11	45 - 49	6
12	40 - 44	1

For the scales: let us say 1 cm represent 4 units on the vertical axis and 1.5 cm represent 5 units on the horizontal axis.

Activity 6.5

Use the table given in the activity 6.4 to construct a cumulative frequency curve.

3.6 CUMULATIVE PERCENTAGE CURVE

You have finished activity 6.5 look at the curve you have obtained. Recall that you obtained the curve using cumulative frequencies against the real upper limits. Now, if you convert the cumulative frequencies to cumulative percentages and then use the values against the exact upper limits, you will get the same curve. So in some situations, we may wish to use the cumulative percentages instead of cumulative frequencies in the construction of the curve. In this case the cumulative frequencies are converted to get the shallow S-curve. The advantage here is that one can very quickly approximate the percentage of the total number of cases which fall below certain scores.

Activity 6.6

Construct a cumulative percentage curve using the following data.

S/No	Class Interval	F
1	85 –	89
2	80 –	84
3	75 –	79
4	70 –	74
5	65 –	69
6	60 –	64
7	55 –	59
8	50 –	54
9	45 –	49
10	40 –	44

4.0 CONCLUSION

In this unit, you have been exposed to the various types of graphical representation of statistical data, and their applications in representing the data. As teachers who generate data often, you can choose any of these methods at any time, depending on what you want, to represent your data for easy interpretations.

You can go through these methods again for your practice and familiarise yourself with them. Once again they are Bar chart, pie chart, histogram, frequency polygon, cumulative frequency curve and cumulative percentage curve.

5.0 SUMMARY

In this unit, you have been able to go through with illustrations and constructions, the graphical methods of representing statistical data. These are:

- i. The bar graph, which involves constructing rectangular bars using the frequencies against the items to represent the data.
- ii. The pie chart which involves using proportions that represent the frequencies.
- iii. The histogram which is a bar graph frequency distribution in which the bars are not separated because the data are continuous. Its construction involves plotting the frequencies against the scores or class boundaries of corresponding class intervals.
- iv. Frequency polygon which is a graph that considers the frequencies against the class marks or mid points of the class intervals.
- v. Cumulative frequency curve which is otherwise called Ogive is also a graph, but that of the cumulative frequencies against the scores or the exact limits of the class intervals.
- vi. The cumulative percentage curve, which can be used sometimes in the place of cumulative frequency curves, involves converting the cumulative frequencies into cumulative percentages and using them to plot against the exact upper limits of the class intervals. The points are joined to get a shallow S-curve.

6.0 TUTOR MARKED ASSIGNMENT

The age distribution (in years) of a group of 200 individuals in a community is given in a table below.

S/No	Class Interval	F
1	1 – 10	8
2	11 – 20	10
3	21 – 30	20
4	31 – 40	25
5	41 – 50	38
6	51 – 60	48
7	61 – 70	30
8	81 – 90	16

9	81 – 90	4
10	91 – 100	1
		200

- a. Complete the composite table
- b. Use the table to construct a histogram
- c. Use the table to construct a frequency polygon
- d. Use the table to construct an Ogive

7.0 REFERENCES

Ogomaka, P. M. C. (1990) descriptive educational statistics: A guide to research. Owerri, Top books.

Ughamadu, K. A; Onwuegbu, O. C. and Osunde, A. U. (1990) Measurement and Evaluation in Education; Onitsha Enba.

Unit 7

MEASURES OF CENTRAL TENDENCY

1.0 INTRODUCTION

The measures of central tendency or location are a set of bench marks or typical scores which make precise and brief presentation or description of a set of data. There are basically three measures of central tendency which are of interest to the statisticians. These are the mean, the median and the mode. This unit will therefore expose you to the methods of handling or calculating them.

2.0 OBJECTIVES

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

- i. Define the mean and calculate the mean from a given set of scores.
- ii. Explain the median and find the median from a given set of scores.
- iii. Describe the mode and locate the mode in a given set of scores.

3.1 THE MEAN

You are familiar with the arithmetic average which is used to find the average performance of the students in your class, or the average performance of students in deferent school subjects. This is the same with the mean which is an interval statistics and which is generally most reliable, most stable and most widely used measure of central tendency and which takes into account every score in the distribution. It can be used in computation for more sophisticated statistical analyses. It is equal to the sum of the scores divided by the number of scores. The symbol is \bar{x} and the formula is $\bar{x} = \Sigma X/N$ where \bar{x} = mean

ΣF = Sum of
 X = raw score
 N = number

Example 7.1

Find the mean of the following set of scores:

10, 21, 15, 29, 22, 12, 11, 5, 3, 2

$$\text{Mean} = \bar{X} = \frac{\Sigma x}{N} = \frac{10+21+15+29+22+12+11+5+3+2}{10} = 130$$

$$\therefore \bar{X} = 130/10 = \underline{13}$$

The example above is used when small number of data is given. Most of the times, data can come in a frequency table as in the example below.

Example 7.2

Find the mean of the frequency distribution of scores below.

S/No	X	F
1	10	3
2	9	2
3	8	4
4	7	4
5	6	3
6	5	4
7	4	4
8	3	2
9	2	3

S/No	X	F	FX
1	10	3	30
2	9	2	18
3	8	4	32
4	7	4	28
5	6	3	18
6	5	4	20
7	4	4	16
8	3	2	6
9	2	3	6
Σ		= 29	174

Steps to follow:-

- i. The formula is $\bar{x} = \Sigma fx / \Sigma f$ Therefore add the next column on the table which is fx
- ii. Find $\Sigma f = 29$
- iii. Find $\Sigma fx = 174$
- iv. $\therefore \bar{x} = \Sigma fx / \Sigma f = 174 / 29 = 6$

Example 7.2

Above is an illustration of ungrouped data. When the number of data is large and when the data are grouped the examples above can not be used. So we will now go to the next example using grouped data.

S/No	For Group Data Scores	F
1	55 – 59	1
2	50 – 54	1
3	45 – 49	3
4	40 – 44	4
5	35 – 39	6
6	30 – 34	7
7	25 – 29	12
8	20 – 24	6
9	15 – 19	8
10	10 – 14	2

S/No	Group	Mid Pt X	F	FX
1	55 – 59	57	1	57
2	50 – 54	52	1	52
3	45 – 49	47	3	141
4	40 – 44	42	4	168
5	35 – 39	37	6	222
6	30 – 34	32	7	224
7	25 – 29	27	12	324
8	20 – 24	22	6	132
9	15 – 19	17	8	136
10	10 – 14	12	2	24
	Σ		50	1480

Steps to follow:

- i. Find the mid points of the class intervals
- ii. Find FX by multiplying the mid points with the frequencies
- iii. Find the $\Sigma fx = 50$
- iv. Find the $\Sigma fx = 1480$
- v. Find the $\bar{x} = \Sigma fx / \Sigma f = 1480 / 50 = \underline{29.60}$

The mean can also be calculated using another method called assumed mean or deviation method. Let us take another example to illustrate this.

Example 7.4

Assumed mean method or deviation method

S/No	C. I.	F
1	55 – 59	2
2	50 – 54	2
3	45 – 49	6
4	40 – 44	8
5	35 – 39	12
6	30 – 34	14
7	25 – 29	24
8	20 – 24	12
9	15 – 19	16
10	10 – 14	4

S/No	C. I.	CM	F	X	FX
1	55 – 59	57	2	5	10
2	50 – 54	52	2	4	8
3	45 – 49	47	6	3	18
4	40 – 44	42	8	2	16
5	35 – 39	37	12	1	12
6	30 – 34	32	14	0	0
7	25 – 29	27	24	-1	-24
8	20 – 24	22	12	-2	-24
9	15 – 19	17	16	-3	-48
10	10 – 14	12	4	-4	-16
	Σ	100			-48

Steps to follow:-

- i. Prepare a composite table to include:
 - a. Mid points or class marks (CM)
 - b. Frequencies
 - c. Deviations (X) and F
 - d. X. Note that the deviation is coded positive above the assumed mean and negative below.
- ii. Note the group which is centrally located, or which has about half of the scores. The mid point is taken as the assumed mean. In this case, take 32.
- iii. The deviation is coded O on the assumed mean. Then above it we have 1,2,3...N, and below we have -1,-2,-3...N.

- iv. Multiply the coded deviations with the frequencies.
- v. Find the summation and divide by N
- vi. Use the formula for assumed mean $= \bar{x} = AM + int (\Sigma fx / \Sigma f)$ where \bar{x} = mean, AM = assumed mean, int. or I = class interval size.

$$\begin{aligned}
 X &= AM + int (\Sigma fx / \Sigma f) &&= 32 + 5 (-.48 / 100) \\
 &&&= 32 + (5 \times -.48) &&= 32 + (5 \times -.48) \\
 &&&= 32 + (-2.40) \\
 &&&= 32 + (-2.40) = 32 - 2.40 \\
 &&&= \underline{29.60}
 \end{aligned}$$

Activity 7.1

Using the following grouped data:-

S/No	C. I.	F.
1	95 – 99	2
2	90 – 94	4
3	85 – 89	6
4	80 – 84	10
5	75 – 79	14
6	70 – 74	14
7	65 – 69	50
8	60 – 64	40
9	55 – 59	33
10	50 – 54	8
11	45 – 49	6
12	40 – 44	1

- a. find the mean using grouped data method
- b. find the mean using assumed mean method

3.2 The Median: (\bar{X})

You have learnt that the mean is an arithmetic average. The median is also an average. But it is a positional average. It is the middle point in a scale of distribution of measurement above which half or 50% of the distribution falls and below which half or 50% of the scores lie. The first step in locating the middle point or median is to arrange the scores in order-ascending or descending. You will recall that organisation of data or sequencing involves arranging the scores in order.

Example 7.5

Find the median of the following set of scores: 10, 21, 15, 30, 22, 12, 11, 6, 5, 3, 4.

Steps to follow:

- i. Arrange in order = 3, 4, 5, 6, 10, 11, 12, 15, 21, 22, 30.
- ii. Find the middle score = 11

This is possible when the number N is odd. If the number, N, is even, you will add the two middle numbers and divide by two.

Example 7.6

Find the median of the following set of scores.

13, 15, 12, 17, 16, 16, 19, 14, 20, 11

Steps to follow:

- i. Arrange in order. = 11, 12, 13, 14, 15, 16, 16, 17, 19, 20.
- ii. Find the middle score or scores. i.e. 15 and 16
- iii. The median therefore is $\frac{15+16}{2} = \frac{31}{2} = 15.5$

Activity 7.2

Find the median of the following set of scores.

65, 48, 39, 57, 70, 49, 33, 72, 61, 42, 38, 66, 75, 57, 45, 59, 60, 47, 55, 68

You have seen that the examples above used ungrouped data. Median can also be found in grouped data. Let us take another example with grouped data.

Example 7.7

Find the median of the following groups of scores

S/No	C. I.	F.
1	60 – 64	2
2	55 – 59	2
3	50 – 54	6
4	45 – 49	8
5	40 – 44	12
6	35 – 39	14
7	30 – 34	24
8	25 – 29	12

9	20 – 24	16
10	15 – 19	4

S/No	C. I.	F	C. F
1	60 – 64	2	100
2	55 – 59	2	98
3	50 – 54	6	96
4	45 – 49	8	90
5	40 – 44	12	82
6	35 – 39	14	70
7	30 – 34	24	56
8	25 – 29	12	32
9	20 – 24	16	20
10	15 – 19	4	4

N = 100

Steps to follow:

- i. Prepare the frequency distribution table showing the frequencies and the cumulative frequencies.
- ii. Find the total number of scores $N = 100$.
- iii. Divide N by 2 = $N/2 = 100/2 = 50$
- iv. From the cumulative frequency column, find where 50 lies. You can see that it lies within the group 30-34 where the cumulative frequency is 56. This is the median class
- v. Find the cumulative frequency before the median class = 32
- vi. Subtract it from $N/2 = 50 - 32 = 18$
- vii. Divide the result by the frequency within the median class = $18/24$
- viii. Multiply the result by the class size

These can be summarized using a formula.

$$\text{Median} = \tilde{X} = L + \frac{(N/2 - Cfb)i}{fw}$$

- Where L = the real lower limit of the median class
 Cfb = cumulative frequency below median class
 fw = frequency within the median class
 i = class interval size or simply class size

From the table above: $L = 29.5$, $cfb = 32$, $fw = 24$, $i = 5$

$$\begin{aligned} \therefore \tilde{X} &= L + \frac{(N/2 - Cfb)i}{fw} = 29.5 + \frac{(100/2 - 32)5}{24} \\ &= 29.5 + \frac{(50 - 32)5}{24} = 29.5 + 18/24 \times 5 \\ &= 29.5 + 3.75 = \underline{\underline{33.25}} \end{aligned}$$

Activity 7.3

Using the grouped data below find the median

S/No	Class Interval	F.
1	95 – 99	2
2	90 – 94	4
3	85 – 89	6
4	80 – 84	10
5	75 – 79	14
6	70 – 74	14
7	65 – 69	50
8	60 – 64	40
9	55 – 59	33
10	50 – 54	8
11	45 – 49	6
12	40 – 44	1

3.3 The Mode (\hat{X})

You have learnt that the median is a positional average. The mode is a measure of popularity. It is defined as the point on the scale of measurement with maximum frequency in a distribution. In other words, it is the score value which occurs most frequently in a group of scores.

Example 7.8

Find the mode in the following scores

10, 11, 9, 20, 17, 12, 20, 11, 20, 9

Here the most popular score is 20. It has occurred three times, and more than any other score.

You can see that in the example 7.8 above, we have only one mode. It is therefore called unimodal. But sometimes you may come across a set of scores which has two modes.

Example 7.9

1, 2, 9, 6, 7, 5, 2, 8, 9, 4

In this example, 2 and 9 appeared two times while others appeared once. Therefore, the modes are 2 and 9. This is bimodal; you may also come across, some cases where there are more than two modes. This is called multimodal.

You have noted that the two examples above made use of ungrouped data. When grouped data are given, the mode is the mid point of the class with the highest frequency i.e. the modal class.

Example 7.10

Find the mode of the grouped data below

S/No	Class Int	F
1	60 – 64	2
2	55 – 59	2
3	50 – 54	6
4	45 – 49	8
5	40 – 44	12
6	35 – 39	14
7	30 – 34	24
8	25 – 29	12
9	20 – 24	16
10	15 – 19	4

The mode can be calculated using $L + \left[\frac{d_1}{d_1 + d_2} \right] i$

Where L = lower exact limit of the modal class.

d_1 = difference between the frequency of the modal class and frequency of the class before the modal class.

d_2 = the difference between the frequency of the modal class and frequency of the class above it

i = class size

^

$$\text{The mode } X = L + \left(\frac{d_1}{d_1 + d_2} \right) i$$

From the table $L = 29.5$, $I = 5$

$$\begin{aligned} \therefore \hat{X} &= 29.5 + \left(\frac{24-12}{24-12+24-14} \right) 5 = 29.5 + \left(\frac{12}{12+10} \right) 5 \\ &= 29.5 + \frac{60}{22} = 29.5 + 2.27 = \underline{\underline{32.23}} \end{aligned}$$

Activity: 7.4

Using the following grouped data, find the mode

S/No	C. I.	F
1	95 – 99	2
2	90 – 94	4
3	85 – 89	6
4	80 – 84	10
5	75 – 79	14
6	70 – 74	14
7	65 – 69	50
8	60 – 64	40
9	55 – 59	33
10	50 – 54	8
11	45 – 49	9
12	40 – 44	1

4.0 CONCLUSION

In this unit you have been exposed to the measures of central tendency which are bench marks or typical scores which give precise and brief description of a set of data. These are very important aspects of statistics which you as a teacher can not afford to toy with.

To make your data very precise for interpretation, you will need to learn these measures of location very well.

5.0 SUMMARY

In this unit you have learnt that the measures of central tendency are a set of bench marks which make precise and brief presentation or description of a set of scores. The three basic measures of central tendency are the mean, the median and the mode.

The mean is the most widely used. It is equal to the sum of the scores divided by the number of scores. The symbol is \bar{x} and the formula is $\frac{\Sigma X}{N}$ or $\frac{\Sigma fx}{\Sigma f}$. Or for assumed mean = $AM + \text{int} (\Sigma fx / \Sigma f)$.

The median is the middle point in a distribution. It is a positional average. It can be found in a grouped data using median $\tilde{X} = L + \left[\frac{N/2 - cfb}{fw} \right] i$

The mode is a measure of popularity. It is the point on the scale of measurement with maximum frequency. In a distribution. It shows the highest frequency in a set of scores. There can be unimodal, bimodal or

Multimodal. In a grouped data, it is found using mode $\tilde{X} = L + \left[\frac{d_1}{d_1 + d_2} \right] i$

6.0 TUTOR MARKED ASSIGNMENT

Use the data below to find;

- Mean
- Median and
- Mode

S/No	Class Int	F
1	75 – 79	2
2	70 – 74	4
3	65 – 69	6
4	60 – 64	10
5	55 – 59	25
6	50 – 54	35
7	45 – 49	20
8	40 – 44	15
9	35 – 39	10
10	30 – 34	5

7.0 References:-

Ogomaka, P. M. C. (1990) Descriptive Educational Statistics: A guide to Research. Owerri Topbooks.

Ughamadu, K. A. Onwuegbu, O. C and Osunde, A. U (1990) Measurement and Evaluation in Education. Onitsha, Enba.

UNIT 8**MEASURES OF VARIABILITY/DISPERSION I****1.0 INTRODUCTION**

The measures of dispersion or variability are measures which show the variation or spread or dispersion of values or scores in a given distribution. The homogeneity or heterogeneity of the scores can be established with these measures. In this case, some or all the scores are considered with a view to determining the differences between the scores. The measures which we shall discuss in this unit are the range, the quartiles, the deciles and the percentiles.

2.0 OBJECTIVES

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

- i. Define and calculate the range in a given set of scores.
- ii. Explain and locate the quartiles in a distribution of scores
- iii. Explain and locate the deciles in a set of scores
- iv. Explain and calculate the percentiles of a given set of scores

3.1 THE RANGE

This is the simplest but crude and unreliable method of estimating variability. It is defined as the difference between the highest and the lowest scores in a given distribution. It is usually affected by the presence of two extreme scores. The greater the range, the greater the dispersion or variability. There are two types of range. The common type, most commonly used and simply called the range is technically known as exclusive range. It is the highest score minus the lowest score in a set of scores. It can be found using the formula $X_h - X_L = R$ where X_h represents the highest score and X_L is the lowest Score.

Example 8.1

Find the range in the following set of scores.

66, 59, 72, 62, 57, 54, 66, 79, 14, 65, 64, 95, 59

If you look at the scores very well, you will notice that the lowest score $X_L = 14$ and the highest score $X_h = 95$. Therefore the range R . will be $X_h - X_L = 95 - 14 = \underline{81}$

Activity 8.1

Find the range in the set of scores below
53, 59, 60, 48, 64, 72, 56, 34, 75, 52, 36, 93

Note that you do not need to arrange the scores in any order of magnitude before finding the range.

The other type of range which is not commonly used is called inclusive range. It involves subtracting the real lower limit of the smallest score or observation from the real upper limit of the highest score. It is called inclusive because both the lowest score and the highest score are included in this arrangement. It is mostly used with grouped data.

Find the inclusive range in the grouped data below.

Score	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
Frequency	2	4	8	9	15	14	10	4

Lowest interval = 11 – 15, highest interval = 46-50

The exact limits are 10.5-15.5 and 45.5-50.5

The smallest observation = 10.5, the highest observation is 50.5

Therefore the range is $50.5 - 10.5 = 40$

Note that

- i. The inclusive range is always higher than the exclusive range by 1
- ii. When the range is not qualified in an ungrouped data it is conveniently taken to mean the more commonly known range, the exclusive range.

3.2 THE QUARTILES

In the last unit, you learnt that the Median is a positional score, which occupies the middle point on the score scale. In this same way, the quartiles are positional scores. When we count up from below to include the lowest, or first, quarter of the cases, we find the point called the first quartile. This is given the symbol Q_1 . The first quartile is the score point that sets off the lower quarter or 25% of the group. In the same way, when you count down from above to include the highest, or fourth, quarter of the cases, we locate the third quartile or Q_3 . In other words the 3rd quartile is the score point that sets off the upper 25% or quarter of the scores. The middle quartile Q_2 is the median score point. You will note that the quartiles Q_1 , Q_2 and Q_3 are points on the measuring scale. They are division points between the quarters. We may say therefore of an individual that he is in the highest quarter or 4th

quarter but not in a certain quartile. So the quartiles are points that divide a score scale into four equal parts. These points can be located in a distribution scale.

Example 8.3

Locate the Q_1 and Q_3 in the data given below.

Scores	15	18	21	23	25	27	28	29	32
Frequency	1	1	2	3	6	5	3	4	3

Steps to follow:

i. Complete the table to include the cumulative frequency.

S/No	Score	F	CF
1	32	3	28
2	29	4	25
3	28	3	21
4	27	5	18
5	25	6	13
6	23	3	7
7	21	2	4
8	18	1	2
9	15	1	1

\leftarrow Q_3
 \leftarrow Q_1

ii. Find the 25% or $\frac{1}{4}$ of the number of scores = $\frac{25}{100} \times \frac{28}{1} = 7$

iii. Count from below along the frequency column until you get 25% of the cases. This gives Q_1 . It is between 23 and 25 i.e. $Q_1 = \frac{23+25}{2} = \underline{24}$

iv. Find 75% or $\frac{3}{4}$ of the scores = $\frac{75}{100} \times \frac{28}{1} = \underline{21}$

v. Count from below along frequency column until you get 75% of the cases. This gives Q_3 . It is between 28 and 29 i.e. $Q_3 = \frac{28+29}{2} = 28.5$

3.3 QUARTILE DEVIATIONS

You have learnt that quartiles are points on the score scale that divide the total number of observations or scores in a distribution into four equal groups. You can now locate the points Q_1 , Q_2 and Q_3 . Then, note that the interval from Q_1 to Q_3 contains the middle 50% or half of the scores in a distribution.

It is called the interquartile range. Note also that if the interquartile range is divided by two, we shall have what is called the semi-interquartile range or quartile deviation. This can be found using the formular $\frac{Q_3 - Q_1}{2}$

Example 8.4

Using the results from example 8.3 we can find the quartile deviation or semi-interquartile range.

$$Q_1 = 24, Q_3 = 28.5 \quad \therefore \text{Semi interquartile range} = \frac{Q_3 - Q_1}{2} \\ = \frac{28.5 - 24}{2} = 2.25$$

Example 8.5

Find the semi-interquartile range in the group data below

S/No	C. O.	F
1	55 – 59	1
2	50 – 54	1
3	45 – 49	3
4	40 – 44	4
5	35 – 39	6
6	30 – 34	7
7	25 – 29	12
8	20 – 24	6
9	15 – 19	8
10	10 – 14	2
	N =	50

Steps to follow:

- Find $N/4 = 50/4 = 12.5$
- Counting up the frequency column, locate the 12.5 cases. You see that if we count 2+8, we have 10 cases. It means that we need 2.5 out of the next frequency, which is 6, to complete. We say therefore $2.5/6 \times 5$ (5 is the interval size) = 2.08
- Add the result above to the real lower limit of the class interval i.e. $19.5 + 2.08 = \underline{21.58}$. This is Q_1 .
- For Q_3 , count from the top-down 12.5 cases. Again we have 1+1+3+4 will give us 9. It means that we will need 3.5 to make it up to 12.5 cases. It will be 3.5 out of the next frequency which is 6. We have therefore $3.5/6 \times 5 = 2.92$
- Since we are going down, we deduct 2.92 from the real upper limit of the class i. e. $39.5 - 2.92 = 36.58$. This is Q_3 .
- Find the semi-interquartile Q, using $\frac{Q_3 - Q_1}{2}$

$$= \frac{36.58 - 21.58}{2} = \frac{15.00}{2} = \underline{7.5}$$

Note that the formula used in locating the median can be applied here

$$\text{i.e. } L + \frac{(N/2 - cfb)}{f_w} i$$

$$\text{In the case of quartiles, instead of } N/2 \text{ you will use } N/4 \text{ so we have } Q = L + \frac{(N/4 - cfb)}{f_w} i$$

Activity 8.2

Find the quartile deviation (QD) or semi-interquartile range of the scores from 65 students on the use of English examination in a school.

Scores	52-54	49-51	46-48	43-45	40-42	37-39	34-36	31-33	28-30
Frequencies	6	11	16	8	9	8	2	3	2

3.4 THE DECILES

So far you have learnt how to divide a given set of scores into two equal parts to locate the mid point or the median; you have also learnt how to divide a set of scores into four equal parts to locate the quartiles. This time we shall move to another step. This is to divide into ten equal parts to locate the deciles. Decile points are used to mark off a distribution, thus indicating points of dividing a distribution of scores into tenths. Thus there are 9 deciles i.e. from 1 to 9 which divide a distribution into ten equal parts. D_1 is the first decile and below D_1 lies the bottom 10% of the group. In the same way D_2 is the point in the distribution below which 20% of the cases fall. Like quartiles, deciles are points in a distribution not segments.

3.5 PERCENTILES

Percentiles are ordinal measures. They are score points which divide the distribution into 100 equal parts called percentages. In other words, they are points on the raw score scale below which given percentages of the cases in the distribution fall. For instance, the 80th percentile is the point on the score scale that has exactly 80% of the cases below it. Percentiles are symbolised by the letter P_x , with X denoting the particular percentile. Thus, the 90th percentile is written P_{90} , they are used for decision making when part of a population is to be selected because of its position within the total.

Note that the median corresponds to the 50th percentile, P_{50} and 2nd Quartile Q_2

The 1st quartile corresponds to the 25th percentile P_{25}

The 3rd quartile corresponds to the 75th percentile P_{75} .

You will recall that the formula for calculating median is $X=L + \frac{(N/2-cfb)}{fw} i$.

You also recall the formula for the quartiles = $\frac{(N/4 - cfb)i}{fw}$

you can see that they are almost the same. The formular for calculating the percentiles is the same and it is the adaptation of this same formula. It is used for specific percentile points. The general formula which can be used for any value of P_x is $P_x = L + \frac{(P_n - Cfb)i}{fw}$ where

P_x = score at any given percentile

L = the lower real limit of the interval containing the given percentile (P_x)

P_n = the desired percentage of N ($N/100$)

N = the total frequency

cfb = cumulative frequency below the interval containing the percentile

fw = frequency within the interval containing the desired score

I = interval size

Now, let us illustrate this with an example

Example 8.6

Use the data below to calculate:

- a. P_{95}
- b. P_{50}
- c. P_{40} and
- d. P_{20}

Note: $N/100 = 200/100 = 2$
 $i = 5$

S/No	Class Interval	F
1	95 – 99	1
2	90 – 94	6
3	85 – 89	8
4	80 – 84	33
5	75 – 79	40
6	70 – 74	50
7	65 – 69	24
8	60 – 64	14
9	55 – 59	10
10	50 – 54	8
11	45 – 49	4
12	40 – 44	2
	N =	200

Steps to follow

i. The steps are the same with those of median and quartiles

$$\begin{aligned} \text{a. For } P_{95} &= L + \left[\frac{\left(\frac{200}{100} \times \frac{95}{100} \right) - cfb}{fw} \right] i = L + \frac{(190 - cfb)i}{fw} \\ &= 84.5 + \left[\frac{198 - 185}{8} \right] 5 = 84.5 + \frac{5}{8} \times 5 = 84.5 + 3.125 = 87.625 \end{aligned}$$

ii. For $P_{50} = L +$

$$\begin{aligned} L + \left[\left(\frac{\frac{200}{100} \times \frac{50}{100}}{fw} \right) - cfb \right] i &= L + \left(\frac{100 - cfb}{fw} \right) i = 69.5 + \left(\frac{100 - 62}{50} \right) 5 \\ &= 69.5 + \left(\frac{38}{50} \times 5 \right) = 69.5 + 3.8 = 73.3 \end{aligned}$$

iii. For $P_{40} = L +$

$$\begin{aligned} \left[\left(\frac{\frac{200}{100} \times \frac{40}{100} - cfb}{fw} \right) \right] i &= L + \left(\frac{80 - cfb}{fw} \right) i = 69.5 + \left(\frac{80 - 62}{50} \right) 5 \\ &= 69.5 + \left(\frac{18}{50} \times 5 \right) = 69.5 + 1.8 = 71.3 \end{aligned}$$

$$\begin{aligned} \text{iv. For } P_{20} &= \left[\left(\frac{200}{100} \times \frac{20}{1} \right) - cfb \right] i = L + \left(\frac{40 - cfb}{fw} \right) i = 64.5 + \left(\frac{40 - 38}{24} \right) 5 \\ &= 64.5 + \left(\frac{2}{24} \times 5 \right) = 64.5 + 0.417 = 64.917 \\ &= \underline{64.917} \end{aligned}$$

Activity 8.8

Using the data below, find the following percentiles

- P_{75}
- P_{25}
- P_{40}
- P_{90}
- P_{10}

C. I.	91-99	82-90	73-81	64-72	55-63	46-54	37-45	28-36	19-27	10-18	1-9
Frequency	2	3	5	9	15	18	15	9	5	3	2

4.0 CONCLUSION

In this unit have gone through source of the measures of variability or dispersion. These are the measures used to established the homogeneity or heterogeneity of a set of scores in a distribution scale. It is very important that you study them very well as you will displays use them in the analysis and interpretation of your data.

5.0 SUMMARY

In this unit you have been exposed to some of the measures of variability which are measures that show the spread of the scores in a given distribution. The measures you have seen so far are:

- The range:- which simply shows the difference between the highest and the lowest observations or numbers.
- The quartiles are points which divide the distributions or scores into
- four equal parts called quarters. The formula is $L + \left(\frac{N/4 - cfb}{fw} i \right)$
- The deciles are also points on the distribution that divide the distribution into ten equal parts or tenths. The formula is

$$L + \frac{(N/10 - cfb)i}{fw}$$

- v. The distance from Q_1 to Q_3 is called interquartile range while half of this is called semi-interquartile range or quartile deviation.
- v. Percentiles: which are points on the score scale that divide the distribution into 100 equal parts called centiles or percentages. The formula is

$$L + \frac{(N/100 - cfb)i}{fw} \text{ or } \frac{(P/N - cfb)i}{fw} + L$$

The 1st quartile Q_1 corresponds to the 25th percentile, the 3rd quartile Q_3 corresponds to the 75th percentile while the 2nd quartile Q_2 which is the median corresponds to the 50th percentile and the 5th decile.

6.0 TUTOR MARKED ASSIGNMENT

Using the data below find:

- Q_1, Q_3
- P_{10}, P_{50}, P_{90}
- Quartile deviation

S/No	Class Interval	F
1	80 – 84	2
2	75 – 79	3
3	70 – 74	5
4	65 – 69	10
5	60 – 64	16
6	55 – 59	20
7	50 – 54	30
8	45 – 49	25
9	40 – 44	22
10	35 – 39	18
11	30 – 34	15
12	25 – 29	10
13	20 – 24	8
14	15 – 19	5
15	10 – 14	2

7.0 REFERENCES

Ary, Donald and Jacobs, L. C. (1976) introduction to statistics: Purposes and Procedures. New York, Chicago, San Fransisco, Atlanta, Toronto, London, Sydney, Montreal; Holt Rinehart and Winston.

Guilford, J. P and Fruchter, B (1978) Fundamental statistics in Psychology and Education. International Student Edition, Auckland, Boyota ... Sydney, Tokyo. McGraw-Hill International Book company.

Unit 9**MEASURES OF VARIABILITY/DISPERSION II****1.0 INTRODUCTION**

In unit 7, you learnt the various measures of central tendency and you were told that these measures are a set of bench marks which make precise and brief presentation or description of a set of data. You also learnt that the measures of location are very useful in providing a concise index of the average of a set of scores. But there is more to know about sets of scores. Variability is a universal characteristic of any set of scores with which the teacher, the psychologist or researcher might have to deal. Some distributions may have the same mean yet differ in the extent of variation of the scores around the measure of central tendency.

Therefore, in order to describe a distribution of scores adequately, we shall need both the measures of central tendency and variability. This is because information concerning variability may be as important or more important than information concerning central tendency. The measures of variability provide a needed index of the extent of variation among the scores in a distribution. You have seen the range, the quartile deviation, the deciles and percentiles. We shall now look at the mean deviation, variance and standard deviation in this unit.

2.0 OBJECTIVE

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

- i. Calculate the mean deviation in a given distribution
- ii. Calculate the variance in a set of scores
- iii. Calculate the standard deviation in a given distribution

3.1 MEAN DEVIATION

Deviation of a score involves trying to find out how far that score is away from the mean. Look at the following set of scores for instance; 10, 15, 13, 12, 8. The mean is $\frac{58}{5} = 11.6$. The deviation of 10 from the mean $10 - 11.6 = -1.6$. In the same way, the deviation of 15 from the mean $= 15 - 11.6 = 3.4$ the deviation of 13 from the mean $= 13 - 11.6 = 1.4$ etc.

The mean deviation therefore is the average of the numerical deviations of the observation. In other words the mean deviation considers the average of the various deviations or distances of the individual scores from the mean of

the set of scores. It can also be said to be the average of the modulus or positive values of the differences between the individual scores and the mean of the set of scores. In mathematical notation, mean deviation

$$= \frac{\sum(|X-\bar{X}|)}{N} \text{ or } \frac{\sum f(|X-\bar{X}|)}{N} \text{ where } |X-\bar{X}| \text{ is the positive values or modulus.}$$

Example 9.1

Find the mean deviation of the listed scores below 41, 27, 19, 9, 23, 31, 25, 28, 15, 22, 35.

Steps to follow:

- i. Find the mean of the set of scores. It is $\bar{X} = \frac{\sum X}{n}$
- ii. Using a table find the deviation of the scores from the mean
- iii. Add the positive values of the deviations i.e. $16+2+6+16+2+6+0+3+10+13+3+10 = 74$
- iv. Find the mean deviation $M.D = \frac{74}{11} = 6.727$

S/No	X	$X - \bar{X}$	D
1	41	41 - 25	16
2	27	27 - 25	2
3	19	19 - 25	-6
4	9	9 - 25	-16
5	23	23 - 25	-2
6	31	31 - 25	6
7	25	25 - 25	0
8	28	28 - 25	3
9	15	15 - 25	-10
10	22	22 - 25	-3
11	35	35 - 25	10

Example 9.2

Find the mean deviation of the set of scores below

S/No	X	F
1	20	2
2	19	3
3	18	5
4	17	10
5	16	15
6	15	9
7	14	5
8	13	2

S/No	X	F	FX	$ X - \bar{X} $	F $ X - \bar{X} $
1	20	2	40	3.76	7.52
2	19	3	57	2.27	8.28
3	18	5	90	1.76	8.80
4	17	10	170	0.76	7.60
5	16	15	240	0.24	3.60
6	15	9	135	1.24	11.16
7	14	5	70	2.24	11.20
8	13	2	26	3.24	6.48
Σ		51	828		64.64

Steps to follow:

- i. Find the mean $\bar{X} = \frac{828}{51} = \underline{16.24}$
- ii. Complete the composite table to include the scores, X, frequencies F, FX, $|X - \bar{X}|$ and F $|X - \bar{X}|$
- iii. Find the total sum of F $|X - \bar{X}| = 64.64$
- iv. Find the mean deviation M.D = $\frac{\sum f|X - \bar{X}|}{\sum f}$
 $= \frac{64.64}{51} = 1.27$

Note that when you have grouped scores, you will use the mid points of the class interval as X.

Activity 9.1

Find the mean deviation of the set of scores below.

X	41	27	19	9	23	31	25	28	15	22	35
F	2	8	10	3	12	15	14	12	11	10	4

3.2 VARIANCE

You have seen that the deviation scores, which you have studied in 3.1 above, provide a good basis for measuring the spread of scores in a distribution. But, we can not use the sum of these deviations in order to get an index of spread because this sum in any distribution will be equal to zero. This becomes a problem which we must overcome. To do this, square all the deviation scores. This is to remove all negative scores and make all the scores positive. This is because all squared scores will be positive. Then these squared deviation scores are added to give a measure called the sum of the squared deviation scores which is simply called sum of squares,

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

The variance therefore is a measure of variability which is derived from the deviation of scores from the mean. It is defined as the mean of the squared deviation scores. It is widely used for inferential statistics than for descriptive statistics. The population variance is symbolized by the lower case Greek letter sigma (σ) raise to the second power i.e. σ^2 , while the sample variance is represented by S^2 . For the purpose of this course we shall be using S^2 since the difference is not noticeably high.

1.2 Example 9.3

For listed scores; using 15, 14, 11, 10, 9, 7, 4. Find the variance.

S/No	X	$(X - \bar{X})$	$(X - \bar{X})^2$
1	15	5	25
2	14	4	16
3	11	1	1
4	10	0	0
5	9	-1	1
6	7	-3	9
7	4	-6	36
	70	0	88

Steps to follow-:

- i. Find the mean of the set of scores = $\frac{70}{7} = \underline{10.00}$
- ii. Find the deviations from the mean. $(X-\bar{X})$
- iii. Find the squares of the deviation $(X-\bar{X})^2$
- iv. find the variance using $S^2 = \frac{\Sigma(x-\bar{x})^2}{N} = \frac{88}{7} = \underline{12.57}$

Activity 9.2

Find the variance of the data below which are score of 15 students in a 10-itemed multiple choice test on maths 7, 2, 6, 8, 4, 3, 5, 9, 6, 1, 6, 8, 0, 7, 3,

Example 9.4 for data with frequencies
Find the variance for the data below.

X	19	18	17	16	15	14	13	12	10	9
F	2	3	5	10	15	9	5	3	2	1

Steps to follow:

1. Complete the composite table to include the scores X, F, FX, $X-\bar{X}$, $(X-\bar{X})^2$, $F(X-\bar{X})^2$

X	F	FX	X-X	$(X-\bar{X})^2$	$F(X-\bar{X})^2$
19	2	38	4.13	17.057	34.114
18	3	54	3.13	9.797	29.391
17	5	85	2.13	4.537	22.685
16	10	160	1.13	1.277	12.769
15	15	225	0.13	0.017	0.254
14	9	126	-0.87	0.757	6.812
13	5	65	-1.87	3.497	17.485
12	3	36	-2.87	8.237	24.711
10	2	20	-4.87	23.717	47.434
9	1	9	-5.87	34.457	34.457
	55	818			230.112

- i. Find the mean of the set of scores. It is $\frac{818}{55} = 14.87$
- ii. Find the deviations from the mean of the set of scores.
- ii. Find the squares of the deviations above
- iv. Multiply the squares of the deviation by frequencies.
- v. Find the sum of the squared deviation x frequencies above

vi. Find the variance S^2 using $\frac{\sum f(X - \bar{X})^2}{\sum f} = \frac{230.112}{55} = \underline{4.104}$

Activity 9.3

Find the variance of the data below.

X	25	30	28	22	15	34	14	26	16	20	32	12	8	10
F	7	5	5	9	10	2	8	15	9	12	3	4	1	2

Note that:

- You shall use N when dealing with population and N-1 when dealing with the samples.
- You shall use the mid points or class mark when you are given grouped data.

Example 9.5:

The raw score method.

Find the variance of the grouped data below.

S/N	Class int	F
1	36-40	1
2	31-35	4
3	26-30	7
4	21-25	10
5	16-20	8
6	11-15	5
7	6-10	3
8	1-5	2

Steps to follow:

- Complete the compost table as shown below.

S/N	Class int	F	X	FX	X ²	FX ²
1	36-40	1	38	38	1444	1444
2	31-35	4	33	132	1089	4356
3	26-30	7	28	196	784	5488
4	21-25	10	23	230	529	5290
5	16-20	8	18	144	234	2592
6	11-15	5	13	65	169	854
7	6-10	3	8	24	64	192
8	1-5	2	3	6	9	18
		40		835		20225

ii. Using the formular
$$\frac{n \sum fx^2 - (\sum fx)^2}{n^2}$$

$$\begin{aligned} \text{We have } S^2 &= \frac{40 (20225) - (835)^2}{40^2} \\ &= \frac{809000 - 697225}{1600} = \frac{111775}{1600} \\ &= \underline{69.859} \end{aligned}$$

Activity 9.4

Find the variance of the grouped data given below.

Class int	56-60	51-55	46-50	41-45	36-40	31-35	26-30	21-25	16-20
Frequency	3	4	5	6	10	18	15	11	8

3.3 Standard Deviation.

You have gone through the variance and the methods and processes involved in calculating the variance. If you have learnt it very well, then you will not have any difficulty in mastering the methods and processes involved in calculating the standard deviation. This is because the standard deviation is simply the square root of the variance. It is by far the most commonly used indicator of degree of dispersion and is the most dependable estimate of the variability in the population from which the sample is drawn. It also enters into numerous other statistical formulas which we shall see latter in this course. The standard is a kind of averages of all the deviation form the mean, but it is not a simple arithmetic mean. The symbol is S for sample and σ for population.

To compute the standard deviation, you will have to find the variance, then find the square root of the variance.

You can use any method to get this.

Example 9.6

Find the standard deviation of the scores below.

S/N	X	Dev	
1	15	5	25
2	14	4	16
3	11	1	1
4	10	0	0
5	9	-1	1
6	7	-3	9
7	4	-6	36
	70	0	88

Steps to follow.

- i. Go through the steps in calculating the variance. After getting the variance then find the square root.

$$S^2 = \frac{\sum (x - \bar{x})^2}{N} = \frac{88}{7} = 12.58 \text{ (variance)}$$

$$\therefore S = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{88}{7}} = \sqrt{12.57} = 3.55$$

Example 9.7 using the raw score methods.

10, 8, 7, 6, 3, 2, 1.

Steps to follow:

Find the squares of the raw scores X^2

X	10	8	7	6	3	2	$\sum = 37$
X^2	100	64	49	36	9	4	$\sum = 263$

- ii. Find the sum of the scores and squared scores. i.e. $\sum x = 37$, $\sum x^2 = 263$
 iii. Find the $(\sum x)^2 = 27^2 = 1369$

- iv. Find the standard deviation using $S = \sqrt{\frac{\sum X^2 - (\sum X)^2}{N}}$

$$S = \sqrt{\frac{263 - 372}{6}} = \sqrt{\frac{263 - 1369}{6}}$$

$$S = \sqrt{\frac{263 - 228.167}{6}} = \sqrt{\frac{34.833}{6}} = \sqrt{5.8055}$$

$$= 2.4094605 = \underline{\underline{2.409}}$$

Activity 9.5

Find the standard deviation of the set of data below.

X	15	11	9	7	5	3	1
F	1	1	2	2	1	2	1

4.0 Conclusion

In this unit you have learnt that apart from the usefulness of the measures of central tendency for providing a concise index of the average value of a set of scores, there is more to be studied about a set of scores, variability which is a universal characteristic of any set of scores is a very important attribute with which the teacher, the Psychologist, the social scientist, the research etc. might have to deal. For instance, the measures of achievement, intelligence, personality and or other characteristics may be expected to show variability in any sample of individuals.

Therefore, to describe a distribution of scores very well and adequately too, we need both the measures of central tendency and the measures of variability. This is because the two measures make up two types of descriptive statistics which are indispensable in describing distributions of scores.

5.0 Summary

In this unit you have been told that the measures of variability or dispersion are very necessary for adequately describing quantitative distributions. The three measures which you have studied in this unit are the mean deviation the variance and the standard deviations. The mean deviation is the average of the deviations of the scores or observations from the mean, given that all the deviations are positive. It uses the modulus in computation. Thus mean deviation.

$$MD = \frac{\sum(X - \bar{X})}{N}$$

The variance is an interval scale measure and is the mean square of deviations from the mean. It is in squared units of the original measure and is given by $\frac{\sum(X - \bar{X})^2}{N}$ or $\frac{\sum X^2 - (\sum X)^2}{N}$ or $\frac{\sum fx^2 - (\sum fx)^2}{N}$

- v. The standard deviation which is the most common of all the measures of variability is the square root of the variance. It belongs to the interval scale of measurement and is given by $\sqrt{\frac{\sum(X - \bar{X})^2}{N}}$ or $\sqrt{\frac{\sum X^2 - (\sum X)^2}{N}}$

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

6.0 Tutor Marked Assignment:

Using the data below, find the variance and standard deviation.

X	25	24	23	22	21	20	19	18	17	16
F	1	2	3	6	11	16	7	9	8	2

7.0 References:

- Ary, Donald and Jacobs, L.C (1976). Introduction to statistic: purposes and procedures. New York Chicago..... Sydney, Montreal . Holt Rinehart and Winston.
- Guilford, J.P and Fruchter, B. (1978) Fundamental Statistics in Psychology and Education. International students Edition. Auckland, Bogotá..... Sydney, Tokyo. McGraw- Hill international Book company
- Ogomaka, P.M.C (1990). Descriptive Educational Statistics: A guide to Research. Owerri Top book.

UNIT 10

1.0 Introduction:

Earlier, you learnt how to plot graphs like the frequency curve or cumulative frequency curve. You have noticed that different curves have different shapes. One of man's most interesting discoveries was the determination of a relationship between measurements of many types of natural phenomena and the mathematical laws of chance if most distributions of many natural events are plotted on a frequency curve, the shape will be like a bell. This is called normal curve. In other words, measurement observed in physical and psychological phenomena will produce a normal curve. For instance, if the heights of randomly selected people in a community are taken and plotted on a graph paper, it will give a normal curve. But apart from the normal curve, some other shapes which are not normal may be observed in some cases. In this unit we shall look at the different shapes of curves that can be observed.

2.0 Objectives:

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

- i. Describe the general nature of normal curves.
- ii. Identify the different types of curves.
- iii. Explain Skewness and Kurtosis
- iv. Compute the co-efficient of skewness and kurtosis

3.1. The Normal Curve

Any symmetrical bell- shaped type of curve is known as a normal curve. The concept of normal curve is very basic in statistics. This is because; the frequency distributions of many natural events have shapes similar to that of a normal curve. In other words many physical and psychological phenomena when shown in a frequency distribution curve will resemble the normal curve. Take for instance, the weights of the girls in a school, the heights of men in a church, the achievement scores of students in a class etc. if you get these measurements and plot them on graphs. They will be similar to the normal curve.

Activity 10.1

Get the scores of the students in a class in one subject. Use the scores to plot a graph of frequency against the scores.

You have seen that the shape of the curve is bell shaped? Similarly, the distributions of scores on many psychological tests, such as IQ tests and tests of school achievement, are approximated by the normal curve

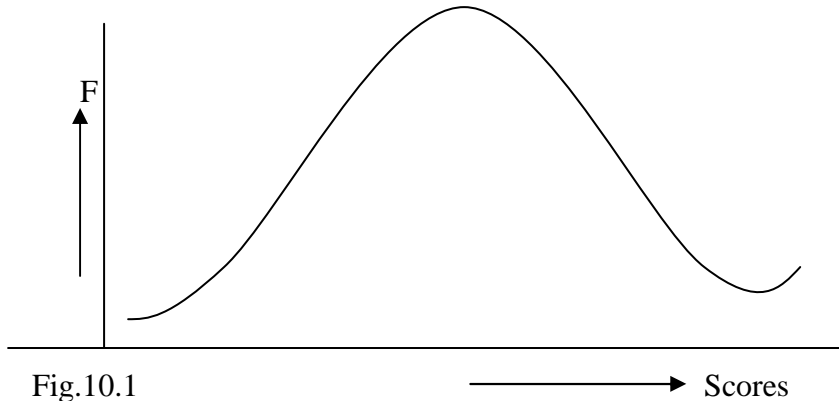


Fig.10.1
The Normal curve.

Although the fit may not be perfect, but you will see that the distribution of scores closely approximates the normal curve.

The normal curve is actually a graph of a rather complex mathematical relationship. This relationship results in a graph which is unimodal and symmetrical, when plotted. There is an equation which is used to define the normal curve. It relates the height of the curve to the score values. It is as follows:

$$y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

Where y = Height of the curve corresponding to the particular value of X .

X = a score value corresponding to a particular height.

π = a constant which is equal to 3.1416 or $\frac{22}{7}$

e = a constant which is equal to 2.7183

μ = the mean of the X variables

σ = the standard deviation of the X variables.

This formula is used to construct a normal curve, and it imparts the characteristics of a bell-shape to the curve. But the form of the curve will also depend on the mean and the standard deviation. If these change, the shape of the curve will also change, but it is still very similar to a normal curve. However, you are not asked to master this formula, since you will not have

to use it in any computations,. It is included only to inform you that there is a mathematical formular for constructing a normal curve but you have to note that this one equation can result in many normal curves. Each time the mean and standard deviations changes a different normal curve is produced.

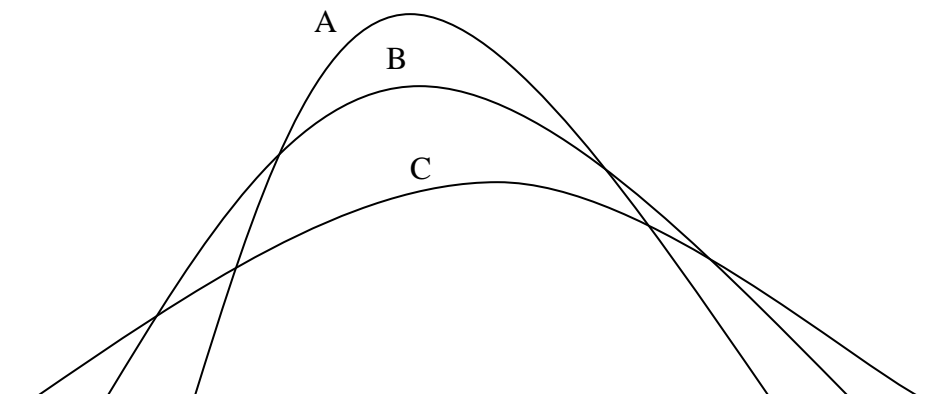


Fig 10.2 Three normal curves.

From the figure above, you will see that the three curves are normal curves, yet they are different in shape and appearance. Curve A is narrow and the ordinate is relatively long. Curve C is wide and the ordinate is relatively short. B is not too narrow nor too wide.

Activity 10.2.

Collect the fooling data:-

- i. The result of a class of students' examination in any one subject of your choice.
 - ii. Collect the weights of the same students in the class.
 - iii. Collect the heights of the same students in the class.
- Plot three graphs with the data on the same graph. Take note of the shapes for comparison.

3.2 The Properties of a Normal Curve.

- i. A normal curve is symmetrical with its maximum height at the mean. It is often described as a bell-shaped curve while some describe it as a well-weathered manure pile.
- ii. The mean, median and mode fall at the same point.
- iii. The height of the curve decreases as one moves to the left and right of the Point of maximum height.

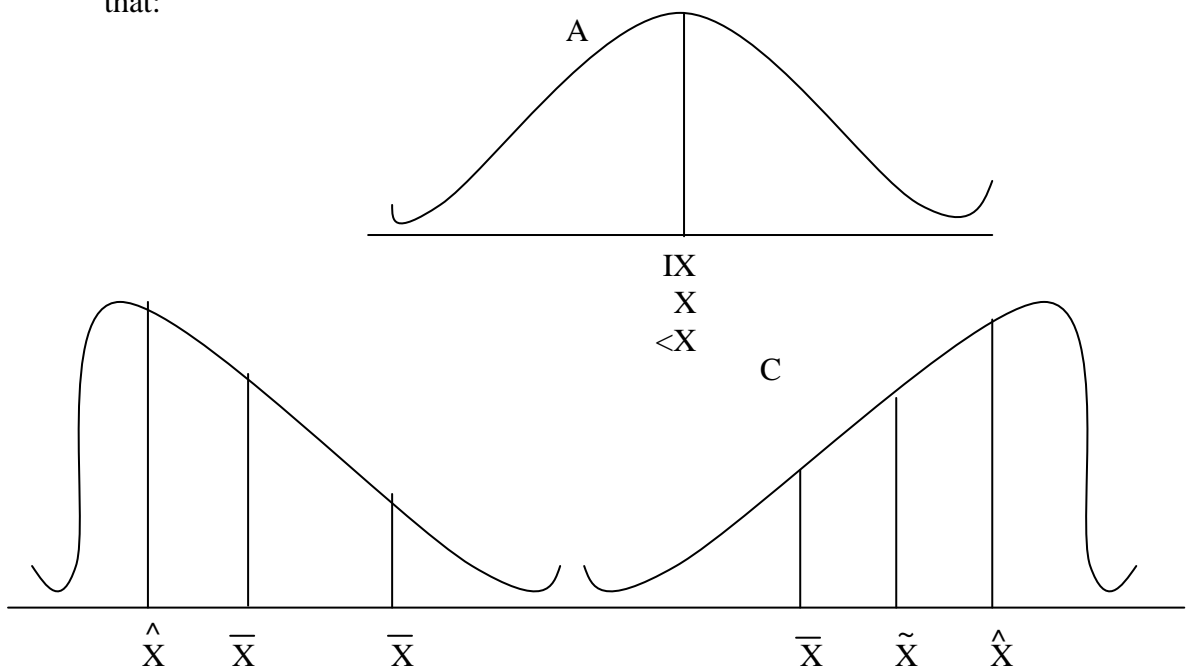
- iv. Although the height of the curve continues to decrease as one moves farther and farther from the mean, it never actually reaches zero. Therefore the theoretical range of the normal curve is from plus infinity ($+\infty$) to minus infinity ($-\infty$).

3.3 Skewness

You remember that the normal curve is symmetrical but many distributions produce curves that are not symmetrical but asymmetrical. These curves lean or bend either to the left or to the right. Such curves are said to be skewed. Skewness, therefore is the degree or extent to which a frequency curve is asymmetric. There are two main types of skewness.

When a curve leans to the left from the observers view point and the tail extends out towards to the right, it is said to be positively skewed. On the other hand, if the curve leans to the right and the tail extends outwards to the left, it is called negatively skewed.

Now, take a look at the diagrams below. Diagram A is a normal curve. B is positively skewed and C is negatively skewed. From these figures we can say that:



- The mode, the mean and the median have the same value in a normal curve.
- The median and the mean lie to the right of the mode in the same direction of the skewness in the positively skewed curve.
- In the negatively skewed curve the median and the mean lie to the left of the mode. In other words the mean and the median are less than the mode.

You will note again that in a normal curve the mean, the mode and the median do not differ. But they differ in a skewed distribution. This difference or a function of it may be taken as a measure of skewness. There are some measures of skewness. They include: the Pearson's first and second coefficients of skewness which are:-

$$\text{i. Skewness} = \frac{\text{mean-mode}}{\text{Standard deviation}} = \frac{\bar{X} - \hat{X}}{S}$$

$$\text{ii. Skewness} = \frac{3(\text{mean} - \text{median})}{\text{Standard deviation}} = \frac{3(\bar{X} - \tilde{X})}{S} \quad \text{others are}$$

$$\text{iii. Quartile coefficient of skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

$$\text{iv. 10-90 percentile of skewness} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{10})}$$

Example 10.1

Find the skewness in the data given below.

X	2-4	5-7	8-10	11-13	14-16	17-19	20-22	23-25	25-28
F	2	5	7	10	15	30	15	10	6

Steps to follow.

i. Complete the composite table.

S/No	Class limit	X	F	FX	$X-\bar{X}$	$(X-\bar{X})^2$	$F(X-\bar{X})^2$
1	2-4	3	2	6	-14.01	196.280	392.5602
2	5-7	6	5	30	-11.01	121.2201	606.1005
3	8-10	9	7	63	-8.01	64.1601	449.1207
4	11-13	12	10	120	-5.01	25.1001	251.0010
5	14-16	15	15	225	-2.01	4.040	60.6015
6	17-19	18	30	540	0.99	0.9801	29.4030
7	20-22	21	15	315	3.99	15.9201	238.8015
8	23-25	24	10	240	6.99	48.8601	488.6010
9	26-28	27	6	162	9.99	99.8001	598.8006
				100	1701		3114.99

ii. Find the mean $\bar{X} = \frac{\sum fx}{\sum f} = \frac{1701}{100} = \underline{17.01}$

iii. Find the standard deviation $S = \sqrt{\frac{\sum F(X - \bar{X})^2}{\sum F}}$

$$= \sqrt{\frac{3114.99}{100}} = \sqrt{31.1499} = 5.5812095 = \underline{5.58}$$

iv. Find the mode. From the table it is 18 (i.e. $16.5 + (\frac{15}{15+15})3$)

v. Find the median = $L + \frac{(N/2 - cfb)}{fw}i = 17.5 + \frac{(50-39)3}{30}$

$$= 17.5 + 1.10 = \underline{18.60}$$

vi. Coefficient of skewness = $\frac{\bar{X} - \hat{X}}{S} = \frac{17.01 - 18}{5.58}$

$$= \frac{-0.99}{5.58} = \underline{-0.177}$$

OR $\frac{3(\bar{X} - \hat{X})}{S} = \frac{3(17.01 - 18.60)}{5.53}$

$$\frac{-477}{5.53} = \underline{-0.863}$$

From the result using the Pearson's coefficient of skewness, you can see that the curve is negatively skewed.

Activity. 10.3

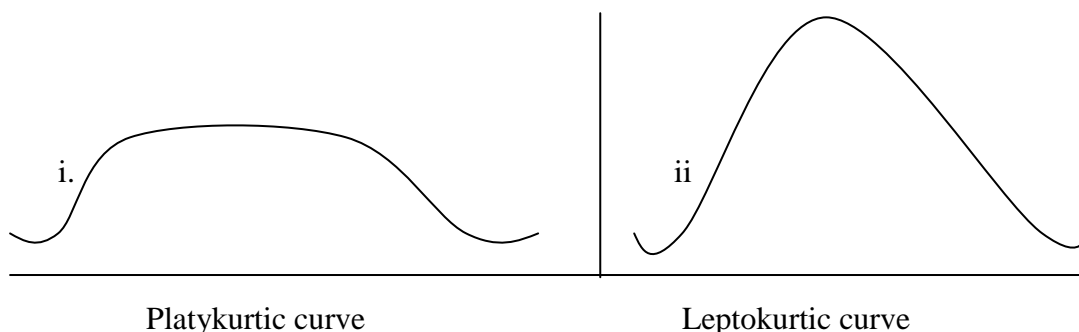
Find the coefficient of skewness in the distribution below.

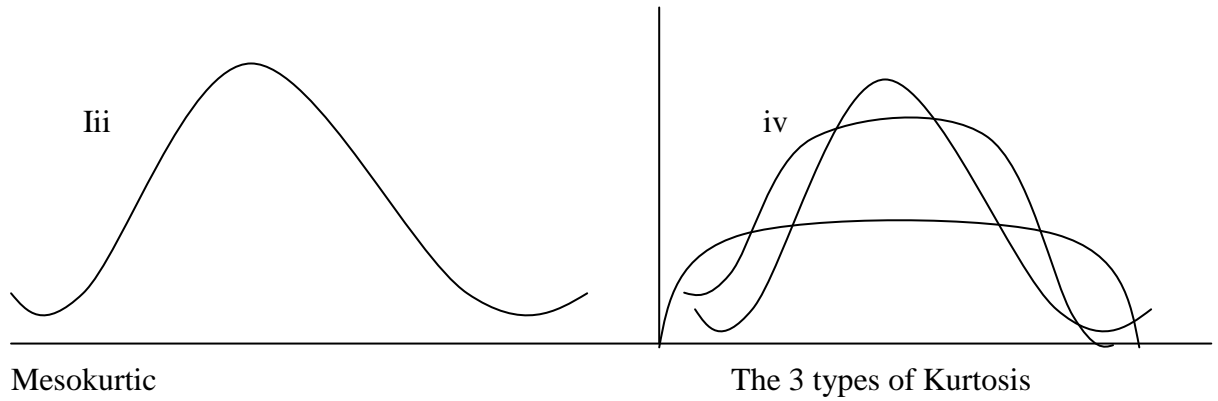
Class	5-7	8-10	11-13	14-16	17-19	20-22
Freq	8	48	26	16	12	8

3.4 Kurtosis

So far you have seen that normal curves are symmetrical and can be used as the basis for certain comparisons in handling curves. You have also seen that curves which are not normal may be skewed either to the left or to the right. There is yet another characteristic of the form of curves. This is called kurtosis. The word kurtosis is derived from a Greek word *kyrtos*, which means curved. Kurtosis therefore describes the peakness or flatness of a curve around the mode in a distribution of scores. The three types of kurtosis are:

- i. Platykurtic (platy means flat in Greek.) This has a broad, relatively flat appearance. It is relatively flat topped.





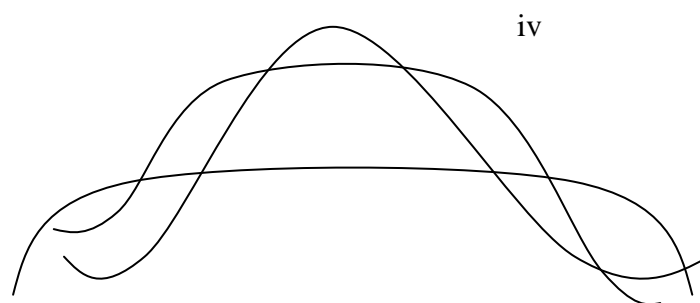
- ii. Leptokurtic: Lepto means thin in Greek. It is relatively highly peaked in the middle. It has thinner tails.
- iii. Mesokurtic: Here the kurtosis is the same as in the normal curve.

The quantitative indices of kurtosis of a distribution can be calculated using the semi-inter quartile range and the nintieth and tenth percentiles. This index is symbolized by the Greek letter K(Kappa) and is given by

$$K = \frac{Q}{(P_{90}-P_{10})} = \frac{1/2 (Q_3-Q_1)}{(P_{90}-P_{10})} \text{ or } \frac{P_{75} - P_{25}}{(P_{90} - P_{10})}$$

Example 10.2

Class	26-28	23-25	20-22	17-19	14-16	11-13	8-10	5-7	2-4
Freq	6	10	15	30	15	10	7	5	2



Steps to follow:

- i. Complete the composite table.

S/No	Class int	F	CF
1	26-28	6	100
2	23-25	10	94
3	20-22	15	84
4	17-19	30	69
5	14-16	15	39
6	11-13	10	24
7	8-10	7	14
8	5-7	5	7
9	2-4	2	2

$$\text{ii. Find } Q = L + \left(\frac{\frac{N/4 - cfb}{fw}}{4} \right) I = \frac{N}{4} = \frac{100}{4} = \underline{25}$$

$$\text{iii. Find } Q_1 = L + \left(\frac{25 - 24}{15} \right)^3 = 13.5 + \left(\frac{1}{15} \times \frac{3}{1} \right) = 13.5 + 0.2 = \underline{13.7}$$

$$\text{iv. Find } Q_3 = L + \left(\frac{75 - 69}{15} \right)^3 = 19.5 + \left(\frac{6}{15} \times 3 \right) = 19.5 + 1.2 = \underline{20.7}$$

$$\text{v. Find } P = L + \left(\frac{N/100 - Cfb}{fw} \right) I = \frac{N}{100} = \frac{100}{100} = \underline{1}$$

$$\text{vi. Find } P_{90} = L + \left(\frac{90 - 84}{10} \right)^3 = 22.5 + \left(\frac{6}{10} \times \frac{3}{1} \right) = 22.5 + 1.8 = \underline{24.3}$$

$$\text{vii. Find } P = L + \left(\frac{10 - 7}{7} \right)^3 = 7.5 + \left(\frac{3}{7} \times \frac{3}{1} \right) = 7.5 + 1.29 = \underline{8.79}$$

$$\text{viii. Find } K = \frac{1}{2} \left(\frac{Q_3 - Q_1}{P_{90} - P_{10}} \right) = \frac{1}{2} \frac{20.7 - 13.7}{24.3 - 8.79} = \frac{7}{15.51} = \frac{3.5}{15.51} = 0.2256608 = \underline{0.23}$$

Activity 10.4

Find the index of kurtosis in the distribution below

Class	7-5	8-10	11-13	14-16	17-19	20-22
Freq	8	48	26	16	12	8

4.0 Conclusion

You have gone through the measures of central tendency. You have seen that when a set of data is appreciably or greatly skewed the median is better than the mean. In this unit you have learnt how to find out the degree to which a set of data is skewed, and to categorize measures of skewness. You have also seen how to find out spread or bunched up of a set of scores which is technically referred to as kurtosis. As a teacher or researcher or even social scientist, you are often confronted with large masses of data, usually scores of some type which require interpretation, if they are to be useful you have to do summarizing of the data by using the graphical presentation. This will show you at a glance the degree of skewness and kurtosis, or when the curve produced is a normal curve.

5.0 Summary

In this unit, you have learnt that the normal curve is a frequency curve of a theoretical distribution which is unimodal and symmetrical with the mean, median and the mode at the same point. The weight is greater at this point and decreases on both sides of this point to form a bell-shaped curve. The theoretical ranges of the normal curve are from $-\infty$ to $+\infty$ but most of the area lies between $+3$ and -3 . you have now known that one of the most immediately obvious characteristics of the form of a graphed frequency distribution is its symmetry or lack of symmetry or balance. A curve is symmetrical in shape if one side is a mirror image of the other. But when one side is not a mirror image of the other, it is asymmetrical and this is characterized by a high point or lump that is off-centre and by tails of distinctly unequal length. This is called a skewed curve. The lump indicates the scores with the highest frequencies. Skewness can be positively or negatively.

Apart from skewness, another characteristic of a normal curve is the measure of kurtosis. Kurtosis is the peakness or flatness around the mode of a distribution. There are three types of kurtosis. These include Leptokurtic, Mesokurtic and Platykurtic curves. The normal curve is the mesokurtic, the highest peaked curve is referred to as platykurtic, while the flat topped is the leptokurtic curve .

6.0 Tutor Marked Assignment:

Compute the coefficient of skewness and index of kurtosis of the distribution below.

Class	7-5	8-10	11-13	14-16	17-19	20-22
Freq	8	48	26	16	12	8

8.0 References

Ary, Donald and Jacobs, L.C (1976). Introduction to statistics: purposes and procedures. New York Chicago..... Sydney, Montreal . Holt Rinehart and Winston.

Ogomaka, P.M.C (1990). Descriptive Educational Statistics: A guide to Research. Owerria Top book.

Ughamad, K.A, Onwuegbu, O. Ci Osund, A.U (1990) Measurement and Evaluation in Education. Onitsha . Emba.

UNIT 11

SOME MEASURES OF ASSOCIATION AND AGREEMENT

1.0 Introduction

So, far we have focused on these statistical procedures used for describing single variables or for analyzing what we may call univariate distributions. You have learnt how to compute the measures of central tendency and variability, but these always come from one variable, such as test scores, etc. but we need statistical methods that can be used to investigate relationship that may exist between two variables in a population or samples.

This is because of the understanding that scientific progress depends upon finding out what things are co-related and what things are not. However, no single statistical procedure or method has opened up so many new avenues of discovery in psychology, Education and possibly the behavioral sciences in general, as that of correlation. In this unit therefore, we shall look at the concept of correlation, bivariate frequency distributions, the Pearson's product moment correlation and its computations.

2.0 Objectives

After completing this unit, you should be able to:

- i. Explain the concept correlation.
- ii. Construct and interpret a scatter gram.
- iii. Describe the differences between positive and negative correlation.
- iv. Use a given paired scores to compute a Pearson product moment correlation coefficient.
- v. List the assumptions of the Pearson product moment correlation.

3.1 The Concept of Correlation

Some of the times, we are faced with such questions as: Is there a relationship between students' achievement in mathematics and their achievement in the sciences? Does socioeconomic status affect school achievement? Is there any relationship between aptitude in Engineering and performance in engineering courses? What is the relationship between time used in Mathematics drills and students' achievement in mathematics? Is there any relationship between the scores of candidates in Common Entrance Examination and their scores in JSCE etc. These questions can be answered using a statistical procedure or technique called correlation? In other words correlational methods are used

for determining relationships between pairs of variables. Correlational methods are used with bivariate distributions. i.e. Distributions where two variables are involved. Thus, bivariate data, unlike univariate data, consist of observations that are paired on some logical basis. Have you noticed that some variables tend to be related? Some statistical indices have been developed to enable us concisely describe such relationship. These are called correlation indices. By convention, the two variables involved are labeled X and Y. A coefficient of correlation is single number that tells us to what extent two things or two variables are related, to what extent variations in the one thing go with variations in the other. You will have to note that without the knowledge of how one variable varies with another, it would be difficult to make predictions. For instance, if we are able to establish that the higher the interest in a particular subject, the higher the level of performance in that subject. We can therefore take the scores to predict the level of performance in a particular subject by particular students based on their interest level.

You will have to note again that if two sets of scores do not have a common source we cannot employ correlation. In other words there must be logical bases for pairing the variables before correlation can be employed.

Once again, remember that correlation simply involves the statistical procedure or technique or method used for describing the extent of linear association or relationship or 'going –together in some linear trend or pattern of distributions of measures of two attributes or variables or constructs possessed by a population or samples of individuals

Activity 11.1

List 10 pairs of variables whose relationships can be investigated using correlational methods.

3.2 Scatter Grams

Scatter diagrams or simply scatter grams were used before calculators and even computers, were as available as they are now. They were also used when samples to be correlated were large, or even moderate in size. The common procedure was to group data in both X and Y and to prepare a scatter gram or correlation diagram to provide some shortcuts in calculation. It is also a way to show correlation visually. A scatter gram therefore is a graph in which a single dot is used to locate each individual on two dimensions. The pattern formed by the dots show the correlation.

To construct a scatter gram we first lay out the scale for one variable on the abscissa and the scale for the other on the ordinate. By now you are very familiar with the construction of graphs having constructed many of them earlier in this course. But note that by convention the variable on the abscissa is labeled X and the variable on the ordinate is labeled Y.

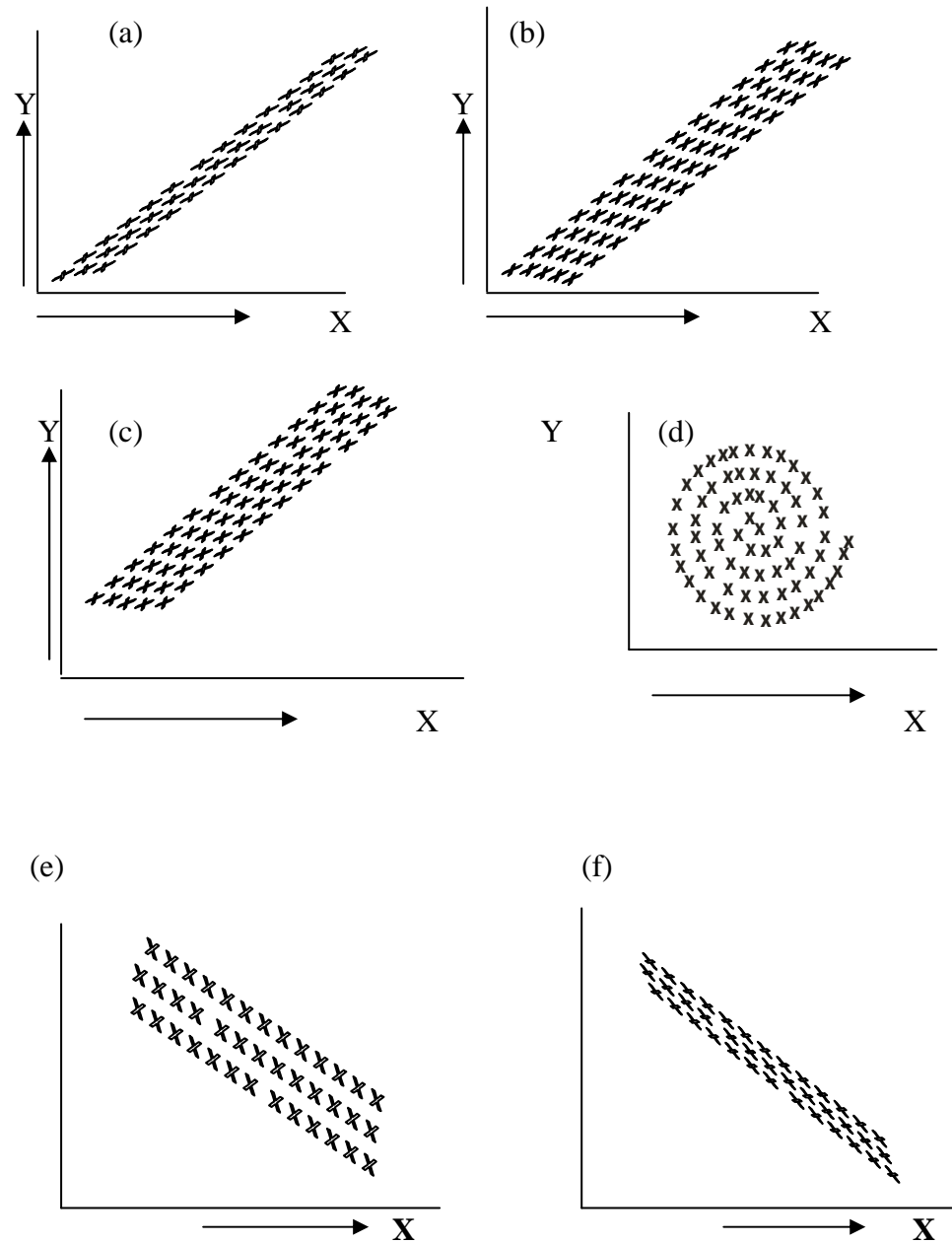


Figure 11.1 Types of Scatter grams

You will have to note that in reading or interpreting the scatter gram, you must bear in mind that when high scores on one variable are associated with high scores on a second variable and low scores on the one associated with low scores on the other, the variables are said to be positively correlated, or to show positive correlation. Fig 11.1 above gives you the ideas of the shapes of some scatter diagrams. From the figures: (a) shows high positive relationship. (b) Moderate positive relationship and (c) shows no or zero relationship. (e) Shows moderately negative relationship while (d) shows high negative relationship

You will also take note that in a positive correlation the dots or x marks in the scatter gram spread from lower left to upper right, while in a negative correlation the marks spread from upper left to lower right. Note also that in some cases, the variables show no tendency to vary or change. In other words, some individuals scoring high on one variable and scoring neither systematically high or low on the other variable. The result is that there is no or zero correlation between the two variables. The marks are spread at random on the scatter gram.

In Education or behavioural sciences we rarely have perfect association while in the physical sciences the association is perfect and gives straight lines. Example the graph of $d = \frac{m}{v}$ or $m = dv$ will produce a straight line but

if you plot the graph of peoples heights against their weights, it will not give you a straight line.

Activity 11.2

List 5 pairs of variables which can produce perfect relationship or association and 5 pairs of variables which can not produce perfect association.

3.4 Bivariate Frequency Distribution

You have learnt how to plot graphs, so talking about X and Y axes may not be new to you. We are going to use the same method to set up a two-way grouping of data having a table prepared in columns and rows. A bivariate frequency distribution is another way to show a correlation visually. Values of the X variable are shown on the abscissa.

While values of the Y variable are shown on the ordinate. In other words, there are columns for the dispersions of Y scores within each score or class interval for the X scale, and rows for the dispersions of x scores within each of the intervals for the Y scale. Along the top of the table are listed the score limits for the class intervals for X scores. Along the left hand margin are listed the score limits for the intervals of Y scores. A tally mark shows each

score combination. Bivariate frequency distributions can be constructed with either ungrouped data or with grouped data.

Example 11.1

Construct a bivariate frequency distributions of scores in two tests of students given below.

A

Class Int	60-64	65-69	70-74	75-79	80-84	85-89	90-94
F	5	8	16	12	17	12	8

B

Class Int	60-69	70-79	80-89	90-99	100-109	110-119	120-129	130-139	140-149
F	5	10	10	14	10	8	10	6	5

What is the interpretation?

Steps to follow:

- i. set up the two way table as shown below to show the number of rows and columns required.

		Test B									
		60-69	70-79	80-89	90-99	100-109	110-119	120-129	130-139	140-149	
Test A	90-94										8
	85-89										12
	80-84							++++			17
	75-79										12
	70-74				++++						16
	65-69										8
	60-64										5
		5	10	10	14	10	8	10	6	5	78

- ii. Fix the classes as follows test A on the right and test B on top of the table.
- iii. Fix the frequencies on the opposite sides
- iv. Match the scores and tally as shown

The interpretation is that the scores are positively related.

Activity 11.3

Construct a bivariate frequency distributions of the test below and give your interpretation.

Test A

Class Int	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
F	2	3	5	8	10	7	4	1

Test B

Class Int	2-4	5-7	8-10	11-13	14-16	17-19	20-22	23-25	26-28	29-31	32-34
F	1	3	3	3	4	5	8	4	5	2	1

3.4 The Pearson Product Moment correlation coefficient.

You have seen that mere inspection of a scatter gram furnishes you with some general information on the relationship between two sets of measures for a given group. It can give you idea on the type and direction of relationship, but it does not give you the degree or extent of relationship. Therefore a numerical index indicating precisely the degree of relationship is much more helpful and highly required.

Different correlation indices have been developed by different people. The most widely used, is the index which is used when both variables are expressed as interval data. This was developed by an English statistician called Karl Pearson and is called the Pearson Product moment coefficient of correlation. It is symbolized by the Greek letter rho ρ , while the statistic is represented by r. there are two methods for computing the Pearson r. These methods are

- i. the deviation methods and
- ii. the raw score methods

3.4.1 The Deviation method.

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

Where x = deviation from the mean of X scores

y = deviation from the mean of Y scores

Example 11.2

Compute the Pearson r for the two sets of data

X	9, 13, 6, 18, 14, 12, 11, 7, 2, 6, 14, 15, 5, 8,
Y	23, 40, 10, 48, 25, 30, 15, 10, 5, 45, 40, 35, 12, 27.

Steps to follow:

- i. Set up a composite table as shown below.
- ii. Find the mean of X scores = $\frac{\sum X}{N} = \frac{140}{14} = \underline{10}$

iii. Find the mean of Y scores = $\frac{\sum y}{N} = \frac{365}{14} = 26.07 = \underline{26.1}$

iv.

S/No	X	Y	X- \bar{X} χ	Y- \bar{Y} Y	χy	χ^2	y^2
1	9	23	-1	-3.1	3.1	1	9.61
2	13	40	3	13.9	41.7	9	193.21
3	6	10	-4	-16.1	64.4	16	259.21
4	18	48	8	21.9	175.2	64	479.61
5	14	25	4	-1.1	-4.4	16	1.21
6	12	30	2	3.9	7.8	4	15.21
7	11	15	1	-11.1	-11.1	1	123.21
8	7	10	-3	-16.1	48.3	9	259.21
9	2	5	-8	-21.1	168.8	64	445.21
10	6	45	-4	18.9	-75.6	16	357.21
11	14	40	4	13.9	55.6	16	193.21
12	15	35	5	8.9	44.5	25	79.21
13	5	12	-5	-14.1	70.5	25	198.81
14	8	27	-2	0.9	-1.8	4	0.81
Σ	140	365			587.0	270	2614.94

v. Find the deviations of X and Y scores and complete the composite table.

vi. Applying the formular $\frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$ we have $\frac{587}{\sqrt{270 \times 2614.94}} =$

$$\frac{587}{\sqrt{706033.8}}$$

$$= \frac{587}{840.25} = 0.698 = \underline{0.70}$$

3.4.2 **The Raw Score Method.** $r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$

Example 11.3

Compute the Pearson r for the same sets of data using the raw score method

$$\begin{array}{l} X | 9, 13, 6, 18, 14, 12, 11, 7, 2, 6, 14, 15, 5, 8, \\ Y | 23, 40, 10, 48, 25, 30, 15, 10, 5, 45, 40, 35, 12, 27. \end{array}$$

Steps to follow:

- Set up a composite table as shown below.
- Find the summation of X scores

- iii. Find the summation of Y scores
 iv. Complete the composite table as shown.

S/No	X	Y	XY	X ²	Y ²
1	9	23	207	81	529
2	13	40	520	169	1600
3	6	10	60	36	100
4	18	48	864	324	2304
5	14	25	350	196	625
6	12	30	360	144	900
7	11	15	165	121	225
8	7	10	70	49	100
9	2	5	10	4	25
10	6	45	270	36	2025
11	14	40	560	196	1600
12	15	35	525	225	1225
13	5	12	60	25	144
14	8	27	216	64	729
Σ	140	365	4237	1670	12131

- v. Apply the formula $r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$

$$r = \frac{14 \times 4237 - (140 \times 365)}{\sqrt{(14 \times 1670 - 140^2)(14 \times 12131 - 365^2)}}$$

$$= \frac{59318 - 51100}{\sqrt{(14 \times 1670 - 140^2)(14 \times 12131 - 365^2)}} = \frac{8218}{\sqrt{13642020}} = \frac{8218}{11763.589}$$

$$= 0.6985963 = \underline{0.70}$$

Activity 11.4

Using both the deviation and raw score methods calculate the correlation coefficient of the scores below.

Maths	28	46	11	34	9	43	21	30	17	25	40	5	48	32	22	16	44	14	37
Phy.	30	33	22	38	7	40	18	26	15	24	36	9	44	30	20	12	46	27	21

3.5 Assumptions of the Pearson r.

Before you finish this unit, you should be informed about some restrictions that should be observed in the use of the Pearson product moment coefficient of correlation. You have been told that the Pearson r is a meaningful index of the relationship between two variables but the data must meet certain underlying assumptions.

These basic assumptions are:

- i. The relationship between the two variables is linear or rectilinear. In other words the relationship is one where the plotted values of the two variables X and Y tend to scatter along a straight line rather than along a curved line.
- ii. The two distributions are similar in shape. In other words they should be skewed towards the same direction otherwise the Pearson r will underestimate the relationship between the variables.
- iii. The scatter gram is homoscedastic. That is to say that the width of the patterns of dots is the same in all the parts of the scatter gram.
- iv. The scores have been obtained in independent pairs each pair being unconnected with other pairs
- v. The two variables correlated are continuous.

3.6 Factors Influencing the Correlation Coefficient

Now that you have seen the assumptions underlying the Pearson r, let us move further to look at the factor influencing the correlation coefficient. Whenever you are interpreting a correlation coefficient you should always consider the nature of the population in which the two variables were observed. The correlation coefficient observed between two variables will vary from one population to another because:

- i. the basic relationship is different in different population or
- ii. the variability in the populations differs. Or
- iii. the correlation of the two variables is influenced by their relationship with a third variable.

4.0 Conclusion:

In this unit you have learnt some of the measures of association. You have learnt that these measures seek to find out the tendency of scores of two variables to change together either in the same direction or in the opposite direction. This is made possible because of the understanding that scientific progress depends upon finding out the relationship between things.

Correlation is one of the best statistical methods used for finding out these relationships especially in psychology, Education and the behavioural sciences. In the physical sciences such as Biology, Chemistry or Physics, we have measures which give perfect relationship. e.g. The longer a piece of metal the heavier, therefore the volume of the metal. But the height of a man can not be in perfect relationship with the weight of the man. So in the social or behavioural sciences we rarely have perfect association..

5.0 Summary

In this unit you have gone through the concept of correlation which refers to the extent to which two variables are related in a population or sample. You have seen that correlation can be illustrated graphically using the scatter gram or the bivariate frequency distribution where the values of the two variables X and Y for each member are plotted as points or dots or marked X. Mere inspection of the construction will show the type and direction of the association. If the plotted points run from lower left to upper right, it indicates positive relationship but if the points run from upper left to lower right, it is negative correlation, and if the points are scattered in a random fashion all over the graph, It is indicative of zero correlation between the variables. When the plotted points are close to a straight line, it shows perfect correlation. When the points are removed from a straight line, the degree of relationship is less. But to get the index of the degree of relationship, we apply the coefficient of correlation which is a numerical index and the most commonly used is the Pearson product moment correlation coefficient which can use the deviation method or the raw score method.

6.0 Tutor Marked Assignment

Use any of the pearson's method to find the correlation coefficient of the two sets of data below.

X	6	7	7	8	8	9	10	11	12	12	13	13	15	16	18	19	20	21	21
Y	20	16	18	17	17	16	15	13	14	13	10	14	10	9	8	6	7	4	6

7.0 References:

Ary, Donald and Jacobs, L.C (1976). Introduction to statistic: purposes and procedures. New York Chicago..... Sydney, Montreal . Holt Rinehart and Winston.

Guilford, J.P and Fruchter, B. (1978) *Fundamental Statistics in Psychology and Education*. International students Edition. Auckland, Bogotá..... Sydney, Tokyo. McGraw- Hill international Book company

Ogomaka, P.M.C (1990). *Descriptive Educational Statistics: A guide to Research*. Owerria Top book.

Ughamadu, K.A, Onwuegbu, O. Ci Osunde, A.U (1990) *Measurement and Evaluation in Education*. Onitsha . Emba.

Units 12

SOME MEASURES OF ASSOCIATION AND AGREEMENT II

1.0 Introduction

In the last unit you went through the Pearson product- moment correlation coefficient, which we described as the best known and the most frequently used index of relationship. But there are data or situations to which the Pearson r . cannot be applied, and there are instances in which it can be applied, but in which for practical purposes, other procedures are more expedient. The Pearson r . is most defensibly computed when the two variables X and Y are measured on continuous metric scales and the regressions are linear. Many data are in frequencies or are in nominal scales, in this case the Pearson r . cannot be applied. These and other situations such as if X or Y variable are measured:

- i. on an interval or ratio scale, like height
- ii. on an ordinal scale like rank in class
- iii. Dichotomously on a nominal scale, like male female.
- iv. Or on a dichotomy where an underlying normal distribution is assumed (i.e. artificial dichotomy) like pass- fail: in various combinations result in a number of different types of correlation coefficients. In this unit we shall look at two of them- spearman Brown and point biserial correlations.

2.0 Objectives

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

- i. Define a spearmen –Brown correlation
- ii. Calculate and interpret a spearman –Brown correlation
- iii. Define a point-biserial correlation coefficient
- iv. Calculate a point- biserial correlation coefficient

3.1 Types of Values and Correlation Methods.

You have learnt that various types of values or data are suitable for different correlational computations. In other words, you are aware that correlation coefficients and the computational techniques or models for obtaining them do vary. These variations depend on the type of values assigned to or taken by the variables being correlated. The type of values taken by or assigned to variables that may be correlated include:

- i. Continuous values (raw scores).
- ii. Ranked values (ranks)
- iii. Naturally dichotomized values
- iv. Artificially dichotomized values
- v. Categorized values (three or more).

Diagrammatically, Pairs of these variables and the type of correlation coefficient that can be used are shown below.

Types of values a variable takes	i. Continuous or raw scores	ii. Ranks	iii. Naturally dichotomized	iv. Artificially dichotomized	v. Three or more categories
i. Continuous/ raw scores	Pearson r.		Point biserial r _{pbi}	Biserial r _{bi}	
ii. Ranks		Spearman rho ℓ			
iii. Naturally dichotomized	Point biserial r _{pbi}		Phi coefficient \emptyset		
iv. Artificially dichotomized	Biserial r _{bi}			Tetrachloric coefficient r _{tet}	
v. Three or more categories					Contingency coefficient c

3.2 Spearman – Brown Rank- order Correlation coefficient

In the last unit you were told that the Pearson product moment correlation coefficient is the most widely used and that the others are adaptations of it. This is true with the spearman- Brown Rank- order correlation coefficient which was developed first by a British psychologist Charles spearman and made popular by both spearman and Brown. It is denoted by the Greek letter rho ℓ . When data from both of the two variables to be correlated are measured on an ordinal scale or rank order scale, the spearman rank coefficient is the technique generally applied.

As a teacher, there are so many situations in which you may have data from ordinal scale for correlations. Such situations may arise when there are questions concerning the relationship between variables on which students can be ranked e.g. questions on interest, sociability, cooperativeness socioeconomic status, ability, attitudes towards issues, adjustments, performance in class among others. In this case the spearman Rank order is used.

You have learnt that the spearman rank is designed for ranked data, it can also be used with interval data that have been expressed as ranks. In this case it is an alternative to the Pearson r , especially when the data are not large i.e. not more than 30.

3.3 The Computation of Spearman Rank Coefficient

We have said that the spearman rank correlation makes use of ranks. The use of ranks instead of the original raw scores results in a marked simplification in the formular for the correlation coefficient. The formular is given by

$$\rho_{\ell} = 1 - \frac{6\sum d^2}{n(n^2-1)} \text{ or } 1 - \frac{6\sum d^2}{(n+1)(n)(n-1)}$$

Where d^2 = difference in subjects rank on the two measures squared
 n = number of subjects in the sample.

Example 12.1

The scores of 10 students in two subjects' physics and Technical Drawing are given below. Compute the correlation coefficient using the spearman rho.

Phy	45	50	80	68	10	42	65	50	25	70
T.D	60	90	60	72	30	88	70	60	40	75

Steps to follow:

- i. Set up a composite table as shown below.

S/No	Scores in		Ranks in		D	D ²
	Phy	T.D	Phy	T.D		
1	45	60	7	7	0	0
2	50	90	5.5	1	4.5	20.25
3	80	60	1	7	-6	36
4	68	72	3	4	-1	1
5	10	30	10	10	0	0
6	42	88	8	2	6	36
7	65	70	4	5	-1	1
8	50	60	5.5	7	-1.5	2.25
9	25	40	9	9	0	0
10	70	75	2	3	-1	1
						97.50

- ii Rank the subjects in physics
- iii Rank the students in T.D (Note the ranks where there is a tie)
- iv Find the difference in the ranks
- v. Find the squares of the differences and.
- vi Find the sum of the squares

vii Using the formular $\ell = 1 - \frac{6\sum D^2}{n(n^2-1)}$ we have

$$1 - \frac{6 \times 97.50}{10(10^2-1)} = 1 - \frac{585}{10 \times 99} = 1 - \frac{585}{990} = 1 - 0.590909 = 0.409091$$

$$= \underline{0.41}$$

Activity 12.1

Calculate the spearman- Brown correlation coefficient of the scores of some students in two subjects X and Y give below.

X	47	71	52	48	35	35	41	82	72	56	59	73	60	55	41
Y	75	79	85	50	49	59	75	91	100	87	70	92	54	75	68

3.3 Point- Biserial Correlation Coefficient rpbi

So far, you have gone through correlation coefficients which make use of two variables that are measured on continuous scales. But some of the times you may be confronted with a situation where you have to deal with one continuous variable measured on an interval or ratio scale and the other variable is dichotomous. A genuine or naturally dichotomized variable has only two possible values such as male- female; graduate – non graduate married- unmarried, urban-rural, smoker- non smoker, good-bad, old-young, fat- thin, long-short etc. These are measured on a nominal scale. Therefore, when you have a continuous variable such as school achievement test versus a naturally dichotomized variable, the correlation coefficient to use is the point- biserial correlation coefficient.

The point-biserial is simply a Pearson product- moment coefficient of correlation computed from data where one variable is dichotomized and the other is normally distributed. The formula, which is derived for it, although some what simpler, is mathematically equivalent to the Pearson formula.

It is give by

$$rpbi = \frac{\bar{X}_p - \bar{X}_q}{St} \sqrt{Pq} \text{ where}$$

St = Standard deviation of the whole set or sample of scores on the continuous variable.

\bar{X}_p = Mean score on the continuous variable of the sub- sample belonging to the natural dichotomy P.

\bar{X}_q = Mean score on the continuous variable of the sub- sample belonging to the natural dichotomy q.

P = Proportion of the number of individuals in sub-sample P to the number of individuals in the whole sample and

q_1 = Proportion of the number of individuals in sub-sample q to the number of individuals in the whole sample.

Example 12.2

Calculate the point-biserial correlation coefficient from a continuous and a genuine dichotomous variable given below.

Individual	1	2	3	4	5	6	7	8	9	10	11	12	12	14
Achievement Test	60	56	51	58	49	48	55	45	47	55	45	50	52	61
Sex: MIFO	1	1	1	0	1	0	0	0	1	0	1	1	1	1

Steps to follow:

- i. Count the number of males $N_1 = 9$
- ii. Count the number of females $N_0 = 5$; $N = 14$
- iii. Find the mean score for males = $\frac{471}{9} = \underline{52.33}$
- iv. Find the mean score for females = $\frac{261}{5} = \underline{52.20}$
- v. Find the standard deviation of the whole set of scores using $\sqrt{\frac{\sum (X - \bar{X})^2}{n}}$

X	$X - \bar{X}$	$(X - \bar{X})^2$
60	7.7	59.29
56	3.7	13.69
51	-1.3	1.69
58	5.7	32.49
49	-3.3	10.89
48	-4.3	18.49
55	2.7	7.29
45	-7.3	53.29
47	-5.3	28.09
55	2.7	7.29
45	-7.3	53.29
50	-2.3	5.29
52	-0.3	0.09
61	8.7	75.69
Σ	732	366.86
\bar{X}	52.3	

$$= \sqrt{\frac{366.86}{14}} = \sqrt{26.204286} = 5.1190122$$

$$= \underline{5.12}$$

- vi. Find the proportion of males = $9/14 = 0.643$
vii. Find the proportion of the females = $5/14 = 0.357$
viii. Find rpb using the formula $rpb = \frac{\bar{X}_p - \bar{X}_q}{St} \sqrt{Pq}$

$$= \frac{52.33 - 52.20}{5.12} \sqrt{0.643 \times 0.357}$$

$$= \frac{0.13}{5.12} \sqrt{0.229551} = \frac{0.13}{5.12} \times 0.479$$

$$= 0.0253906 \times 0.479 = 0.0121621$$

$$= \underline{0.012}$$

Note that the point- biserial correlation coefficient like all other types of correlation coefficients has a theoretical range of +1 to -1. But the size of the coefficient here is dependent upon the proportions P and q in the two categories of the dichotomous variable.

The r_{pb1} can reach ± 1 when p and q are 0.50, that is $p = q$. If the proportions differ from p and $q = 0.50$ it is mathematically impossible for the r_{pb1} to reach ± 1 . Other measures of association or correlation coefficients will be discussed later.

Activity 12.2

The scores of boys and girls in an integrated science performance test for 20 students are given in a table below. Compute the point biserial correlation coefficient.

S/No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Score	70	53	37	68	45	56	38	32	61	48	40	60	30	52	49	63	55	44	62	47
Sex	A	G	B	G	B	G	B	B	G	G	B	B	B	G	B	B	G	G	B	B

4.0 Conclusion

Pearson's product-moment coefficient is the standard index of the amount of correlation between two variables and we prefer it whenever its use is possible and convenient. But you have seen that most of the times, there are data to which this kind of correlation methods cannot be applied, and there are instances in which it can be applied but in which, for practical purpose, other procedures are used. In this unit, you have learnt two of such methods- the Spearman-Brown and the point biserial coefficients of correlation.

5.0 Summary

In this unit you have examined two other correlation coefficients which have been developed for use with particular types of data. These correlation coefficients are derivations of the Pearson product moment correlation coefficient which is the most widely used. The Spearman-Brown rank order correlation coefficient is designed for use with naturally ordinal data or with interval data which have been expressed as ranks. The point-biserial correlation coefficient is designed for the correlation when the data are measured from one continuous variable and one naturally dichotomous variable. It assumes that the dichotomous variable is genuine.

9.0 Tutor Marked Assignment

- Given that a sample of 51 students used as control group in an experiment have the following data: No of boys = 27, No of girls = 24 mean score of boys = 67.8, mean score of girls = 56.6 proportion of boys = $27/51 = .471$. proportion of girl = $.529$ standard deviation of the whole set of scores = 13.2 find the r_{pb1}

2. calculate the spearman- brown correlation coefficient of the data below.

S/No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	12	19	2	17	11	18	6	15	10	3	14	1	9	16	13	4	8	20	7	5
Y	35	30	35	25	23	22	40	25	37	33	30	40	27	33	37	32	33	28	31	29

7.0 References:

Ary, Donald and Jacobs, L.C (1976). Introduction to statistic: purposes and procedures. New York Chicago..... Sydney, Montreal . Holt Rinehart and Winston.

Guilford, J.P and Fruchter, B. (1978) Fundamental Statistics in Psychology and Education. International students Edition. Auckland, Bogotá..... Sydney, Tokyo. McGraw- Hill international Book company

Ogomaka, P.M.C (1990). Descriptive Educational Statistics: A guide to Research. Owerri Top book.

Unit 13

STANDARD SCORES

1.0 Introduction

In the school systems in Nigeria, teachers use raw scores to compare students' performance or academic achievements. This is not very good. It would be wrong to compare a students' performance in different subjects based on raw scores. The spread of scores obtained by students in a particular class in different subjects might be different. Again, there is no common scale of measurement or comparison. In the circumstance it would be more meaningful and useful to bring scores obtained by a student in different subjects or in different test within the same subject to a common scale. Scores by different students in the same or different subjects or different tests can similarly be brought to a common scale. When scores are brought to a common scale, we say they are standardized. In other words it is necessary to convert measurements into standard form which means finding standard scores. With standard scores we can then compare and add meaningfully the scores of students. In this unit you will go through the standard scores which are very common and widely used. They are the Z-score, T score and the stanine scores.

2.0 Objectives

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

- i. Convert given raw scores to T- score
- ii. Convert given raw scores to Z-score
- iii. Explain the stanine scores.

3.1 Standard Scores

You have seen that comparing students using the raw scores is not correct. The standard scores are used very effectively in bringing the students to the same scale and therefore for comparing students performances. The conversion to standard score takes into account some important considerations. It is used to overcome the anomaly involved in ranking of raw scores. Because there are differences in the strength of the question papers, and because of the different difficulty levels of the different items on the different question papers, the standardization of the scores is very necessary.

Standard scores locate an individual in terms of how much above or below the mean his score falls. They are interval measures which serve a purpose similar to that of the ordinal percentile ranks. Standard scores have two characteristics. They are:

- i. They are a direct transformation of raw scores and therefore reflect the magnitude of a score
- ii. Since they are interval measures they can be used in a wide variety of mathematical computations

3.2 The Z- Score

This is the basic standard score. It is a type of standard score norm in which both the raw scores, the mean and the standard deviation are considered in the process. It is a standard score in which a deviation from the mean is expressed in terms of the deviation of a raw score from the mean divided by the standard deviation. It is symbolized by the small letter z, while the formula is:

$$z = \frac{X - \bar{X}}{S} \text{ where } X = \text{raw score}$$

\bar{X} = mean of the distribution
S = standard deviation of the distribution.

Note that the mean is the reference point and the standard deviation is the basic unit for measuring distance from that point. It gives a mean of 0 and a standard deviation of 1.

Example 13.1

Given the scores of 10 students in a test as follows 40, 50, 80, 60, 30, 25, 90, 75, 40, 60. Transform the scores to z-scores.

Steps to follow.

- i. Calculate the mean of the set of scores.
- ii. Calculate the standard deviation of the set of scores.
- iii. Find the deviations from the mean of the scores.
- iv. Divide the sum of deviations by the standard deviation.

X	X	$X - \bar{X}$	$(X - \bar{X})^2$
1	40	-15	225
2	50	-5	25
3	80	25	625
4	60	5	25
5	30	-25	625
6	25	-30	900
7	90	35	1225
8	75	20	400
9	40	-15	225
10	60	5	25
\sum	550		4300
\bar{X}	55.0		

$$S = \sqrt{\frac{4300}{10}} = \underline{20.74}$$

Using $z = \frac{X - \bar{X}}{S}$, we have

$$1. \quad z_{40} = \frac{40 - 55}{20.74} = \frac{15}{20.74} = \underline{-0.723}$$

$$2. \quad z_{50} = \frac{50 - 55}{20.74} = \frac{-5}{20.74} = \underline{-0.241}$$

$$3. \quad z_{80} = \frac{80 - 55}{20.74} = \frac{25}{20.74} = \underline{1.205}$$

$$4. \quad z_{60} = \frac{60 - 55}{20.74} = \frac{5}{20.74} = \underline{0.241}$$

$$5. \quad z_{30} = \frac{30 - 55}{20.74} = \frac{-25}{20.74} = \underline{1.205}$$

$$6. \quad z_{25} = \frac{25 - 55}{20.74} = \frac{-30}{20.74} = \underline{-1.446}$$

$$7. \quad z_{90} = \frac{90 - 55}{20.74} = \frac{35}{20.74} = \underline{\underline{1.688}}$$

$$8. \quad z_{75} = \frac{75 - 55}{20.74} = \frac{20}{20.74} = \underline{\underline{0.964}}$$

$$9. \quad z_{40} = \frac{40 - 55}{20.74} = \frac{-15}{20.74} = \underline{\underline{-0.723}}$$

$$10. \quad z_{60} = \frac{60 - 55}{20.74} = \frac{5}{20.74} = \underline{\underline{0.241}}$$

Activity 13. 1

Calculate the Z-scores of the following set of scores: 60, 90, 50, 72, 30, 88, 70, 65, 40, and 75.

3.3 T- Score.

From the example above, you have seen that a z- score of 1.205 means that the raw score has a spread or is located 1.205 standard deviations above the mean. In the same way, a z score of -1.446 indicates that the score is located 1.446 standard deviations below the mean. You have also noticed that the z- scores are very low and sometimes have negative scores and decimalized scores. A z-score distribution can therefore be transformed to a new distribution where there are no decimal points and negative values. The decimal point is eliminated by multiplying the z- score by some convenient constant, while the minus sign is eliminated by adding another constant to each z-score. One of the most popular and convenient transformations is to convert the z-score to a distribution which has a mean of 50 and a standard deviation of 10. it is called T-score or Z-score. It is given by the formula, $T = 10z + 50$ or $T = 10\left(\frac{X - \bar{X}}{s}\right) + 50$

Example 13.2

Transform the raw scores below to T-scores. 45, 50, 80, 68, 10, 42, 65, 50, 25, 70.

Steps to follow

- i. Calculate the mean of the set of scores

- ii. Calculate the standard deviation of the set of scores
- iii. Find the deviations from the mean of the score
- iv. Divide the sum of the deviations by the standard deviation
- v. Multiply the result (in iv) above by 10 and
- vi. Add 50 to the result in v above.

X	X	$X - \bar{X}$	$(X - \bar{X})^2$
1	45	-5.5	30.25
2	50	-0.5	0.25
3	80	29.5	870.25
4	68	17.5	706.25
5	10	-40	1640.25
6	42	-8.5	72.25
7	65	14	210
8	50	-0.5	0.25
9	25	-25.5	650.25
10	70	19.5	380.25
\sum	550		4560.50
\bar{X}	55.0		

$$S = \sqrt{\frac{4560.50}{10}} = \underline{21.36}$$

Using $T = 10 \left(\frac{X - \bar{X}}{S} \right) + 50$ we have

$$1. \quad T_{45} = 10 \frac{45 - 50.5}{21.36} + 50 = 10 \frac{-5.5}{21.36} + 50 = 47.43$$

$$2. \quad T_{50} = 10 \frac{50 - 50.5}{21.36} + 50 = 10 \frac{-0.5}{21.36} + 50 = 49.77$$

$$3. \quad T_{80} = 10 \frac{80 - 50.5}{21.36} + 50 = 10 \frac{29}{21.36} + 50 = 63.58.$$

$$4. \quad T_{68} = 10 \frac{68 - 50.5}{21.36} + 50 = 10 \frac{17.5}{21.36} + 50 = 58.19$$

$$5. \quad T_{10} = 10 \frac{10 - 50.5}{21.36} + 50 = 10 \frac{-40.5}{21.36} + 50 = 31.04$$

$$6. \quad T_{42} = 10 \frac{42 - 50.5}{21.36} + 50 = 10 \frac{-8.5}{21.36} + 50 = 46.02$$

$$7. \quad T_{65} = 10 \frac{65 - 50.5}{21.36} + 50 = 10 \frac{14.5}{21.36} + 50 = 56.79$$

$$8. \quad T_{50} = 10 \frac{50 - 50.5}{21.36} + 50 = 10 \frac{-0.5}{21.36} + 50 = 49.77$$

$$9. \quad T_{25} = 10 \frac{25 - 50.5}{21.36} + 50 = 10 \frac{-25.5}{21.36} + 50 = 38.06$$

$$10. \quad T_{70} = 10 \frac{70 - 50.5}{21.36} + 50 = 10 \frac{19.5}{21.36} + 50 = 59.13$$

Activity 13.2

Calculate the T-scores for the data below. 50, 13, 80, 45, 60, 65, 70, 22, 38, 55, 18, 80, 75, 84, 42,

3.4 Stanine Scores

You are familiar with some of the examination bodies that operate at the ordinary level in this country- Nigeria. Some of these examination bodies like the West African Examination Council WAEC, National Examinations Council, NECO, National Business and Technical Education Board NABTEB, among others, use a type of transformation of raw scores called stanine scores. These are a standard score system which provides a single digit score scale running from 1 to 9. In other words stanine are number grades ranging from 1 to 9 and the percentage of cases in the stanine are 4,7,12,17,20,17,12,7,4 respectively. This gives a normal distribution where the highest score corresponds to stanine of 9 and the least to a stanine score of 1. The mean is assigned a value of 5 and the standard deviation a value of 2. This system was originally used by the Air force during the World War II. The stanine simply means scores with nine categories, 1 to 9. The lowest stanine 1 represents a score that is 2 or more standard deviations below the mean while the highest stanine 9 represents a score that is 2 or more standard deviations above the mean. But WAEC uses a reversed type of stanine. For WAEC, stanine of 1 represents the lowest score. Thus $A_1 = \text{Stanine 1}$, $B_2 = \text{stanine 2}$, $B_3 = \text{stanine 3}$, $C_4 = \text{stanine 4}$, $C_5 = \text{Stanine 5}$, $C_6 = \text{stanine 6}$, $E_7 = \text{stanine 7}$, $P_8 = \text{stanine 8}$ and $F_9 = \text{stanine 9}$. This system has a very good advantage because it can be used to compare results from a very wide population whose characteristics may not be the same.

Activity 13.3

Make a collection of Examination bodies you know and examine how they treat their raw scores. Compare them with that of WAEC.

4.0 Conclusion

In this unit, you have seen that it is wrong to compare students with the use of raw scores. You have therefore been exposed to some of the most popular standard scores which you can use at any time in comparing your students. As a teacher, you have to use them most of the times. You need therefore to get used to them now.

5.0 Summary

In this unit you have learnt that the standard scores are a direct transformation of raw scores. Differences in standard scores have the same meaning in any part of a distribution. You have also noted that standard scores tell us the number of standard deviations above or below the mean a score is located.

The z-score is the basic standard score with a mean of 0 and standard deviation of 1 the characteristics are:

- i. The unit of measurement is the standard deviation.
- ii. Raw scores above the mean of the original distribution will have positive z- values while scores below the mean will have negative z- values.
- iii. The mean of a z-score distribution is zero
- iv. The standard deviation is 1
- v. A z- score distribution has the same shape as the original raw score distribution
- vi. z-score distribution is an interval scale.

The T-score is a conversion of the z-score which eliminates both the negative values and the decimal points.

The stanine is a transformation of raw scores which makes use of 9 (nine) categories or nine standard scores. It is widely used by examination bodies such as WAEC, NECO, and NABTEB etc.

6.0 Tutor Marked Assignment.

Calculate the z-score and T-score of the data below. 32, 15, 28, 23, 18, 27, 29, 25, 21, 35.

7.0 References:

Aryl, Donald and Jacobs, L.C (1976). Introduction to statistic: purposes and procedures. New York Chicago..... Sydney, Montreal. Holt Rinehart and Winston.

Ughamadu, K.A (1994) Understanding and implementing continuous Assessment Second Edition Benin City. World of Books Publishers.