POSTGRADUATE DIPLOMA IN EDUCATION (PDE)

MODULE 1

PDE 110: STATISTICAL METHODS IN EDUCATION

PDE 110 - STATISTICAL METHODS IN EDUCATION

INTRODUCTION

Educational Research usually is an attempt to answer educational questions in a systematic, objective and precise manner. In order to do this, the research is designed and carried out. Measurements of various shades are carried out. From this, a jumbled mass of numbers are obtained. In order to make meaning out of these numbers, they are organized. From these array of data, calculations are made and relationships described. In answering the research question, some decisions are made. In carrying out all these, the research as needs **Statistics** apart from the appropriate research design.

Statistics means different things to different people. This is because statistics has its tentacles virtually in every pie of human endeavour. Whenever the need for sound judgment and decision making arises in any life situation, reliance on statistics is considered wise. This is because *figures don't lie though liars can figure*!

This module will introduce you to the use of statistics in Educational Research and Decision Making.

UNIT ONE: THE MEANING OF STATISTICS

INTRODUCTION

The word "*Statistics*" conveys a variety of meanings to people. To some, it is a collection of tables, charts, data or numbers. To others, it is an advanced component of mathematics. However, to the researcher, it is a tool for collecting, presenting, and analyzing data which will be used in decision making. Statistics here is seen in its vigorous analytical applicational perspective.

In this Unit, we shall explore the meaning of statistics and some of the concepts associated with it so that we can clearly understand its use, significance and purpose in education.

OBJECTIVES

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

(1) Define the term “Statistics” correctly.

(2) Distinguish between statistics and statistic.

(3) Discuss the place of statistics in education.

(4) Explain the relationship between statistics and probability.

(5) Explain clearly some basic statistical concepts and notations.

WHAT IS STATISTICS?

Everyday, we are bombarded with statements such as:

> *The number of accidents recorded on our road between September and January this year is more than that of the same period last year.*
>
> *The Federal Government is to reduce the civil service workforce by 33% in its reform agenda.*
>
> *The following statistics was provided as the allocation to states in this Quarter etc.*

In all these cases, statistics is used to inform the public. The use of statistics probably begin as early as the First Century A. D., when governments used a census of land and properties for tax purposes. This was gradually extended to such local events as births, deaths and marriages.

The science of statistics, which uses a sample to predict or estimate some characteristics of a population, began its development during the nineteenth century. Statistics is defined as ***the science comprising rules and procedures for collecting, organizing, summarizing, describing, analyzing, presenting and interpreting numerical data which are used in making decisions, valid estimates, predictions and generalizations***.

Apart from using statistics to inform people, it plays a significant role in modern day business and educational decision making and forecasting. Statistical methods offer us the opportunity to evaluate an uncertain future using limited information to assess the likelihood of future events occurring.

Because of this contemporary use of statistics, it has three distinct parts – Descriptive Statistics, Inferential Statistics and Experimental Statistics. In **descriptive Statistics**, the event or outcome of events are described without drawing conclusions. It is concerned only with the collection, organization, summarizing, analysis and presentation of an array of numerical qualitative or quantitative data. Descriptive statistics include the ***Mean, Mode, Median, Standard Deviation, Range, Percentile, Kurtosis, Correlation Coefficient, and Proportions*** etc.

Experimental statistics relates the design of experiments to establishing causes and effects. Such designs as experimental, Quasi-experimental, (factorial, Block, ANOVA, and ANCOVA etc belong to this group).

Inferential Statistics builds on the descriptive statistics by going a step further to make interpretation. The focus of inferential statistics is surmising the properties of a population from the known properties of a sample of the population. Based on probability theory, valid and reliable decisions, generalizations, predictions and conclusions can be made using this statistics.

Inferential statistics find usefulness in stochastic (random) process, queuing theory, game theory, quality control etc. Statistical procedures like **chi-square, t-test**, **f-test** etc belong to inferential statistics.

As a student of education, you need to study statistics because of its usefulness in making predictions and taking decisions on educational matters. Downier and Heath (1970) indicated the following basic reasons for studying statistics:

(i) **Daily Use**: Statistics is of immediate and practical utility. They help the educator to get work down quickly and efficiently. They help the educator in forecasting, testing, record keeping, test reporting and interpretation etc.

(2) **Problem Solving**: When action researches are conducted to solve immediate problems, statistical methods are applied to the data. Issues bordering on curriculum improvement, deciding on a better method of teaching or predicting students' enrolment and the required school plant will involve the use of statistics.

(3) **Theoretical Research:** Theories predict what we expect to observe in specific circumstances. Most researches in the behavioural sciences are now very sophisticated and are therefore more quantitative. Theories therefore serve to organize the information. In order to test these theories in education and the social sciences, we resort to statistical methods. The advantages of statistical methods in research include:

  (i) They permit the most exact kind of description.

  (ii) They force us to be definite and exact in our procedures and in our thinking.

  (iii) They enable us to summarize our results in a meaningful and convenient form. It gives order to our data in order that we can see the forest as well as the individual trees.

  (iv) They enable us to draw general conclusions in accordance to the accepted rules. It further establishes how much faith can be put on the conclusion and how far we can extend our generalization.

  (v) They enable us to predict "how much" of a given event will occur under specified conditions known and measured.

  (vi) They enable us to analyze some of the causal factors underlying complex and otherwise bewildering events. Causal factors are usually best uncovered and proved by means of experiments. In education and social sciences, this may not be possible in most cases. Statistical methods are therefore often a necessary substitute for and as a constant companion of experiment.

Thus, knowledge of some basic statistical procedures is essential for those proposing to carry out research in order to summarized and interpret their data well and communicate their finding.

(4) **Comprehension And Use of Research**

The competent educator and researcher must be able to read, with understanding, reports of applied and theoretical research. Learning in any field comes largely through reading. In any specialized field, reading is largely a matter of enlarging vocabulary. Reading research reports means encountering statistical symbols, concepts and ideas which must be understood. He should also be able to determine when a given statistical procedure had been appropriately used in order to assess the conclusions reached. To do this, he must be at grips with statistical ideas and methods.

(5) **Employment**

Statistical logic, statistical thinking and statistical operations are necessary components of the teaching profession. To the extent that the teacher uses in his practice the common technical instruments, such as tests, the educator will depend upon statistical background in their administration and in the interpretation of the results. Teachers who are unfamiliar with these procedures may have difficulty in evaluating their students' abilities and achievements. They will also find it difficult to review research in their areas of specialization and to acquire up-to-date information. Knowledge of statistics is also advantageous in other employment situations like Engineering, Accountancy, and Economics etc. Statistics has a wider application in several human endeavours.

Training in statistics is also training in scientific method. Statistical inference is inductive inference – the making of general statements from the study of particular cases. Many instances of this are encountered in life and on teaching.

ACTIVITY I

1. Give a clear definition of statistics.
2. Explain lucidly four reasons for studying statistics in education.

THE PURPOSE OF STATISTICS

As earlier mentioned, statistics is used in a variety of forms by different people. However, the primary purposes of statistics are to:

(i) reduce large array of data to manageable and comprehensible form;

(ii) aid in the study of populations and samples;

(iii) aid in making reliable inferences about events based on observational data; and

(iv) help in arriving at valid and reliable decisions and generalizations.

EDUCATIONAL STATISTICS

Educational statistics is simply the application of the science of statistics to solve problems connected with various facets of education. It helps us to organize, summarize, present and interpret results and data from educational measurements. Through it, the degrees of association between educational variables are measured and inferences or predictions made in order to accomplish certain educational tasks.

According to Boyinbode (1984), such educational tasks may include the organization and presentation of data, the measurement and description of individual or group performance, the measurement of relationships, the design of experiments and testing of the significance of its results, the drawing of inferences or the formation of models and educational forecasting.

There are various role players in education – educational managers and administrators, Teachers, Guidance and Career Counsellors, Examiners and Examining Bodies, Researchers, Parents and students. Each of these stakeholders in education will need information in order to perform their roles well. Reliable information will be arrived at through the use of statistics. Also, for them to manipulate the information to a useful and productive end will involve the use of statistics. Thus, each uses statistics in specific ways to achieve specific educational tasks. It is therefore not surprising that statistics is used in education in the following areas:

- Determination of educational needs of the community ---- population, age distribution, state finance, priorities, manpower, growth rate, existing institutions, personnel etc.

- Planning for physical resources (School Plant) i.e. when determining the number of classrooms, the formular below is often applied:

$$R = \frac{C \times P \times D - W}{P \times D}$$

Where

C = *number of streams in the school*

P = *number of periods held per day*

D = *number of school days per week*

W = *number of periods per week spent outside normal classroom teaching for recess, PHE, Gardening, Break, practicals in the laboratory etc.*

- **Planning for Human Resources**

Accurate projections should be made based on population. From these, the number of classes, teachers, students and other non-teaching staff would be determined. The total number of teachers required in a school used to be $1\frac{1}{2}$: 1 between teachers and number of class streams. However, in recent times, *the number of pupils enrolled for each subject offered in the school, the number of periods per subject per week, the level of difficulty of each subject, the level of academic attainment of students in each subject and the content volume of each subject* are input variables in the equation.

The other important areas where statistics is applied in education include:

- Educational Budgeting
- Inspection and school record/keeping
- Test development, test scoring and test reporting
- Continuous assessment and record keeping and reporting.

In all these areas, statistics is applied to solve educational problems by various stake holders. Statistics is therefore of immense importance in education.

ACTIVITY

Briefly discuss the role of statistics in education.

SOME BASIC STATISTICAL CONCEPTS AND NOTATIONS

Variables and Constants:

A variable is a characteristic or property that can take on different values. It refers to a property where by the members of a group or set differ from one another.

Individuals in a class may differ in sex age, Intelligence, height etc. These properties are variables.

Constants on the other hand do not assume different values

Variables could be those that vary in quality or those that vary in quantity.

Quantitative Variables take values that very in terms of magnitude. They are easy to measure and compare with one another. These may be scores obtained in a test, weight, height, age, distance, number etc.

Qualitative Variables are those that differ in kind. They are only categorized. The differences are usually in kind such as marital status, gender, nationality, social economic status, educational qualifications etc.

Quantitative variables may be discrete or continuous.

A **discrete variable** is one which can take only a finite set of values, implying that fractional values are usually not allowed. These variables are generated by a counting process usually in whole numbers i.e. the number of goals scored in a football match, the number of teachers in a school, number of girls and boys in a class etc.

A **continuous variable** is that which can take on any value over a range of feasible values. Measured data can be whole numbers or fractions 1 – 2 weight, height, distance values etc.

Variables could also be dependent or independent depending on their functions in a given context. A variable that is dependent in one context may be independent in another.

The **Independent Variable** is one that is manipulated or treated. The effect of this manipulation is manifested on the **dependent variable**. The value of the dependent variable thus depends on that of the independent variable. Also, the value of the dependent variable is usually predicted from that of the independent variable. When comparing the effects of two teaching methods on students' learning achievement, the teaching methods are the independent variables while learning achievement is the dependent variable.

Note that in graphing, the dependent variable is placed on the vertical, Y – axis while the independent variable is placed on the horizontal, X – axis.

There are two types of independent variables – Treatment or Active variables and Organismic or Attribute variables.

Treatment or Active Variable is defined as one that can be directly manipulated by the researchers and to which he or she assigns subjects. This group includes method of teaching, method of grouping and reinforcement procedures.

Organismic or Attribute Variables are those variables that cannot be actively manipulated by the researchers. These variables are sometimes called assigned variables and they are characteristics of individuals that cannot be manipulated at will. Such independent variables as age, sex, aptitude, social class, race, and intelligence level had already been determined but the researchers can decide to include or remove them as variables to be studied.

Confounding Variables: confounding variables are those aspects of a study or sample that might influence the dependent variable or the outcome measure and whose effect may be confused with the effects of the independent variable. There are two types of these – Intervening and Extraneous variables.

Intervening variables are those variables that cannot be measured directly or controlled but may have an important effect upon the outcome. They are usually modifying variables that interfere between the cause and the effect.

These may include anxiety, fatigue and motivation. These variables cannot be ignored in experiments and must be controlled as much as practicable through the use of designs.

Extraneous variables: These are variables not manipulated by the researchers (uncontrolled variables) that may have a significant effect on the outcome of a study. These may include such variables as teacher competence or enthusiasm, the age, socio-economic status or academic ability of the students in the study.

Though it is impossible to eliminate all extraneous variables in a classroom research, using robust experimental designs enables the researcher to neutralize their influence to a large extent. Some other methods include removing the variable, randomization, matching cases, group matching or balancing cases and analysis of covariance.

Data: This is a collection of information, qualitative or quantitative

Distribution: This is the arrangement of a set of numbers classified according to some property.

Population: This refers to the group of measurements that are of interest i.e. the aggregate of units to be covered. This may be people, objects, materials, measurements or things.

Populations could be finite or infinite. When the population is not too large and can be easily counted then it is finite i. e. number of students in a school, number of candidates that wrote an examination etc.

However, when the members of a pupation are so large say like the grains of sand or number of women in West Africa, we say it is infinite.

Sample: This is a part or subset of a population. It is any subgroup or sub aggregate drawn by some appropriate method from a population. The sample is usually the portion of the population appropriately selected for observation.

Parameter: This is a descriptive measure or characteristic, true value of a population. When such characteristics as mean, standard deviation or variance of a population is computed they are called parameters.

Statistic: This refers to a descriptive measure or characteristic of a sample

When we calculate the average age of candidates who wrote JME, we are talking of a parameter. However, if we compute the average age of candidates from a given school or state then the average age is a statistic. Note that this is also different from statistics as a discipline.

Different Symbols are used to denote statistics and parameter:

characteristics	Parameter	Statistic
Mean	μ	$\bar{X}$ or M
Standard Deviation	δ	SD or S
Variance	δ^2	SD^2 or S^2

Dichotomy: A categorical variable with only two categories- i. e. Male/Female

Categorical Variable: A nominal variable on which positions and scores are not recorded as numbers.

Scores: Any position on a numerical variable

Skewness of a Distribution: This is a distribution having a longer tail at one end than at the other. It is an asymmetrical distribution.

Kurtosis: This is the extent of peakedness in a distribution

Normal Distribution: This is a symmetrical distribution having its mean, mode and Median equal. Also, the frequencies of the variable extend equally both to the left and to the right of the mode.

Parametric Tests: These are tests whose efficacy tests whether the variable being studied is at least approximately normally distributed.

Non-Parametric Tests: These tests are developed without reference to the distribution of variables.

x: a variable

f: frequency of occurrence or observations

n: the sample size in the number of observations selected from a population (number of occurrence

N = $\sum$ F: total number of observations comprising a population of interest.

$\sum$: or pronounced as sigma i. e. is a summation sign which instructs us to "take the sum of add

$\sum$ (2, 4, 6) = 10

$\sqrt{\ }$ = square root sign directs us to find the square root of a number i.e

$\sqrt{36}$ = 6

x^2 = this directs us to raise a quantity to the indicated power.

ACTIVITY

Clearly distinguish between

(i) statistics and "statistic"

(ii) continuous and discrete variable

(iii) parametric and non-parametric tests.

(iv) discrete and continuous variables

(v) qualitative and quantitative variables

(vi) dependent and independent variables

SUMMARY

- In this unit, we have defined statistics as the science which comprises the rules and procedures for collecting, organizing, summarizing, describing, analyzing, presenting and interpreting numerical data which are used for decision making, predictions and generalizations.

- The importance of statistics in general terms were discussed.
- Educational statistics is the application of statistics in the field of education.
- The uses and purposes of statistics in education were enumerated.
- Some basic statistical concepts with some statistical not notations were also explained.

ASSIGNMENT

Discuss one application of statistical methods in teacher education.

REFERENCES

Avy, Donal et al (1979): *Introduction to Research in Education* U. S. A. Holt, Rineheat and Winston, Inc.

Best, J. W and Kahn, J. V. (1986): *Research in Education*. London: Practice Hall Inter.

Boyinbode I. R. (`1984*): Fundamental Statistical Methods in Education and Research*, Ile-Ife, DC SS Books.

Gay L. G. (1970) - *Education Research: Competencies for Analysis and Application*. Ohio. Charles E. Merill.

Guilford, J. P. and Fruchter, B. (1973): *Fundamental Statistics in Psychology and Education*.

McCall, R. B (1980): *Fundamental Statistics for Psychology*, U. A. A. Harcourt B. Jovanovich Inc.

UNIT TWO: DESCRIPTIVE STATISTICS

OBJECTIVES

We have seen that various educational data can be obtained in various ways. These data must be summarized and presented in a form that is easily understood. Statistics is used to do this. The type of statistics will depend largely on the nature of data involved.

In this unit, we shall discuss the various scales of measurement, ways of organizing these data and presenting them and the calculations of some statistics.

OBJECTIVES

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

1. describe the four scales of measurement;
2. describe the organization and presentation of data using charts and graphs
3. define terms associated with frequency distribution;
4. construct a frequency table for any set of data;
5. draw a histogram ro represent a given set of data;
6. draw frequency polygons from frequency distributions;
7. draw a frequency curve for a large set of data; and
8. identify the different types of frequency curves.

SCALE OF MEASUREMENT

Quantification has been defined as a numerical method of describing observations of materials or characteristics. When a defined portion of the material or characteristic is used as a standard for measuring any sample, a valid and precise method of data description is provided

Measurement is a fundamental step in the conduct of a research. Measurement is defined as ***the process through which observations are translated into numbers. It is the assignment of numerals to objects or events according to certain rules***. Starting with variables, some rules are then used to determine how these variables will be expressed in numerical form. It may be through tests or actual measurements. The nature of the measurement process that produces the numbers determines the interpretation that can be made from them and the statistical procedures that can be meaningfully used with them. Scientists distinguish among four levels of measurement as categorized in the scales of measurement which are ***Nominal, Ordinal, Interval*** and ***Ratio***.

NOMINAL SCALE

Nominal data are counted data. Each individual can only be a member of mutually exclusive category and not the other. All members of each category include notionally, gender, socio-economic status, occupation, role, religious affiliation etc.

Numbers are often used at the nominal level, but only in order to identify the categories. The numbers arbitrarily assigned to the categories serve mainly as labels or names. The numbers do not represent absolute or relative amounts of any characterization. For instance, the numbers given to football players do not represent their degree of skillfulness but just for recognition and positions.

The identifying numbers in a nominal scale can not be arithmetically manipulated through addition, subtraction, multiplication or division. However, those statistical procedures based on mere counting such as reverting the number of observations in a category can be used.

Thus, with this type of scale, we can only find the mode, percentages, draw charts and may perform chi-square test and some special types of correlation.

ORDINAL SCALE

Nominal scales show that things are different but ordinal scale shows the direction of differences. It shows relative position of one thing to another but can not specify the magnitude of the interval between two measures. Ordinal scales, thus only permit the ranking of items or individuals from highest to lowest. The criterion for highest to lowest ordering is expressed as relative position or rank in a group: 1^{st}, 2^{nd}, 3^{rd} …..nth. This is why ordinal scale is also called **rank order**. Ordinal measures have no absolute values and real differences between adjacent ranks may not be equal. Neither difference between the number nor their ratio has meaning. When numbers 1, 2, 3 and so on are used, there is implication that rank 1 is as much higher than rank 2 as 2 is than 3, and so on.

In ordinal measurement, the empirical procedure used for ordering objects must satisfy the criterion of ***transitivity postulate***. This postulate holds that the relationship must be such that "if object ***a*** is greater than object ***b***, and object ***b*** is greater than object ***c*,** then object a is greater than object ***c***.

This is written as "if ($a > b$) and ($b > c$), then ($a > c$). Other words such as stronger than, precedes and has more attribute than can be substituted for greater than in other situations.

The arithmetical observation of addition, subtraction, multiplication and division cannot be useful with ordinal scales. The statistics that can be used with nominal scale can also be used with the ordinal scale.

INTERVAL SCALE

This is an arbitrary scale based on equal units of measurements which indicates how much of a given characteristic is present. It provides equal intervals from an arbitrary origin.

An interval scale not only orders objects or events according to the amount of the attribute they represent but also establishes equal intervals between the units of measure.

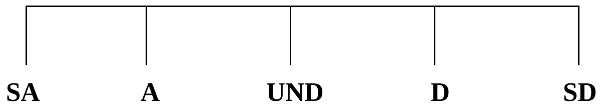
Equal differences in the numbers represent equal differences in the amount of the attributes being measured. The difference in the amount of the characteristics possessed by person with scores of 60 and 65 is assumed to be equivalent to that between persons with scores of 70 and 75. **The limitation here is the lack of a true zero**. The zero point is arbitrary. Interval scale lacks ability to measure the complete absence of the trait and a measure of 30 does not mean that the person has twice as much of the trait as someone who scored 15.

You should note that in most cases where we use interval scales, the intervals are equal in terms of the measuring instrument itself but not necessarily in terms of the ability we are measuring.

Common example of interval data include time and temperature as measured on Centigrade and Fahrenheit scales, scores obtained in achievement tests and other examples.

We can also force ordinal scale into an interval scale as in the case of ratings like:

a)	1.	Strongly agree	b)	Excellent
	2.	Agree		Good
	3.	Undecided		Average
	4.	Disagree		Weak
	5.	Strongly disagree		Poor



If this is regarded as a continuum, where it is possible to choose any point, then we can regard it as interval scale.

Because interval scale lack true zero, multiplication and division of the numbers are not appropriate. This is because ratios between the numbers on an interval scale are measureless.

However, additions and subtractions are possible. Any statistical procedures based on adding may be used with their scale along with the procedures earlier mentioned to be appropriate for the lower level scales. These include mean, standard deviations, t-tests, pearson r, analysis of variance, etc.

THE RATIO SCALE

The fourth and final type of scale is the ratio scale. It provides a true zero point as well as equal intervals.

The numerals of the ratio scale have the qualities of real numbers and can be added, subtracted, multiplied divided and expressed in ratio relationship e.g. 10g is one half of 20g. 30cm is three times 10cm etc.

Examples of ratio data are usually found in the physical sciences and seldom if ever obtained in education and behavioural sciences.

In education, these are limited to educational performance and other physiological measurements. All types of statistical procedures are appropriate with a ratio scale.

ACTIVITY

Describe each type of the measurement scales and give a situation when each can be applied.

THE ORGANIZATION OF DATA

It is always difficult to make sense out of a large data that have not been arranged. This may be data from your research work or students' scores on tests. You need a method to organize the data in order to interpret them. Organizing research data is a fundamental step in statistics. There are two ways of organizing such data:-

(i) arranging the measures into frequency distributions and

(ii) presenting them in graphic forms.

When you have an ungrouped raw data that is few, it is wise to arrange them in descending order of magnitude to produce what is known as an **array of data**. This process of arranging the raw data to get an array of data is called **Ranking**. For example, scores of the students in your class on Statistics are as follows:-

Musa	-	70	Lawrence	-	45
David	-	50	Ade	-	52
Audu	-	88	Ofodile	-	48
Hanatu	-	60	Benedict	-	55
Bunu	-	90	Osun	-	40

Ranking
90
88
70
60
55
52
48
45
40

This is an array of Raw Data.

This array provides a more convenient arrangement. The highest score being 90 and the lowest 40.

From this, the Range can be easily calculated.

The **Range** is the *difference between the highest score, H and the Lowest score, L* which is 90 – 40 = 50

FREQUENCY DISTRIBUTION

Given below is another array of the raw scores of 60 students in another statistics test:

55	60	70	53	40	38	50	49	58	57
52	52	49	44	44	52	49	47	44	55
50	50	50	42	49	52	47	44	53	52
53	52	49	57	53	46	46	55	55	47
53	47	46	55	42	49	50	46	52	58
47	50	47	50	58	53	60	52	57	57

In order to make meaning out of this array of data, you will arrange them from highest to lowest. A systematic arrangement of individual measures from lowest to highest or vice-versa is called a **Frequency Distribution**.

Rank ordering the scores from highest to lowest.

70							
60	60						
58	58	58					
57	57	57	57				
55	55	55	55	55			
53	53	53	53	53	53		
52	52	52	52	52	52	52	52
50	50	50	50	50	50	50	
49	49	49	49	49	49		
47	47	47	47	47	47		
46	46	46	46				
44	44	44	44				
42	42	42					
40							
38							

This may also be put in a frequency distribution table as given below:-

SCORES	TABLES	FREQUENCIES
70	I	1
60	II	2
58	III	3
57	IIII	4
55	~~IIII~~	5

SCORES	TABLES	FREQUENCIES
53	~~IIII~~ I	6
52	~~IIII~~ III	8
50	~~IIII~~ II	7
49	~~IIII~~ I	6
47	~~IIII~~ I	6
46	IIII	4
44	IIII	4
42	II	2
40	I	1
38	I	1

FREQUENCY DISTRIBUTIONS

When summarizing large masses of raw data, it is often good to distribute the data into classes or categories and to determine the number of individuals belonging to each class frequency.

Definition: A tabular arrangement of data by classes together with the corresponding class frequency is called a frequency distribution or frequency table.

As a preliminary of a full scale traffic survey, it was necessary to have some information about the number of occupants of cars entering a certain town on Saturday afternoons, and an occupancy count was made on each of 40 cars. The result were:

1, 3, 2, 2, 3, 1, 1, 2, 2, 1,

1, 4, 3, 1, 3, 2, 3, 2, 2, 2,

1, 2, 5, 1, 3, 1, 2, 1, 3, 1,

4, 1, 1, 3, 4, 2, 2, 1, 1, 4.

Are these variety discrete or continuous?

These are discrete varieties but figures like these dazzle you and you find yourself not able to make any meaning out of numerical data like these just by mere looking at them.

A simple picture of the occupancy of the cars is obtained if the data is given in the form of a table, showing the number of cars with 1 occupant, the number with 2 occupants, and so on. To tabulate the data in this way, you will probably find it easiest to work your way systematically through the 40 counts assigning each to the appropriate category using a tally mark as shown and working with blocks of five to facilitate the final totalling. Infact, the observer might as well have recorded this data in this way in the first place.

Number of Occupants	Tally Stokes	Numbers of cars
1.	~~IIII~~ ~~IIII~~ ~~IIII~~	15
2.	~~IIII~~ ~~IIII~~ II	12
3.	~~IIII~~ IIII	8
4.	IIII	4
5.	I	1

		40

Table 2.1

This is a simple example of a frequency distribution (frequency table). The variate (which will henceforth be denoted by X) is in this case "number of occupants". The number of cars with X occupants shows the frequency with which that value of X occurred. F is usually used for frequency.

In order to get a better picture, raw data can also be grouped into Group Frequency Distribution.

In doing this, we have to decide on the number of Groups required as well as the size of each interval. There is no fixed number of Groups that is appropriate. However, it is advised that between 5 and 20 groups are enough depending on the range of the scores.

GROUPED DATA

Let us consider the result of life-testing of 80 tungsten filament electric lamps. The life of each lamp is given to the nearest hour.

854	1284	1001	911	1168
1357	1090	1082	1494	1684
1355	1502	1251	1666	778
1550	628	1325	1073	1273
1608	1367	1152	1393	1399
1199	1155	822	1448	1623
1058	1930	1365	1291	683
811	1137	1185	892	937
963	1279	1494	798	1599
1281	590	960	1310	1848
1200	845	1454	919	1571

1710	1734	1928	1416	1465
1026	1299	1242	1508	705
1084	1220	1650	1091	210
1399	1198	518	1199	2074
945	1215	905	1810	1265

We now present this data into a grouped frequency table.

NOTE THE STEPS

1. The range (2074-210) is found and divided into 10 groups.
2. Each group has width of 200.
3. The tally method is used to determine the frequency in each group or class.
4. Always check up that the sum of the frequencies is equal to the number of observations in the data, 80 in this case. Table 2.2 shows the grouped data.

STOP! and compare this grouped data with the ungrouped form of the same data. What differences do you observe? We readily observe characteristics of the distribution clearer and faster with the grouped data, and further statistics are readily facilitated, as you will see later in this unit.

Table 1.2 Gr***ouped Frequency Distribution***

Life **X**	**Tally Marks**	**Number of lamps** **F**
201 - 400	I	1
401 - 600	II	2
601 - 800	~~IIII~~	5
801 - 1000	~~IIII~~ ~~IIII~~ II	12
1001 – 1200	~~IIII~~ ~~IIII~~ ~~IIII~~ II	17
1201 – 1400	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	20
1401 – 1600	~~IIII~~ ~~IIII~~ II	12
1601 – 1800	~~IIII~~ II	7
1801 – 2000	III	3
2001 – 2200	I	1 --------- 80 ---------

Table 2.3

TERMS USED IN FREQUENCY DISTRIBUTIONS

The table below is a frequency distribution of masses (to the nearest kg) of 100 male students at a certain College of Education in Nigeria.

Masses of 100 Students at a Certain College of Education

Mass **(Kg) x**	**Number of Students** **F**
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8
	100

Table 2.4

You should notice that with groups for this table defined in the way shown, there is always a gap between the right hand endpoint of one group and the left hand endpoint of the next one (i.e. between 62 and 63, 65 and 66 etc) This may appear to make the data more of a discrete one than continuous one.

However, a life recorded 62 kg would in reality have been between 61.5 and 62.5kg (see rounding off numbers in Unit 1) and similarly 63 Kg covers true values between 62.5 and 63.5 Kg. Thus, in reality the data is a continuous one. The true end points of the groups are as shown with continuous coverage along the time scale.

End Points given as values as measured	End of points Given as true values
60 - 62	59.5 - 62.5
63 - 65	62.5 - 65.5
66-68	65.5 - 68.5

Table 2.5

It is important to choose groups whose end points do not coincide with actual observed data. The above explanation brings us to what is called:

Class Boundaries: These numbers above indicated by the points 59.5, 62.5 etc are called ***class boundaries or true class limits***. The smaller number 59.5 is the ***lower class boundary*** and the larger number 62.5 is the ***upper class boundary***.

How to calculate class boundaries will be discussed later.

Class Intervals and Class Limits: A symbol defining a group such as 60-62 in the above table is called ***class interval or class***. The end numbers 60 and 62 are called ***class limits***; and the larger number is called the ***upper class limit*** while the small one is the ***lower class limit***.

An open class interval: is one which has no upper class limit or no lower class limit such as the class "75 years and over".

The Size of a Class Interval: The size or width of a class interval also referred to as the ***class width, class size*** or ***class strength*** is the difference between the lower and upper class limits. For example in the data of table 2.5 the class interval is.

62.5-59.5 = 65.5-62.5 = 3 etc or

63-60 = 66-63 = 3 or (Since the classes are of equal size)

65-62 = 68-65 = 3

Calculation of Class Boundaries. Class boundaries are obtained by adding the upper class limit of one class to the lower class limit of the next higher class and dividing by 2. For example, the upper class boundary of the first class (60-62) of the data given in table 2.4 is

$$\frac{62 + 63}{2} = 62.5 =$$ The lower class boundary of the second class (63-65).

The upper class boundary of the second class (63 - 65) $\frac{65 + 66}{2}$ = 65.5 = lower

class boundary of the third class (66 - 68) and so on.

The lower class boundary of the first class (60 - 62) is $\frac{59 + 60}{2} = 59.5$

CLASS MARK: The ***class mark*** also called the ***class midpoint*** or ***class centre*** is obtained by adding the lower and upper class limits and dividing by two. Thus, the class mark of the class 60- 62 is

$$\frac{60 + 62}{2} = 61$$

ARRAYS: An array is an arrangement of numerical data in ascending or descending order of magnitude. The difference between the largest and smallest numbers is the ***range*** of the data.

Example: Looking at table 2.5 of the length of life of 80 lamps.

Find

a. The lower limit of the 4th class.

b. The upper limit of the 5th class.

c. The class mark of the 3rd class.

d. The class boundaries of the 8th class.

e. The size of the 6th class.

f. Are all the classes of the same size?

g. The frequency of the 7th class.

h. Which class has the highest frequency?

Solution

a. The 4th class is 801-1000

The lower limit is 801

b. The 5th class is 1001-1200

The upper limit is 1200

c. The 3rd class is 601-800

The class mark is $\frac{601+800}{2}$ = 700.5

d. The eighth class is 1601 - 1800

The lower class boundary is $\frac{1600 \; +1601}{2}$ = 1600.5

The upper class boundary is $\frac{1800 \; +1801}{2}$ = 1800.5

e. The 6th class is 1201 - 1400 and its size is 1400.5-1200.5 = 200

f. To determine if all the classes are equal 600-400=200=800- 600=1000-800 etc OR 401-201 =200=801-601 etc. All the classes are of equal size.

g. The 7th class is 1401-1600 and its frequency is 12.

h. The 6th class 1201-1400 had the highest frequency of 20.

Table 1.4: Marks obtained in Mathematics by 80 Students

Marks *x*	Frequency *F*
50 - 54	1
55 - 59	2

Marks *x*	Frequency *F*
60 - 64	11
65 - 69	10
70 - 74	12
75 - 79	21
80 - 84	6
85 - 89	9
90 - 94	4
95 - 99	4
	80

With reference to this table, determine:

a. The lower limit of the 6th class

b. The upper limit of the fourth class

c. The class mark of the tenth class.

d. The class boundaries of the fifth class.

e. The size of the 9th class.

f. Are all the classes of equal size?

g. What is the frequency of the 6th class?

GRAPHICAL REPRESENTATION OF DATA - THE HISTOGRAM

Data may be presented in two dimensional graphs to make more comparison than is possible with textual matter alone.

There are a number of graphs for doing this. These include line graphs, Bar graphs, Pictographs, Pie graphs, Histogram, Frequency Polygons, Ogive and the smooth curve

Histograms, Frequency Polygon and the smooth curve are most commonly used in education.

HISTOGRAM

The histogram is a graph which uses bars to depict the way two variables are related. Each bar has as their bases the class interval and its length the class frequency.

Example

Let us consider the frequency distribution of the length of life of the lamps earlier discussed:

Definition: The chart of a frequency distribution is called a histogram.

EXAMPLE

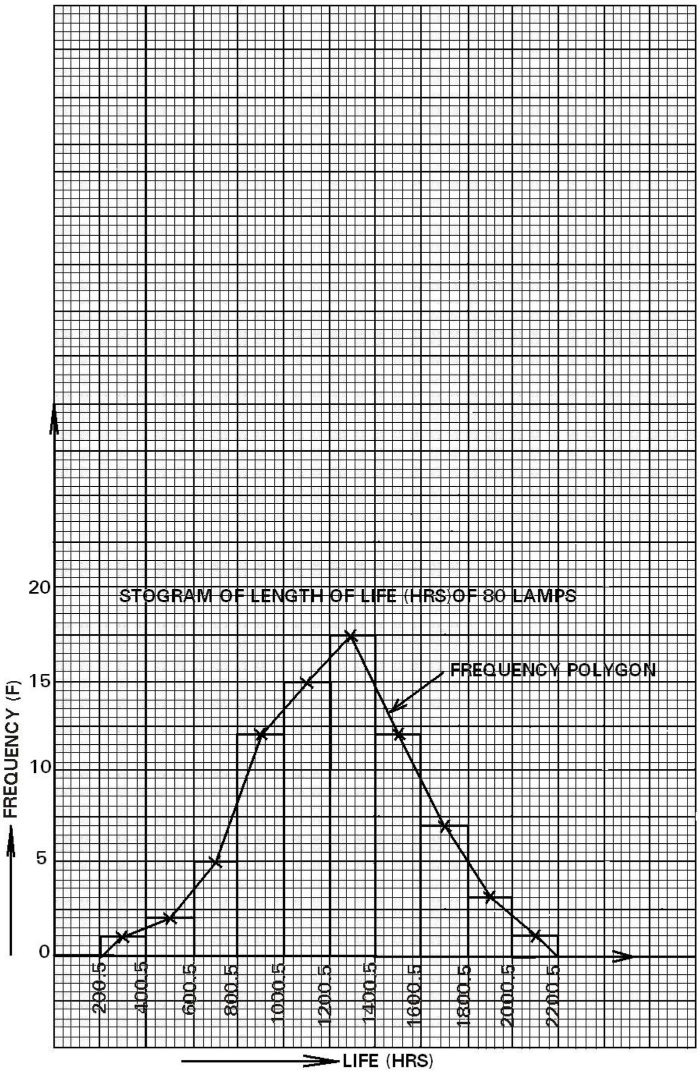


Fig. 1.1

The diagram of the length of life in (hours) of 80 lamps is as shown in Fig.1.1 The base of each rectangle extends from lower class boundary to the upper on a scale representing the variable, in this case the length of life in hours. **The true class boundaries must be used, so that horizontal scale representing length of life is covered continuously with no breaks in between the rectangle.** Notice that the width of the rectangles are equal as shown. This is because the frequency distribution has equal class interval. The bars are of different heights because the frequencies of each class are different.

Looking at the histogram, you will notice that the height of the rectangles represents the frequencies (where classes are of equal size). Note also that the left hand edge on each rectangle represents the lower class boundary and the right hand edge represents the upper class boundary. For the class 1201-1400 AB represents 1200.5 and CD 1400.5 which are the lower and upper class boundaries respectively.

In general, when you combine n classes, the frequency (the height) of the new class becomes $\frac{1}{n}$ of the sum of frequencies.

Thus, if classes two and three are combined, the frequency becomes $\frac{1}{2} \times (2+5) = 3.5$.

FREQUENCY POLYGONS AND FREQUENCY CURVES

Frequency Polygons: The graph of a frequency distribution is called a frequency polygon. The graph is obtained by plotting the class frequencies against the class marks. It can also be obtained by connecting midpoints of the tops of the rectangles in the histogram (where the histogram is already drawn).

ACTIVITY:

Draw the frequency polygon from the histogram of the length of life (in hrs) of 80 lamps.

Solution: All we need do here is join the midpoints of the already drawn histograms. The extremes are adjusted accordingly.

Frequency Curves

Most data are sample of a large population. Where the population is very large many observations are possible, it therefore becomes theoretically possible (for continuous data) to choose class intervals very small and still have quite a number of observations falling within each class. Thus, the frequency polygon for a large population will have so many small broken line segments that they closely approximate curves which we call **frequency curves**.

Frequency Curves can be obtained by smoothing frequency polygons. For this reason a frequency curve is sometimes called a **smoothed frequency polygon**. The smoothing removes irregularities in the curve but still approximates the same area.

ACTIVITIES

1. (a) Arrange the numbers 12,56,42,21,5,18,10,3,61,34,65,24 in an array and

   (b) D etermine the range.

2. If the class marks in a frequency distribution of lengths of laurels are 129, 138, 147, 156, 165, 174 and 183mm, find the class interval size, boundaries and limits.

REFERENCES

Avy, Donal et al (1979): *Introduction to Research in Education* U. S. A. Holt, Rineheat and Winston, Inc.

Best, J. W and Kahn, J. V. (1986): *Research in Education*. London: Practice Hall Inter.

Boyinbode I. R. (`1984*): Fundamental Statistical Methods in Education and Research,* Ile-Ife, DC SS Books.

Gay L. G. (1970) - *Education Research: Competencies for Analysis and Application*. Ohio. Charles E. Merill.

Guilford, J. P. and Fruchter, B. (1973): *Fundamental Statistics in Psychology and Education*.

McCall, R. B (1980): *Fundamental Statistics for Psychology*, U. A. A. Harcourt B. Jovanovich Inc.

UNIT THREE: MEASURES OF CENTRAL TENDENCY AND LOCATION: MEAN, MODE, MEDIAN AND GRAPHICAL LOCATION OF MODE, MEDIAN, QUARTILES, DECILES AND PERCENTILES

INTRODUCTION

We have so far been dealing with the qualitative aspects of a distribution. However, some aspects of a distribution can be described in quantitative terms by calculating certain values from it. An average is a value which is typical or representative of a set of data. Since such typical values tend to lie centrally within a set of data arranged in an array, averages are also called measures of central tendency. There are several types of averages. The most common being midranges, the arithmetic mean, the mode and the median. The unit concerns itself with these averages. These measures reveal the position or length of scores in a distribution.

OBJECTIVES

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

(i) define and calculate the mean, median and mode of a distribution;

(ii) make observations about mean, mode and median of a distribution;

(iii) find the median and mode using a graph; and

(iv) locate the quartiles, deciles and percentiles by means of a graph.

THE ARITHMETIC MEAN

When buying electric lamp bulbs, you can pay a little extra to get the "longer life" type. When tested, the lives (in hours) of 5 "standard" bulbs and 5"longer-life"bulbs were as follows:

"Standard"	1281	1090	1555	1494	1823
"Longer life"	2048	2741	2212	3319	3041

Here, it would be useful to have a measure which, for each type of bulb, would give a general indication of the time lasted. This is sometimes termed a "measure of location", as its aim is to indicate where about the observations are located (in this case, on the time scale). These measures are also called measures of central tendency, since their values tend to lie centrally within a set of data arranged in an array. The measure most often used to meet this need is the arithmetic mean.

There are other types of means such as the geometric mean and the harmonic mean but they are not widely used and it is the arithmetic mean which is referred to when the word "mean" is used

Arithmetic Mean = $\frac{sum\ \ of\ all\ observations}{Total\ \ number\, of\ observations}$

Its symbol is $\overline{x}$ (pronounced x bar). If $\sum$ (pronounced sigma) means sum or addition of a series and there are a set of group of N numbers x_1, x_2, x_3 .... x_7 , then

$$\overline{X} \quad = \quad \frac{\sum_{i=1}^{N} Xi}{N} \quad = \quad \frac{\sum X}{N}$$

The symbol $\sum_{t\,-\,1}^{N}\ Xi$ is used to denote the sum of all Xi's from $i\ =\ 1$ to $i\ =\ N$.

Since the mean is an arithmetic average, it is classified as an interval statistic. Its use is appropriate for interval or ratio data **but not** nominal or ordinal data.

Example: Using the data for the 5"standard" electric lamp bulbs gives their mean length of life as

$$\frac{1281 + 1555\ \ + 1491 + 1823}{5} \quad = \quad \frac{7243}{5} \quad = \quad 1448.6 \text{ hours.}$$

What is the mean length of life of the "longer life" bulbs? You should have done it like this

$$\frac{2048 + 2741 + 2212 + 3319\ + 3041}{5} \quad = \quad \frac{13361}{5} \quad = 2672.2 \text{ hours}$$

More generally if x_1, x_2 ... x_n are n values of a variable x, then their Arithmetic mean $\overline{x}$ is given by

$$\overline{x} = \frac{x_1\ +\ x_2 + x_n}{n} \quad = \quad \frac{1}{n}\ \sum xi \quad \text{which is often written as} \quad \frac{1}{n}\sum_{i=1}^{n} xi$$

PROPERTIES OF THE ARITHMETIC MEAN

1. The first property concerns itself with the deviations of the individual observations from the arithmetic mean. For example, find the deviation of each of the lamp data given for "standard" and "longer life" from their arithmetic means.

   "Standard" -167.6, -358.6,106.4,45.4,374.4

   Sum = -526.2+526.2, = 0

   "longer life" -624.2,68.8, 368.8

   sum = -1084.4+1084.4 = 0

You will notice that in each case above, the negative and positive deviations have exactly cancelled out each other. This does not mean there is anything special about the number.

You can pick some data of your own, find the mean and find the sum of the deviations from the mean. what do you get?

You will discover that in each case the sum of the deviations from the mean is always zero.

Definition: The sum of the deviations from the arithmetic mean is always zero. Thus for any *n* observations $x, x_1 \dots xn$:

$$\begin{aligned}\sum_{j=1}^{n}(xj - \bar{x}) &= (x_1 - \bar{x}) + \quad (x_2 - \bar{x}) \;+\; \dots(x_n - \bar{x}) \\ &= (x_1 + x_2 + \dots\dots + x_n) - n\bar{x} \\ &= n\bar{x} - n\bar{x} = 0\end{aligned}$$

2. **Definition:** The sum of the squares of the deviation of a set of numbers *xj* from any number, *a* is a minimum if and only if *a* = $\bar{x}$

   The above definition and calculation of mean from a frequency table will be dealt with later in this Module

3. The mean is the centre of gravity of the distribution of scores i.e. the measurements in any ample are perfectly balanced about their arithmetic mean.

THE MODE

The mode is the value with the highest frequency, that is, the value that occurs most. The mode may not exist and even when it exists it may not be unique as it may be bimodal etc.

The use of the mode is most often associated with discrete variables. In case of continuous variables, the mode is the value where the frequency density is highest. In a frequency table, it is the interval containing the largest number of observations. In the histogram, the modal group is the one corresponding to the highest rectangle. Mode from diagrams and frequency tables will be dealt with in later units.

The mode is the only measure of central tendency that may be appropriately used with nominal scales. The mode is not often very useful as a measure of central tendencies in distributions because it is unstable as samples drawn from the same population may have different modes. Again, a distribution may have more than one mode.

Example: Find the mode in the following set of numbers.

1. The set 2,3,5,7,9,9, 9, 10,11,12,18.
2. The set 3,5,8,10,12,15,16.
3. The set 23, 4,4,4,5,5,7,7,7,9.

Solution

1. The mode is 9
2. Has no mode.
3. Has two modes 4 and 7, it is therefore said to be bimodal.

THE MEDIAN

The median of a set of numbers arranged in an array is the middle value where the data is odd and the arithmetic mean of the two middle values where the data is even.

Example 1. The set of numbers 4,4,5,7,9,11,13,14,16,18

has median $\frac{1}{2}$ (9+11) = 10. In this case, there are even numbers, therefore, the median is the mean of the two middle values.

Example 2: Find the median of: 14,3,2,1,5,1,7,1,7

Solution: Arrange in ascending or descending order: 1,1,1,2,3,5,7,7,14.

$\therefore$ 3 is the median since it lies in the middle of the arrangement.

Note that repeated scores are written one after the other without missions.

The median is defined as that point in a distribution of measures below which 50 percent of the cases lie and the other 50 percent lies above this point.

In computing median, each score is thought of as representing a range i.e. 3 above represents a range of 2.5 to 3.5; 2.5 is the lower limit of 3 and 3.5 is the upper limit. The median is usually located somewhere between the upper limit and the lower limit of an interval.

It should also be noted that median did not take into account the size of individual scores. It is just the point that divides the distribution into two equal halves. Thus, median is an ordinal statistic since it is based on rank. Though median can be computed from interval or ratio data, the interval characteristic of the data is not used. Median is usually preferred when there are extreme scores in the distribution. The use of any other measure of central tendency which takes into account the size of individual scores will result in either an over estimation or under estimation of the typical score, depending on the distribution.

Therefore, in finding the typical score, the median is most appropriate.

Consider the scores.

40, 41, 42, 43, 44, 80, 95

43 is the median and is he most typical score. Any measure of central tendency that takes 80 and 95 into account will be misleading.

QUANTILES OR FRACTILES

A quartile is a point on a number scale which is assumed to under-lie a set of observations. It divides the set of observations into two groups with known proportions in each group. These are the **Quartiles**, **Deciles** and **Percentiles**. These are also called fractiles.

QUARTILES: Just as the median divides the data into two equal halves, so quartile divides the data into four (4) equal parts.

4th quartile (100%)

3rd quartile (75% below)

2nd quartile (median with 50% of values below)

1st quartile (25% of values below)

DECILE: The decile divides the distribution into ten equal parts. The first decile is such that 10 percent of the values are below it. The 3rd decile have 30% of observation below it. The 5th decile is the median.

- 10th

: 9th

- 8th

: 7th

- 6th

: 5th (median)

- 4th

: 3rd

- 2nd

: 1st

PERCENTILE: Percentile divides the distribution into 100 equal parts. The first percentile has 1% of observation below it. The 50th percentile is the median.

The 25th percentile is equal to the 1st quartile while the 75th percentile is equal to the 3rd quartile.

100th percentile	=	4^{th} quartile
75th percentile	=	3^{rd} quartile
50th percentile	=	2^{nd} quartile (Median)
25th percentile	=	1^{st} quartile

Computation of Quantiles or Fractiles will be dealt with in the next unit.

SOME OBSERVATIONS ABOUT MEASURES OF LOCATION

1. The mean but not the median can be affected by extreme values. Example: Take the numbers 1,2,3,4,5.

$$\bar{x} = \frac{15}{5} = 3$$

Median = $\bar{x}$ = 3

If the set of numbers is now changed to 1,2,3,4,80

The mean = $\bar{x}$ $\frac{90}{5}$ = 18 but the median remains 3.

2. Unlike the mean and median, the mode:

(i) may not exist

(ii) is not always unique.

Example: The set of numbers 1,2,3,4,5 has no mode.

1,1,3,3,5 is bimodal and the mode is not unique since it is more than one, 1 and 3,

3. The sum of the deviations from the mean always equal zero.

4. Effect of adding to, subtracting from, multiplying by or Dividing by a constant leads to the following definitions.

(i) For any set of data, if a constant is added to (or subtracted from) each observation, the corresponding measure of location changes by the same constant.

Example: Take the set of numbers 3,4,7,8,8.

Original Data	**Plus 2**	**Minus 2**
3	5	1
4	6	2
7	9	5
8	10	6
8	10	6
30	**40**	**20**

Mean $\bar{x}$	= 6	8	4
Median $\bar{x}$	= 7	9	5
Mode $\bar{x}$	= 8	10	6

(ii) If each observation is multiplied (or divided) by a constant, the corresponding measure of location must also be multiplied or divided by the same constant.

Example: Using the data above.

Original Data	**Mult. by 2**	**Div. by 2**
3	6	1.5
4	8	2
7	14	3.5
8	16	4

8	16	4
30	**60**	**15**

Mean $\bar{x}$	=	6	12	3
Median $\bar{x}$	=	7	14	3.5
Mode $\bar{x}$	=	8	16	4

GRAPHICAL LOCATION OF MODE, MEDIAN, QUARTILES, DECILES AND PERCENTILES

In the early part of this unit, we defined the mean, median and mode. Quartiles, deciles and percentiles were also defined.

Quickly recall and write down the definition given to each of the above

If you have forgotten, go back to those sections and check. How to find each of them were also discussed, we are now going to find each one of them using the graphical method.

THE MEDIAN

As earlier defined, the median of a set of numbers is the middle value of the set when arranged in order of magnitude. If the set has an even number of items, the median is taken as the mean of the middle two.

Geometric Representation Geometrically, the median is the value of X (abscissa) corresponding to that vertical line which divides a histogram into two parts having equal areas.

Example: The length of 40 laurel leaves are given below.

length mm	Frequency
118 - 126	3
127 - 135	5
136 - 144	9
145 - 153	12
154 - 162	5
163 - 171	4
172 - 180	2

	40

Obtain the median from a histogram.

Solution: See (Fig.5.1) The **X** (abscissa) corresponding to the line CD which divides the histogram into two equal areas is the median. Since area corresponds to frequency in a histogram. CD is such that the total area to the right and left of it is half the total frequency (20)

Thus, ADCB and DFEC corresponds to frequencies of 3 and 9.

Then AD = $\frac{3}{12}\times$ AF $=\frac{3}{12}$ (9) = 2.25. The median is therefore the value OA +AD which equals 144.5 + 2.25 = 146.75 $\approx$146.8mm

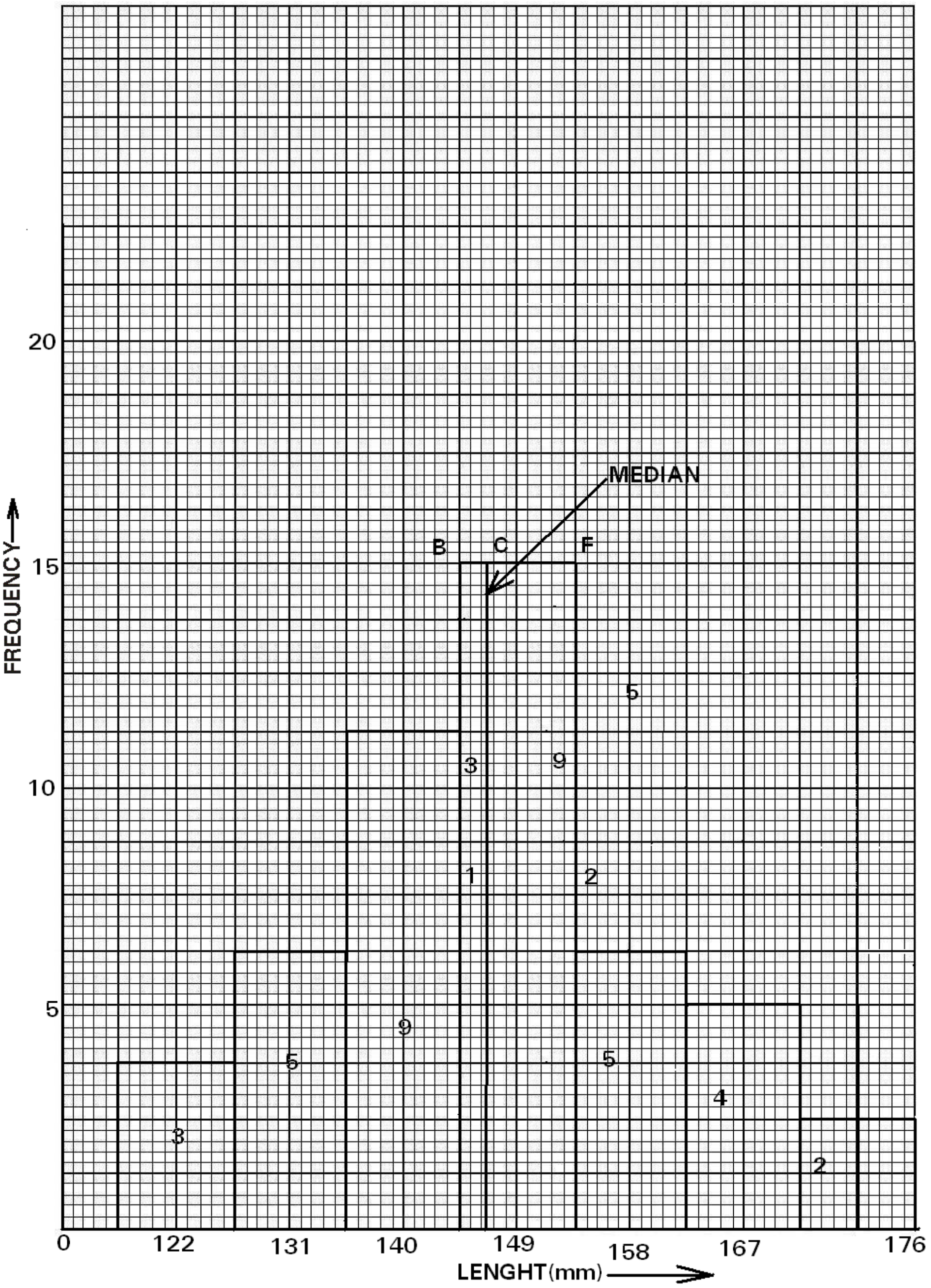


Fig. 3.1

GRAPHICAL REPRESENTATION

Graphically, the median is the value of the X (abcissa) corresponding to that vertical line which (correspond) to the 50th percentile point on the cumulative frequency and divides the frequency into two equal halves.

Example 2: Using the data of the 40 laurel leaves in example 1 above, obtain the median length of the 40 laurel leaves graphically.

Solution: The first step is to draw a smooth cumulative frequency curve or percentage ogive for the given data (see graph - Fig,3.2).

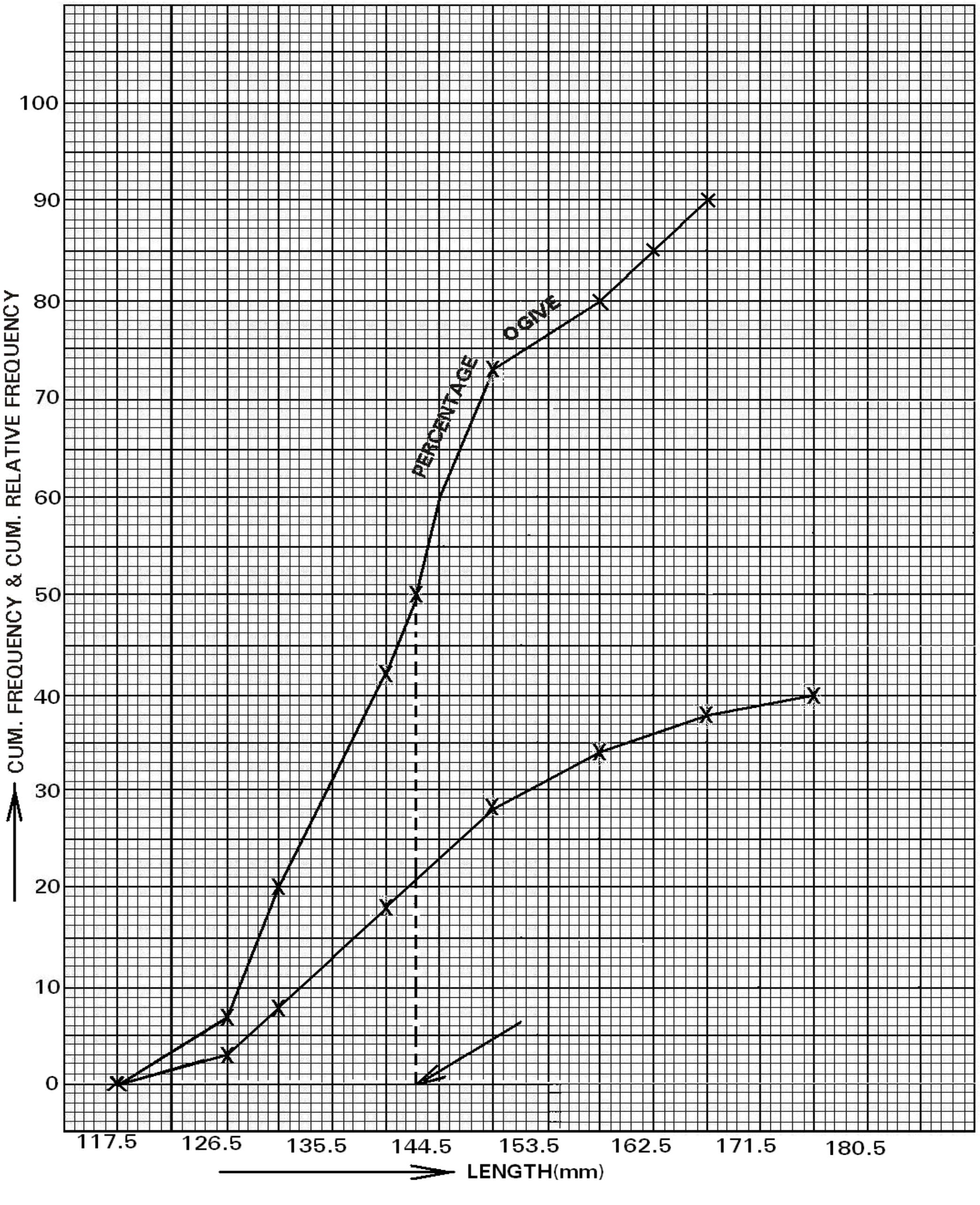


Fig. 3.2

We know that the median should be the 20th item (where N/2) or the 20.5th item (where N+1/2) is used. Therefore, the value corresponding to a frequency of 20 or 20.5 is read along the X axis which is 146.75mm as shown on the graph of Fig.3.2.

A percentage ogive could be drawn and the value corresponding to 50th percentile is read. It is again 146.75mm as shown on the same graph.

THE MODE

As already defined, the mode of a set of values is that one which occurs with the greatest frequency.

Geometric Representation

The mode of a set of values can be obtained from the histogram of the distribution. To illustrate this, we present the following example.

Example: Find the modal age of adult males in a certain company from the following distribution.

Ages	Frequency
21-25	2
26-30	14
31-35	29
36-40	43
41-45	33
46-50	9

Solution: The fourth class 36 – 40 is the modal class since it has the highest frequency. The mode therefore must live within the modal class. To find the mode only the histogram of three classes need be drawn, that is the histogram of the class before the modal class (31-35), the modal class (36-40) and the class after the modal class (41-45). See graph of Fig.3.3.

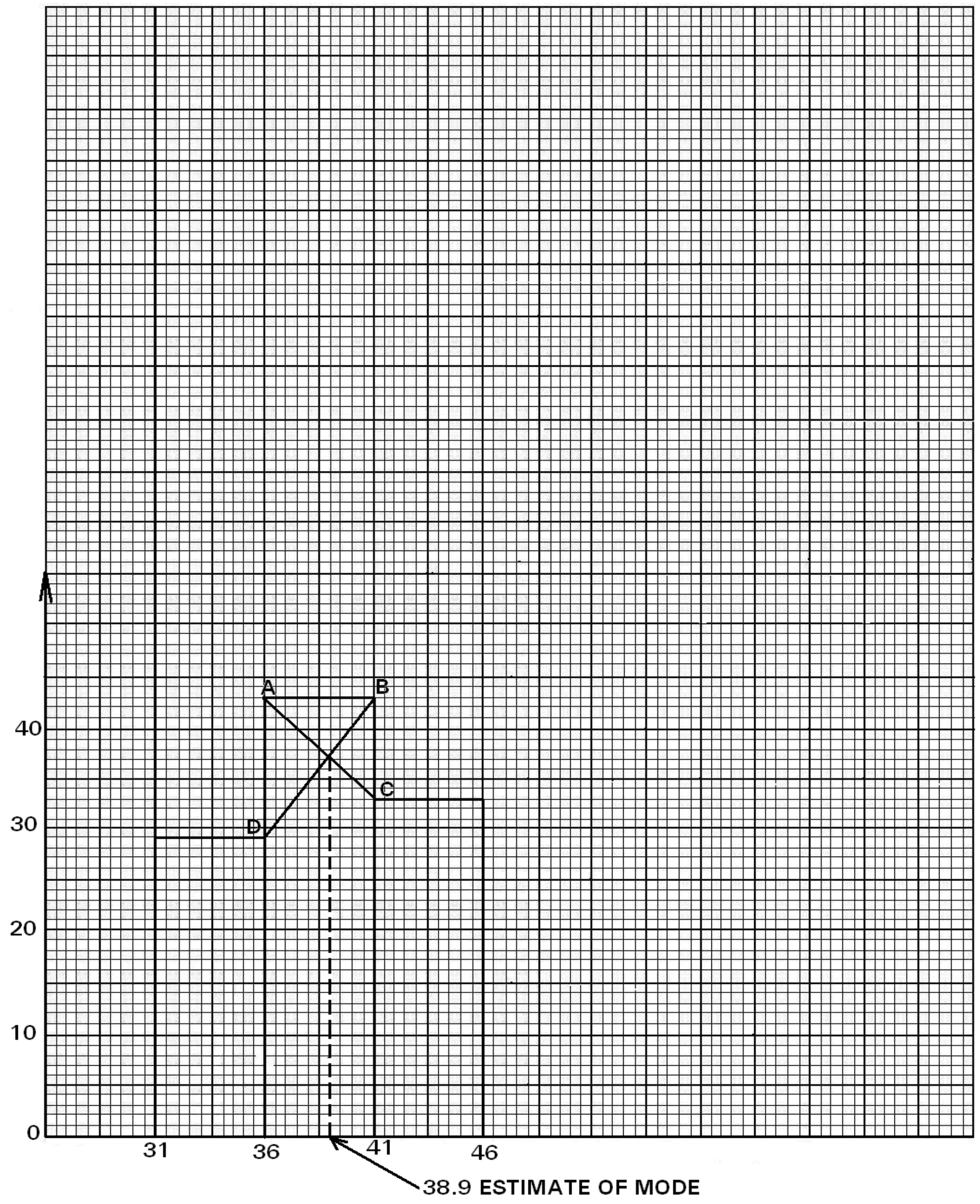


Fig. 3.3

The line AC and BD are drawn. The mode is determined by the X (abcissa) value of their intersection. In this case, the mode is found to be 38.9

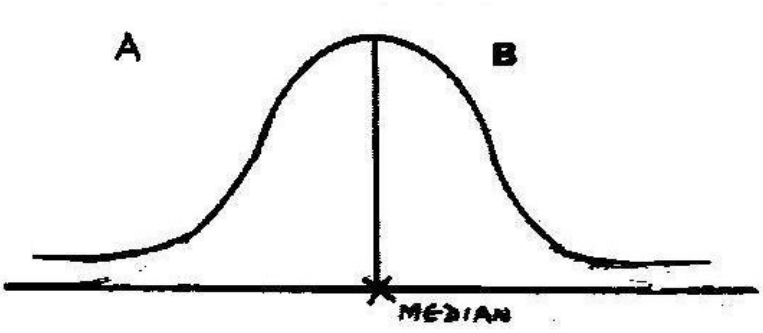
QUARTILES

ACTIVITY

1. What are quartiles? Define quartiles, deciles and percentiles.

If you cannot remember these definitions again, go back to the beginning of this unit and study them again.

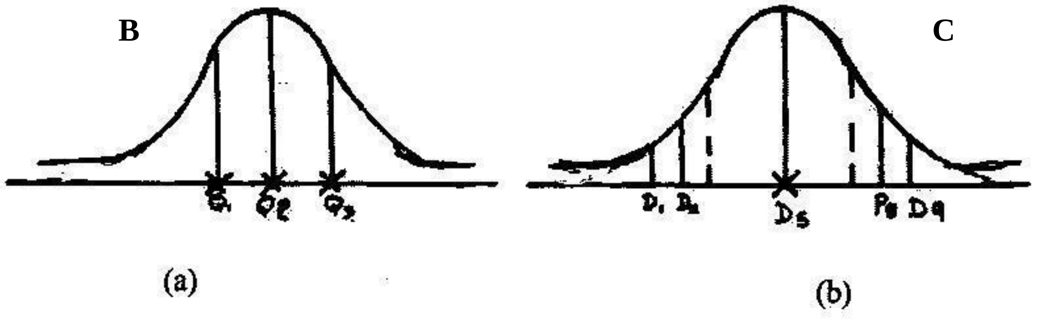
Just as the median splits the area under the curve into equal portions (see diagram below), so also can a frequency curve be splited.



Area A = Area B. (Fig.3.4)

Extending this idea, we can split a frequency curve into as many equal portions as we wish. You will recall that the general name given to those values that split a curve into equal parts are called quantiles. You will also recall the following:-

1. The three values that split a distribution into four equal portions are known as quartiles. In order of magnitude, they are usually represented by Q1, Q2, Q3 and called the first, second and third quartiles, respectively.



(a)

(b)

Fig.3.5

The second Quartile is the median since it divides the area under the curve into two equal portions.

2. The nine values that split a distribution into ten equal portions are known as deciles and are represented by D_1,D_2 -------- D_9. The fifth decile D_5 being the median. See Fig.5.4 (b)

3. The ninety-nine values that split a distribution into one hundred equal portions are known as percentiles and are represented by P_1, P_2,...P_{99} where again P_{50} is the median.

GRAPHICAL REPRESENTATION OF QUANTILES

Example 1: The following data gives the weight of 1200 duck eggs.

Weight (mid point in grams)	57	60	63	66	69	72	75	78	81	84	87	90	93	9 6
No of eggs	7	13	68	144	197	204	208	160	101	54	25	13	4	2

Find the median, quartiles, D8 and P_{37}; the 8th Decile and 37th percentile using graphical method.

Solution:

All we need do is to draw the percentage ogive of the distribution. From the percentage ogive, it becomes relatively easy to find the median which is the 50th percentile. The first quartile is the 25th percentile. The second quartile is the 50th percentile (the median) the third quartile is the 75th percentile. The eight decile D_8 is the 80th percentile and P_{37} the 37th percentile.

(See solution on graph (Fig.2.5) Table is as shown in Table 5.1.

Wight (grams)	**f**	**cf**	**pcf**
58.5	7	7	0.5
61.5	13	20	1.7
64.5	68	88	7.3
67.5	144	232	19.3
70.5	197	429	35.8
73.5	204	633	52.8
76.5	208	841	70.1
79.5	160	1001	83.4
82.5	101	1102	91.8
85.5	54	1156	96.3
88.5	25	1181	98.4
91.5	13	1194	99.5
94.5	4	1198	99.8
97.5	2	1200	100

Table 3.1

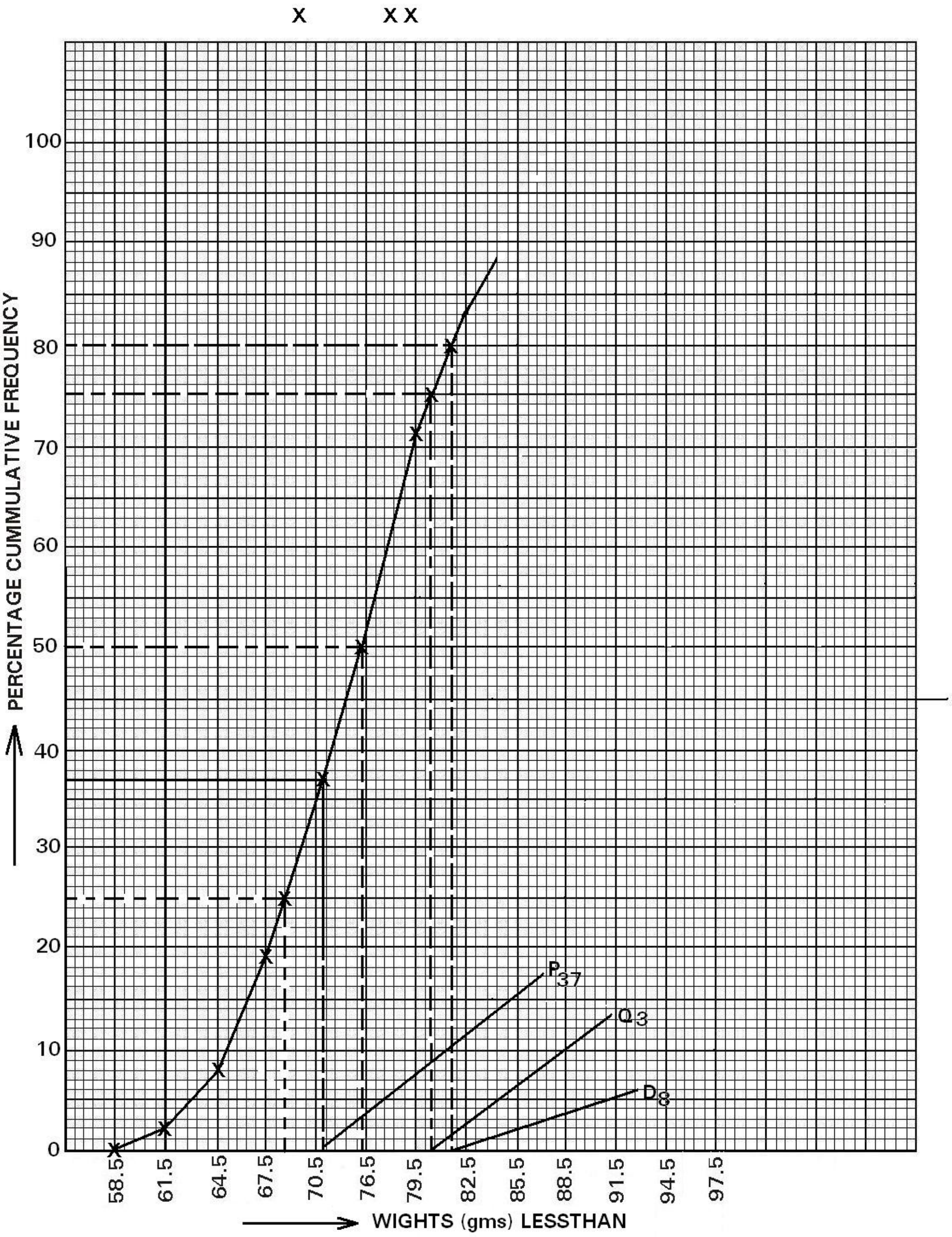


Fig. 3.6

Median	=	73.35gms
1st quartile	=	71.10gms
2nd quartile	=	73.35gms.

3rd quartile	=	77gms
8th decile (D8)	=	79gms
P37,37th percentile	=	70.7gms.

ACTIVITY

1. The annual salaries of five men were N5,500, N4,800, N7,000 N8,000 and N32,000.
   a. find the arithmetic mean of their salaries.
   b. find their median salary.
   c. would you say the mean is typical of the salaries?
   d. which of the two (a) or (b) gives a more reliable average and why?
2. The grades of a student in eight examinations were 50,60,75,85,67,60,56 and 72.
   a. Find the mode of the grades and
   b. Find the median of the grades.
   c. Is the mode unique?

DERIVATION AND USE OF FORMULAE FOR THE MEASURES OF CENTRAL TENDENCY FOR A FREQUENCY DISTRIBUTION

In the early part of this unit, we learnt about measures of central tendency and how to derive them from a set of numbers. Later, we also learnt how to locate them graphically. In this Section, we will learn how to derive them from a frequency distribution.

What is the arithmetic mean or mean of a set of numbers?

If you cannot state what mean is, go back to the opening section of this unit. You will recall that the mean is the sum of all the items in a group divided by the numbers of items in that group.

We had learnt also how to calculate the mean for a set or group. Let us now see how to calculate the mean for a frequency distribution.

MEAN FOR A FREQUENCY DISTRIBUTION

For a discrete frequency distribution taking values (x_1 x_2 ……….. x_n) with corresponding frequencies (f_1, f_2, ……….. f_n), the mean $\bar{x}$ is given by

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} fi}$$

Proof: Now x_1, occurs exactly f_1 times x_2 occurs f_2 times, so that the sum total of all items is $f_1 x_1 + f_2 x_2 + \ldots\ldots\ldots\ldots + f_n x_n = f_i x_i$

and the total number of items is clearly $f_1 + f_2 + \ldots\ldots + f_n = \sum_{i=1}^{n} fi$

But $\bar{x}$ is defined as the sum of all items divided by the number of items hence

$$\bar{x} = \sum_{i=1}^{n} f_i x_i \Big/ \sum_{i=1}^{n} f_i$$

Note: That $\sum$ is a summation notation. The process of adding x_1, x_2, x_3 ….. x_n, can be written as $x_1 + x_2, + x_3 + \ldots + x_n$ and using the $\sum$ notation can be written as

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

which means *i* takes the values 1,2,3 … up to ***n*** and means

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

i.e. the sum of all observations x_1, x_2 … up to and including x_n.

Worked Example

Example. A group of 10 has a mean of 36 and a second group of 16 has a mean of 20. Find the mean of the combined group of 26.

Solution:

X	*f*	*fx*
36	10	360
20	16	320
	26	680

$$\frac{\sum fx}{\sum f} = \frac{680}{26} = 26.15$$

Continuous Frequency Distribution

For a continuous frequency distribution or grouped discrete distribution, the last method cannot be directly used since we do not have distinct *x* values but ranges of values of *x*.

What is done in this case is to simply take the midpoint of the class to represent *x* value and proceed in the usual way as in the last example.

Example 1: The weights in (Kg) of 65 female adults of a certain female adult school is shown in the frequency distribution below. Find their mean weight.

Solution:

Class Weight (Kg)	**Midpoint** ***X***	**Frequency** ***f***	***fx***
5.00 - 5.49	5.245	12	62.940
-5.50 - 5.99	5.745	32	183.840
6.00 - 6.49	6.245	11	68.695
6.50 - 6.99	6.745	8	53.960
7.00 - 7.49	7.245	2	14.490
	65	383.925	

$$\sum f = 65, \quad \sum fx = \mathbf{383.925}$$

$$\bar{x} = \frac{383.925}{65} = 5.91$$

The mean weight is 5.91 Kg

Example 2:

178 people were asked how many coins they had in their pockets and the following results were obtained.

No of Coins	0 - 4	5 - 7	8 - 10	11 - 12
No of people	6	8	8	8

Find the mean number of Coins.

Solution:

Class	Midpoint *(x)*	Frequency *f*	*(fx)*
0 - 4	2	6	12
5 - 7	6	8	48
8 -10	9	8	72
11-12	11.5	4	46
		-------	-------
		26	**178**
		====	====

$\therefore \bar{x} = \frac{178}{26}$ = 6.85 = 7 coins to the nearest whole number of coins

NOTE:

1. The fact that we have unequal class intervals makes no difference to the calculation for the mean.
2. The calculated mean (6.85) is not a typical member of the distribution since the data comprises of whole numbers. However, when calculating statistical measures for discrete distributions, we often give the answer in continuous form unless otherwise specified in which case an approximated mean value such as (7) can be used.

THE CODING METHOD

When dealing with large awkward values of a variable, the calculation of the mean by the methods so far employed can become tedious, for this reason the coding method is introduced.

The method involves subtracting (or adding a number from each of the original values and, if possible and convenient, dividing (or multiplying) these new values by another number to obtain a set of *x* values which should be more manageable. We say that the *x* values have been coded (or transformed) into *x* values. We then find the mean of the *x* values $\bar{x}$ and by using a suitable decoding formula obtain $\bar{x}$.

Definition: If (a) the set $(x_1, x_2\, x_3\, -, x_3)$ is transformed to $(x_1, x_2\, -, x_n)$

or (b) the frequency distribution $\frac{x_1\ x_2\ --\ x_n}{f_1\ f_2\ --\ f_n}$

is transformed to $\frac{x_1\ x_2\ --\ x_n}{f_1\ f_2\ --\ f_n}$ by

means of the coding formula $\bar{x} = \frac{x - a}{b}$

and x is found, we obtain $\bar{x}$ by means of the decoding formula $\bar{x} = a + bx$

NOTE: *a* and *b* are chosen for convenience in order to make the x values as simple as possible.

Example:

Find the mean of the set (15,21,24,27,30,33,36,39,42) using a method of coding.

Solution:

Subtract 27 (a central value) from each item. This is shown in the table below

x	$x - a =$ $x - 27$	$\frac{x-a}{3} = \frac{x-27}{3} = x_1$
15	- 12	- 4
18	- 9	- 3
21	- 6	- 2
24	- 3	- 1
27	0	0
30	3	1
33	6	2
36	9	3
39	12	4
42	15	5

Thus the coding is $\frac{x-27}{3}$ where $a = 27$ and $b = 3$.

$$\bar{x} = \frac{-4-3-2-1+0+1+2+3+4+5}{10} = \frac{5}{10} = 0.5$$

MEDIAN

Recall This

The median of a set of numbers x_1, x_2, --- xn is defined as *the middle value of the set when arranged in order of magnitude and the mean of the two middle values if the set has an even number of items.*

For a Frequency Distribution

For a discrete frequency distribution taking the values (x_1, x_2.….x_n) with corresponding frequencies (f_1, f_2, ….f_n) the median is the $\frac{\sum f + 1}{2} th$ value when the values are ranked.

Here there is distinction as to whether there is even or odd number of items. The $\frac{\sum f + 1}{2}$ is sometimes replaced by

$\frac{\sum f}{2}$ if $\sum f$ is fairly large.

It is usually desirable to include a column of cumulative frequencies when calculating the median for a discrete frequency distribution as shown in the following example.

Example 1:

Find the median of the following discrete distribution.

x	0	1	2	3	4	5	6
f	6	4	10	20	20	30	10

Solution:

(x)	*(f)*	*(Cumf)*
0	6	6
1	4	10
2	10	20
3	20	40
4	20	60
5	30	90
6	10	100

	100	

$N = \sum f = 100.$

$$\frac{N + 1\ th}{2} = \frac{101}{2}_{th} = 50.5th$$

The 50.5th falls at x = 4, the fifth row using the cumulative frequency column.
Hence the median is 4.

Grouped Data

When dealing with a continuous (or grouped discrete) distribution, we can only estimate a value for the median.

Example:

Consider the following distribution.

x	*f*	*Cumf*
10-19.9	2	2
20.29.9	14	16
30-39.9	38	54
40-49.9	23	77
50 59.9	6	83
60-69.9	1	84

N = 84. Therefore the median should be the $\left(\frac{84\ +1}{2}\right)th$ = 42.5th item which falls in the class 30-39.9. This class is called the median class. We need to find where in the median class, the median is expected to lie. From the frequency distribution we see that there are 16 items up to 29.9 and 54 items up to 39.9. We require the 42.5th item.

We therefore need to find m such that there are 42.5 items up to m. Since there are 16 items to 29.5 and 42.5 items to m, there must be 42.5 - 16 = 26.5 items from 29.95 to m. Similarly there must be 54-42.5 items = 11.5 items from m to 39.95. Now there are total of 38 items in the median class, therefore m must lie a fraction $\frac{26.5}{38}$ of the way along 29.95 to 39.95

The actual distance into the class must be 26.5/38 x 10 (since 10 is the class width)

The median therefore lie at a point 29.95 + 10 $\frac{26.5}{38}$ = $\underline{36.92}$

Note that all number in the above expression are well defined quantities, 29.95 is the lower class boundary of the median class. 26.5 is 42.5 -16 that is $\frac{N+1}{2}$ - Cum f up to lower class boundary (*1*cb) of median class. 38 is the median class frequency and 10 is the median class width or interval.

This technique for estimating a median value is called the method of **Interpolation.**

The general formula for working is therefore given for a continuous (or grouped discrete) frequency distribution by.

$$\boldsymbol{m} = \boldsymbol{l}_1 + \left(\frac{\left(\frac{N+1}{2}\right) - (\sum f)_l}{f\ (median)} \right) \boldsymbol{C} \text{ or}$$

$$\boldsymbol{m} = \boldsymbol{l}_1 + \left(\frac{\frac{N}{2} - (\sum f)_l}{f\ (median)} \right) \boldsymbol{C} \text{ where}$$

l_1 = lower class boundary of median class.

N = Number of items in the data

$(\sum f)_{l\,,}$ = Sum of the frequencies of all classes lower than median class

f (median) = frequency of median class

C= median class width

Example: Find the median length of 40 laurel leaves using interpolation formula and interpolation method.

Length mm	f	cf
118-126	3	3
127-135	5	8
136-144	9	17
145-153	12	29
154-162	5	34
163-171	4	38
172-180	2	40
	40	

Solution: We include the cumulative frequency column and find the following:

Using Formula:

$N = 40, \frac{N+1}{2} = 20.5$ or $\frac{N}{2} = 20$

l_1 = 144.5

$(\sum f)_1$ = 17

f median = 12

C = 9

$$m = l_1 + \left(\frac{\left(\frac{N+1}{2}\right) - (\sum f\ )\ 1}{f\ (median)} \right) C$$

$$144.5 + \left(\frac{20.5\ - 17}{12} \right) 9\ = 147.12\text{mm}$$

Using Interpolation:

The median is $\frac{N\ +1}{2}$ item = 20.5th item

Now the sum of the first three classes' frequencies is 17(i.e. 3 + 5 + 9). To give the desired 20.5 we require 3.5 more of the 12 cases in the fourth class. The median must therefore lie $\frac{3.5}{12}$ of the way between 144.5 and 153.5

The median therefore is:

$$144.5\ +\ \frac{3.5}{12}\ (153.5 - 144.5) = 147.12\text{mm}$$

QUARTILES

We just found how to calculate the median using a formula. Let us now look at other quantiles. For small sets of data, the value of calculating quantiles such as deciles or percentiles is not significant. However, this becomes useful for frequency distributions with large number of items. Their location in an ordered set or a frequency distribution is calculated in a manner similar to that of the median.

Since quartiles split a set of distribution into four equal portions, the first and third quartile Q_1 and Q_3 will be 1 $\left(\frac{n+1}{4}\right)^{th}$ and 3 $\left(\frac{n+1}{4}\right)^{th}$ items respectively in a distribution.

Similarly, D_7 will be the 7 $\left(\frac{n+1}{10}\right)^{th}$ item

Also, P_{23} is the 23 $\left(\frac{n+1}{100}\right)^{th}$ item.

In general, if a particular quantile splits a distribution into ***S*** equal parts the ***j***th quantile of the set will be the $j\left(\frac{n+1}{S}\right)^{th}$ item of the size ordered distribution.

Let us now look at the following grouped distribution.

x	***f***	***Cumf***
70-72	5	5
73-75	18	23
76-78	42	65
79-81	27	92
82-84	8	100
	100	

In this distribution, $n = \sum f = 100$

The 1st quartile Q_1 is given by $\frac{100+1}{4}th$ and

the 3rd quartile Q_3 is given by $\frac{3(100+1)}{4}th$ items

It follows therefore that Q_1 is the 25.25th item and Q_3 the 75.75th item.

Since Q_1 occurs in the class 76 to 78 it is the Q_1 class, similarly 79-81 is the 3rd quartile Q_3 class.

The general formula similar to the median interpolation method for obtaining the 1st and 3rd quartiles are as follows:

$$Q_1 = l_1 + \left(\frac{\frac{N+1}{4} - (\sum f)_1}{fQ_1}\right)\boldsymbol{C_1}$$

$$Q_3 = l_3 + \left(\frac{3\frac{N+1}{4} - (\sum f)_3}{fQ_3}\right)C_3$$

Where l_1 and l_3 are the lower class boundaries of the 1st and 3rd quartile classes.

N = Total number of items in the distribution, $(\sum f)_1$, and $(\sum f)_3$ = cumulative frequencies lower than the respective quartile classes.

fQ_1, and fQ_3 = frequencies of 1st and 3rd quartiles; C_1 and C_3 = widths of 1st and 3rd quartile classes.

Example:

Find using an interpolation formula method the median, quartiles and P_{37} of the weight of 1200 ducks given below.

Weight (gms)	***F***	***Cumf***
56 - 58	7	7
59 - 61	13	20
62 - 64	68	88
65 - 67	144	232
68 - 70	197	429
71 - 73	204	633
74 - 76	208	841
77 - 79	160	1001
80 - 82	101	1102
83 - 85	54	1156
86 - 88	25	1181
89 - 91	13	1194
92 - 94	4	1198
95 - 97	2	1200

Solution: The cumulative frequency is calculated above.

(a) Median $= \frac{1200 + 1}{2} = 600.5th$ item

Median class = 71 - 73

$l_1 = 70.5, (\sum f)_1, = 429$, f median = 204

C = 3

Median $= 70.5 + 3\left(\frac{600.5 - 429}{204}\right)$

$= 73.02$

(b) Q_1 is the $\frac{1200 + 1}{4}th$ item $=$ 300.25 the item

Q_1 class $= 68 - 70$

Thus, $l_1 = 67.5$ $(\sum f)_1 = \mathbf{232}$

fQ_1 **=** $197.C = 3$

$Q_1 = 67.5 + 3\left(\frac{300.25 - 232}{197}\right) = 68.54$

$Q_3 = \frac{3(1200+1)}{4} =$ 900.75th item

$l_3 = 76.5$ $(\sum f)_3 = \mathbf{841}$, $fQ_3 = 160$, C $= 3$

$Q_3 = 76.5 + 3\ \frac{(900.75 - 841)}{160} = 77.62$

(c) P_{37} is the 37 $\frac{(1200 + 1)}{100}th$ item 444.37 items

This lies in class 71 – 73

$l_{37} = 70.5$ $(\sum f)_{37} = 429$, $fP_{37} = 204$

$C = 3$

$P_{37} = 70.5 + 3\ \frac{(444.37 - 429)}{209} = 70.72$

THE MODE

The mode of a set of values is defined as the one which occurs with the greatest frequency. For continuous or grouped discrete data, a method similar to interpolation is used.

This is illustrated by the following example. Consider the following distribution.

Class	***f***
21-25	2
26-30	14
31-35	29
36-40	43
41-45	33
46-50	9

The modal class is 36-40 since it is the class with the highest frequency. It is obvious that the modal value should lie in this class. Since the class (44-45) following the modal class is larger than the class (31 - 35) before the modal class, the mode should be larger than the modal class midpoint. The mode is greater or less than the modal class midpoint depending on whether the class following the modal class is larger or smaller than the class previous to the modal class. The figure below illustrates this (see Fig. 4.1).

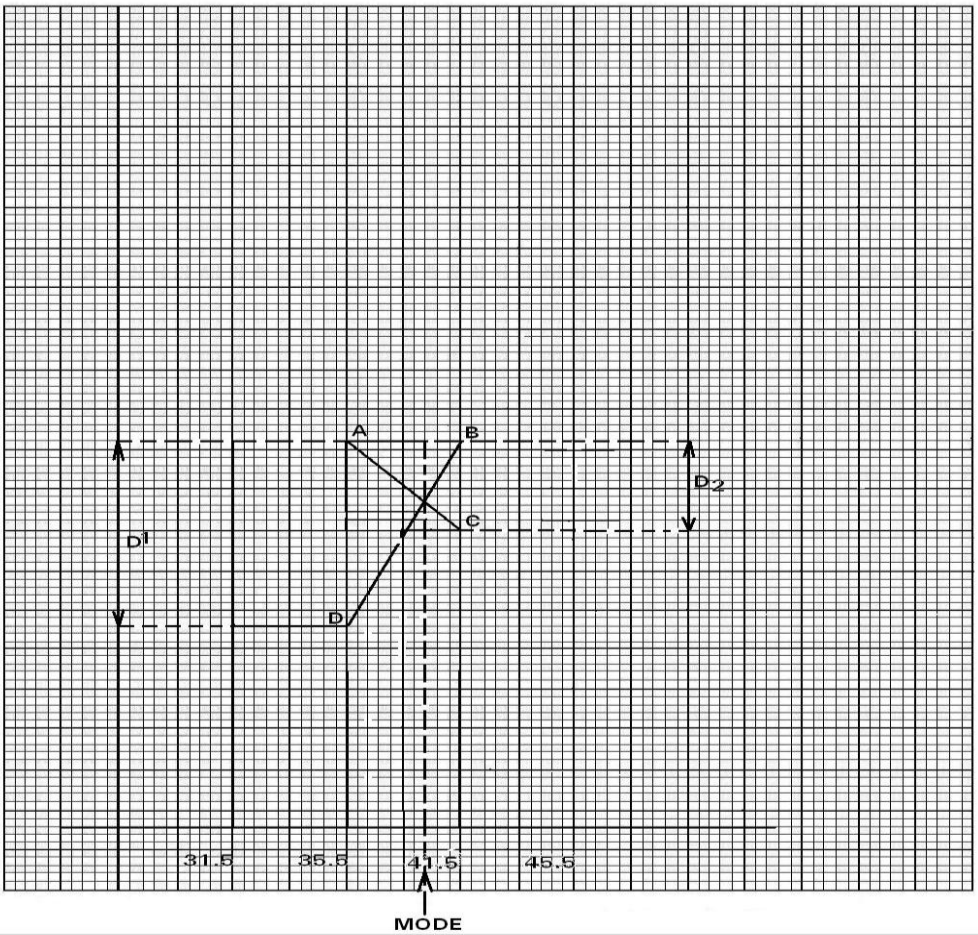


Fig. 3.1

The formula for the mode is given by $l_1 + \dfrac{\Delta_1}{\Delta_1 + \Delta_2}$ where

l_1 = lower class boundary of modal class

Δ_1 = difference in frequencies between modal class and previous class.

Δ_2 = difference in frequencies between modal class and the following class.

c = width of modal class.

NOTE:

The value, $\dfrac{\Delta_1}{\Delta_1 + \Delta_2}$ lies always between 0 and 1

Using the given illustration

$l_1 = 35, \quad \Delta_1 = 43 - 29 = 14$

$\Delta_2 = 43 - 33 = 10$ and $c = 5$

$\therefore$ Mode $= 35 + \left(\dfrac{14}{24}\right)5 = \underline{\mathbf{37.9}}$

Example:

The following are the distribution of marks of 62 students in a statistics test.

x	*f*
9.3 - 9.7	2
9.8 -10.2	5
10.3-10.7	12
10.8-11.2	18
11.3-11.7	14
11.8-12.2	6
12.3-12.7	4
12.8-13.2	1

Find the mode.

Solution: Modal class 10.8 - 11.2

$l_1 = 10.75$

$\Delta_1 = 18 - 12 \;=\; 6 \;\; \Delta_2 \;=\; 18 \;-\; 14 \;= 4$

$c \;=\; 0.5$

$$\text{mode} = 10.75 + \left(\frac{6}{10}\right)(0.5) = 11.05$$

ASSIGNMENTS

1. The weight in kilogrammes, recorded by 50 final year students are as follows:

Weight (Kg)	Number of Students
54 - 57	5
58 61	7
62 - 65	10
66 - 69	12
70 - 73	6
74 - 77	5
78 - 81	4
82 - 85	1

Find the median, Q_1, Q_3, and 60th percentile.

REFERENCES

Avy, Donal et al (1979): *Introduction to Research in Education* U. S. A. Holt, Rineheat and Winston, Inc.

Best, J. W and Kahn, J. V. (1986): *Research in Education*. London: Practice Hall Inter.

Boyinbode I. R. (`1984*): Fundamental Statistical Methods in Education and Research,* Ile-Ife, DC SS Books.

Gay L. G. (1970) - *Education Research: Competencies for Analysis and Application*. Ohio. Charles E. Merill.

Guilford, J. P. and Fruchter, B. (1973): *Fundamental Statistics in Psychology and Education*.

McCall, R. B (1980): *Fundamental Statistics for Psychology*, U. A. A. Harcourt B. Jovanovich Inc.

UNIT FOUR: MEASURES OF VARIABILITY OR DISPERSION, STANDARD SCORES (Z - SCORES AND T – SCORES) AND THE NORMAL CURVE

INTRODUCTION

There is the need to determine the above in any distribution particularly when considering students’ performance.

You are aware that when a teacher or an examiner marks or grades students' or candidates' answer scripts, he/she assigns some marks or scores out of a maximum obtainable score. The fixed maximum obtainable score may be 10, 20. 30, 50, or most often 100. Scores may also be values of a variable (age, height, life span or weight of materials). Scores as presented above are referred to as raw scores. Raw in the sense that such scores are not yet standardized or normed. Performance scores, barring examination malpractices or irregularities, depends upon easiness or difficulty indices of items/tasks and the generosity or severity tendency of the teacher or examiner. Other variable scores may depend upon defects in or errors of reading the calibrations of measuring instruments. All these defects or errors make the interpretations of scores difficult.

Moreso, when a candidate/student gets a score of 70% in an examination, what would you make out of it?

Is the 70% high score in terms of standards of the task undertaken or in relation to the scores of the other candidates/students who also took the examination? Supposing 70% is the highest/greatest or the least score of all the scores earned by all the students, how far apart are the other scores? To overcome the above errors, defects or undue influences on scores, norming and/or standardization of score are devised and used.

OBJECTIVES

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

1. define variability and give its measures
2. calculate standard deviation
3. convert raw scores to z - score and vice versa;
4. transform a given z - score to a T- score and vice versa;
5. convert a raw score overall performance of students in a given set of tests or course as expressed in percentage score and T - scores when necessary statistics are given/obtained.
6. draw the normal curve
7. interpret the areas of the normal curve.

MEASURE OF VARIABILITY (SPREAD) OR DISPERSION

Measures of spread dispersion or variability indicate the degree to which the various points in a distribution deviate from the average. Measures of central tendency only describe a distribution in terms of average value or the typical measure but not the total picture of the distribution. The mean and the median may be identical for some distributions without us knowing their spread. This is why measures of spread are necessary.

For illustration, consider the following distributions of scores of students in two subjects:

		Distribution A	**Distribution B**
		95	76
		90	78
		85	77
		80	71
		75	79
		70	73
		65	72
		60	74
		55	75
$\sum x$	=	675	675
N	=	9	9
$\overline{X}$	=	$\frac{675}{9} = 75$	$\frac{675}{9} = 75$
Md	=	75	= 75

The scores in distribution B is homogonous with little difference between adjacent scores. The scores in distribution are heterogeneous spreading for apart and performance ranged from superior to very poor.

However, the mean and median in both distributions are the same. Therefore, there is the need for the indices that describe the spread or dispersion of scores in a distribution. Several of such measures are available. These include the Range, Quartile Deviation, mean deviation, variance and the standard deviation

Range

The range is the simplest of all indices of variability. It is *the difference between the highest and lowest scores in a distribution.*

The range may be **inclusive** or **exclusive**.

The **exclusive range** is usually quoted as the difference between the largest and the smallest scores in a distribution. However, the **inclusive range** is the difference between the upper boundary of the interval containing the largest score and the lower boundary of the interval containing the smallest score.

The range is a crude way of determining spread as it takes into consideration only the two extremes in a distribution. It is not a stable indicator of the nature of the spread of the measures around the central value. It is usually used in conjunction with such measures as Quartile deviation and standard deviation.

QUARTILE DEVIATION OR SEMI-INTER-QUARTILE RANGE

This is defined as *half the difference between the upper and lower quartile in a distribution*. It is the *difference between the 25th percentile score and the 75th percentile score*.

That is

$$QD = \frac{Q3 - Q1}{2}$$

Recall that the upper quartile, Q3 is the point in a distribution below which 75 percent of the cases lie i.e. 75th percentile.

The lower quartile, Q1 is the point below which 25 percent of the Cates lie i.e. the twenty fifth percentile

$$Q_3 = L + \left(\frac{\frac{3N}{4} - Cfb}{fw} \right) i$$

$$\text{and} \quad Q_1 = L + \left(\frac{\frac{3N}{4} - Cfb}{fw} \right) i$$

where

Q_3	=	the upper quartile
Q_1	=	the lower quartile
N	=	the number of cases in the distribution
L	=	the lower limit of the interval within which the quartile lies
Cfb	=	the communicative frequency below the interval containing the quartile
Fw	=	the frequency of cases within the interval containing the quartile.
i.	=	the interval size.

QD provides a measure of one half of that range of scores within which lie the middle 50% of the cases.

QD, also called semi-inter quartile range, is hardly affected by extreme scores and is more stable. It is therefore preferred to the range. It belongs to the same class as the median with which it is often used.

90TH – TO – 10TH PERCENTILE RANGE (D)

This is defined as the range between the 10th and 90th percentiles in a group of scores

$$D \quad = \quad P_{90} \quad - \quad P_{10}$$

(*D*) is also more stable than the Range because it is affected by a larger number of scores in the distribution than the range.

MEAN ABSOLUTE DEVIATION

This simply is the mean absolute amount by which individual scores in a distribution differ from the mean or the distance of each score from the mean.

The distance of a score from the mean is given by

$$+ \left| Xi - \overline{X} \right|$$

therefore *MD* is given by

$$+ \frac{\sum \left| Xi - \overline{X} \right|}{N}$$

where $\overline{X}$ = the mean

$\left| X - \overline{X} \right|$ = absolute value of the deviations.

MD is no longer in vogue as it has limited statistical use.

Using our earlier example where $\overline{X}$ = 75 to illustrate this, we have

A	x	B	x
X	$\left(X - \overline{X}\right)$	×	$\left(X - \overline{X}\right)$
95	+ 20	76	+ 1
90	+ 15	78	+ 3
85	+ 10	77	+ 2
80	+ 5	71	- 4
75	0	75	0
70	- 5	79	+ 4
65	10	73	- 2
60	15	72	- 3
55	20	74	- 1

$\sum\times \ = \ 675 \qquad \sum\left(X - \overline{X} \right) = 0 \qquad\qquad \sum\times \ = \ 675 \qquad \sum\left(X - \overline{X}\right) \ = \ 0$

$N \ = \ 9 \qquad\qquad\qquad\qquad\qquad\qquad N = 9$

$\overline{X} \ = \ 75 \qquad\qquad\qquad\qquad\qquad\qquad \overline{X} = 75$

It is noteworthy that the sum of the score deviations from the mean equals zero. i.e.

$\sum \left(X - \overline{X}\right) \ = \ 0$ i. e. $\sum x \ = \ 0$

Where *x* is the deviation score.

In other words, the mean is that value in a distribution around which the sum of the deviation scores equals zero. This is an alternative definition of the mean score.

THE VARIANCE AND STANDARD DEVIATION

Variance and standard deviation are measures of variability based on the mean as the point of reference. They also take into cognizance the size and location of individual scores in a distribution.

The main ingredient for computing them is the ***deviation score***. The deviation score (*x*), is the difference between a raw score and the mean i.e.

$$x \ = \ X \ - \ \overline{X}$$

The raw scores below the mean always have negative deviation scores while raw scores above the mean have positive deviation scores.

We have noted that the sum of the deviations from the mean equals zero deviations from the ($\sum x \ = \ 0$). Thus, mathematically, it would be impossible to find a mean value to describe these deviation scores (unless each deviation score yields a positive score). However, as the squares of both negative and positive numbers are positive, the sum of the squared deviation scores will be greater than zero.

The sum of these squares can then be employed to indicate the variability of the scores in a distribution

The mean of the squared deviation scores $\left(\frac{\sum x^2}{N}\right)$ *is called the VARIANCE*

$$6^2 \ = \ \frac{\sum x^2}{N}$$

Where

6^2 = the variance

$\sum$ = the sum of

x = the deviation of each score from the mean $\left(X - \overline{X}\right)$ called deviation score

N = the number of cases in the distribution.

The variance is therefore a value that describes how all of the scores in a distribution are dispersed or spread about the mean. However, since all of the deviations from the mean have been squared to find the variance, it is much too large to represent the spread of the scores. To get the actual picture, the square root of it will have to be calculated.

THE STANDARD DEVIATION

Researchers and educators prefer an index that summarizes the data in the same unit of measurement as the original data. The standard deviation, ***the square root of the variance is*** mostly frequently used as such index to measure the spread or dispersion of scores in a distribution.

Symbolically, this is given as

$$6 = \sqrt{\frac{\sum x^2}{N}} \quad or \quad \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{N}}$$

Using our example, we have

X	x	x^2
95	+ 20	+400
90	+ 15	+ 225
85	+ 10	+ 100
80	+ 5	+ 25
75	0	0
70	- 5	+ 25
65	- 10	+ 100
60	- 15	+ 225
55	- 20	+400
		$\sum x^2 = 1500$

Variance, $6^2 = \frac{1500}{9} = 166.67$

Standard deviation $= \sqrt{\frac{1500}{9}} = \sqrt{166.67} = 12.91$

12-91 does discrites the deviation scores better than 166.67 and the spread of scores in a distribution with arrange of 40.

Though this deviation approach is a clear example of variance and standard deviation it becomes cumbersome when the scores involved are large. Therefore, the following raw scores instead of the deviation scores has been developed. It also eliminates the tedious task of working with fractional deviation scores and gives the same result.

$$\text{Variance, } 6^2 \quad = \quad \frac{N\sum X^2 \quad - \quad (\sum X)^2}{N^2}$$

$$\text{Standard deviation, } 6 \quad = \quad \sqrt{\frac{N\sum X^2 \quad - \quad (\sum X)^2}{N^2}}$$

Note that *X* here is the raw score and X^2 is the square of that score. Using this formular, there is no need to compute the difference between the scores and the mean.

Using the example above

X	X^2
95	9025
90	8100
85	7225
80	6400
75	5625
70	4900
65	4225
60	3600
55	3025

$\sum X \;=\; 675 \qquad \sum X^2 \;=\; 52{,}125$

$N \;=\; 9$

$$6^2 \;=\; \frac{13{,}500}{81} \;=\; 166.67$$

$$6 \;=\; \sqrt{166.67} \;=\; 12-91.$$

The standard deviation is a very useful device for comparing characteristics that may be quite different or may be expressed in different units of measurement.

5D belongs to the same statistical class as the mean i.e. it is an interval or ratio statistic and its computation is based on the size of individual scores in the distribution. It is the most useful measure of variability. *SD* has the following uses:

(1) It is used in comparing the spread of two groups

(2) It is used in comparing the individual in a group to the group as a whole

(3) It is widely used in the computation of other statistics like correlation and in statistical tests of significance.

For group data, the formula for *SD* is given by

$$SD = \sqrt{\frac{N\sum fX^2 - \left(\sum fX\right)^2}{N^2}}$$

Where f is the frequency of the raw score *X* and *N* is the number of

ACTIVITY

(1) Compute the standard deviation of the following scores of 20 students. Group the data using an interval of 5

80	73	69	65
76	72	69	62
74	72	69	62
74	72	68	60
73	70	66	56

(2) Calculate all the measures of variability discussed above for the following distribution: 30,28, 25, 23, 21, 20, 18, 17, 15, 14, 19, 18, 16, 14, 10

OBSERVATION TO NOTE

When discussing the dispersion or spread of a population, you should note that if the variance is small, the scores are close together but if the variance is large, then the scores are more spread out. Likewise, a small standard deviation indicates that scores in the distribution are close together and a large standard deviation shows a distribution with very widely spread scores.

In a normal distribution, an interesting phenomenon is associated with the standard deviation. If you add 3 standard deviations to the mean and then subtracts 3 standard deviations from the mean, the range given encompassed just about all the scores or over 99% of them in the distribution. Symbolically, this is represented as follows:

$$\overline{X} \pm 3SD = 99 + \% \text{ of the scores in the distribution.}$$

Thus, once we know the mean and standard deviation of a distribution, we can adequately describe the set of data.

THE NORMAL CURVE

Normal Distribution

A French Mathematician, **Abraham De Moivre**, discovered that a mathematical relationship explained the probability associated with various games of chance. He subsequently developed the equation and drew the graphic pattern that describes it. This was around 1733.

However, in the nineteenth century, a French astronomer, **LaPlace**, and another German Mathematician, both working independently arrived at the same principle and applied it widely to other physical measurements. This led to the theory of probability density function or the curve of distribution of error. The theory describes the fluctuations of chance errors of observation and measurement i.e. the probable occurrence of certain events.

A probability density function thus indicates the relative likelihood of occurrence of a particular value on the number line by the magnitude of the function associated with that outcome.

This functions was mathematically derived from the formula.

$$Y \quad = \quad \frac{1}{\delta\sqrt{2\Pi}} \quad e^{-\,{}^{x^2}\!/_{25^2}}$$

where

Y = the ordinate for any value of x

II = the ratio of the circumference of any circle to its diameter which equals to 3.1416

e = is the base of the natural logarithm and is equal to 2.7183

x = the deviation of each score from the mean

δ = the standard deviation of the distribution.

When any distribution conforms to this mathematical model, it can be represented by a hypothetical polygon which is bell-shaped with certain characteristics.

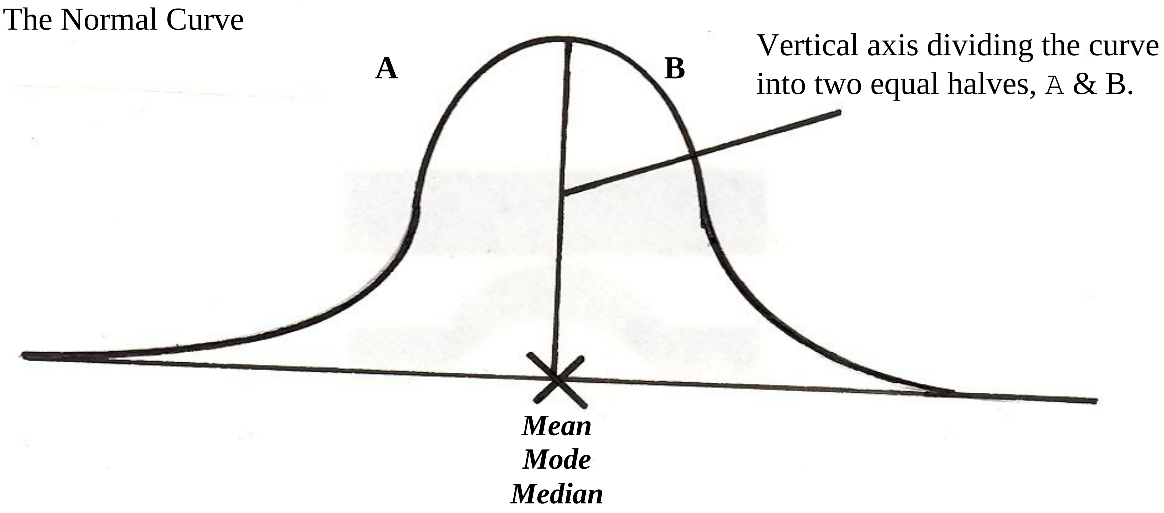

The Normal Curve
A
B
Vertical axis dividing the curve into two equal halves, A & B.
Mean
Mode
Median

Fig. 4.1

From the above diagram you can observe that

(i) The curve is symmetrical around its vertical axis. The percentage of frequencies is the same for equal distance below or above the mean.

(ii) The terms "crowded" or "cluster" around the mean implies that the percentages in any given standard deviation are greatest around the mean and decrease as we move away from the mean.

(iii) The mean, mode and median of the distribution are the same value. The curve is highest at the mean.

(iv) The curve has no boundaries in either direction. The curve, no matter how far it stretches will never touch the base line.

We should note here that while there are some natural human characteristics that are said to be normally distributed in nature, the normal curve itself is never a natural phenomenon. It is just a mathematical model used to explain certain observations in nature.

Some of the human traits that approximate normal distribution include weight, height, age, intelligence, achievement, longevity etc. However, for some of these to be taken as normal distributions, some other factors such as age, race, gender etc would have to be kept constant.

The normal curve is a symmetrical distribution of measures with the same number of cases at specified distances below as above the mean. Approximately 34% of the frequencies in a normal distribution fall between one standard deviation above or below the mean.

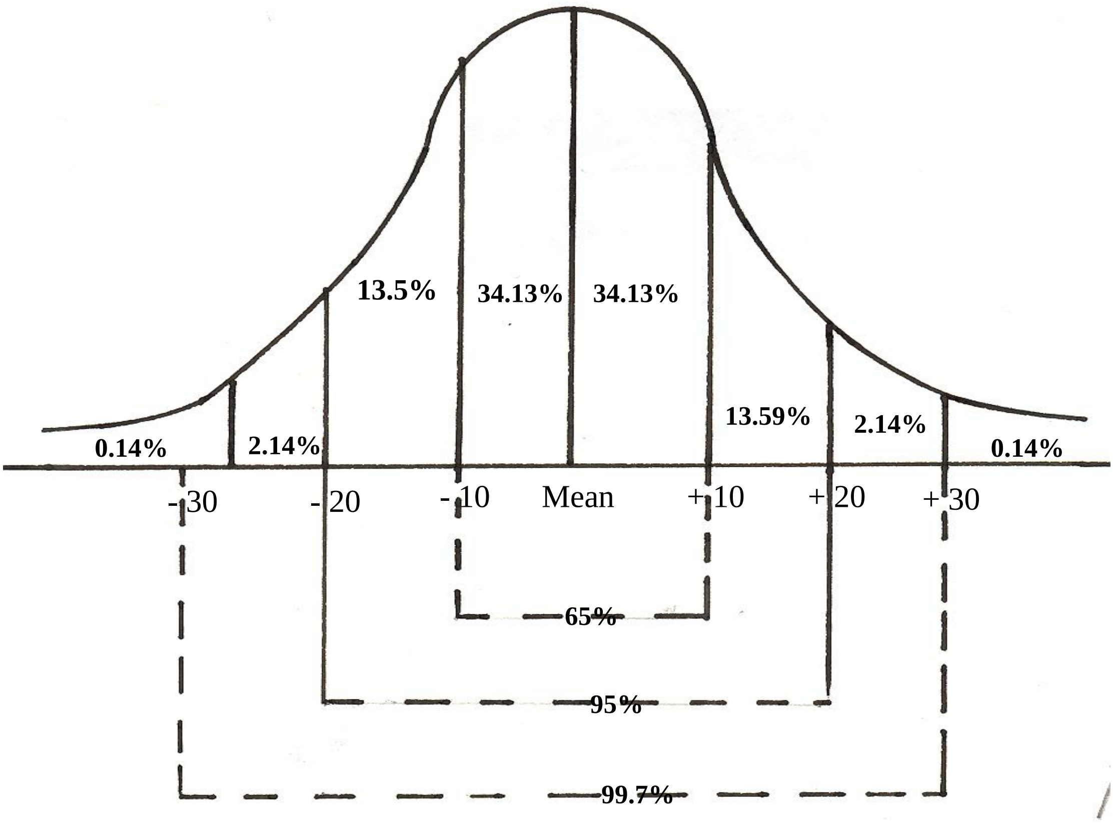


Fig: 4.2 Percentage of Cases falling between successive standard deviation in a normal distribution

The area between one and two standard deviations on either side contains about 14% of the frequencies.

About 2% fall between 2 and 3 standard deviations on both sides.

Only about 0.01% of the frequencies fall above 3 standard deviations below on above the mean.

Reading Table for Areas of Normal Curve

When scores are standardized, then it is possible to determine the percentage of the cases below and above each Z – score in the normal distribution by using the Table. The table shows the areas of the normal curve. (See Appendix)

Column (1) of the table has different Z – values, column (2) gives the areas under the curve between the mean and each Z – value.

Column (3) shows the remaining areas from each 2 – scores to the end of the curve.

The areas in columns (2) and (3) add up to 5000. Column (2) contains the area beyond each Z – score in the deviations opposite to the mean.

Z - SCORE

Because of limitations and errors that are inherent in the interpretation and use of raw scores, standardized scores, Z - scores based on criterion (- task performance and weighting) or ability of a group or composition of objects, have been devised and adopted. Such standard scores include; **stanine scores/ranks**

- (standard nine score/ranks, 9, 8,7,6,5,4,3,2, and 1) WASC results are presented in reversed stanine score/ranks
- percentile scores/ranks,
- z - scores (normal scores) and
- T - scores (-linear transformation of z - scores).

A z-score represents a raw score within a group/distribution of scores. The z - score is calculated/computed by:

i. subtracting the mean of the group of scores from the raw score (which the z - score will represent) and

ii. dividing the difference in (i) above by the standard deviation of the group of scores.

If ***X*** denotes a raw score, $\overline{X}$ denotes the mean of the distribution of scores to which ***X*** belongs and S denotes the standard deviation of the distribution, then

$$Z = \frac{X - \overline{X}}{S}$$

ILLUSTRATION

The table below gives a distribution of percentage scores of students in a mathematics tests.

Group	10-18	19-27	28-36	37-45	46-54	55-63	64-72	73-81	82-90
Freq.	2	3	5	9	12	9	5	3	2

a. Convert (i) 70% and (ii) 35% to z - scores

b. Convert z - scores (i) 2 and (ii) -1.5 to raw scores.

SOLUTIONS

The mean and standard deviation must first be obtained.

Group	*f*	*x*	*fx*	*x*	x^2	fx^2
82-90	2	86	172	36	1296	2592
73-81	3	77	231	27	729	2187
64-72	5	68	340	18	324	1620
55-63	9	59	531	9	81	0729
46-54	12	50	600	0	0	0000
37-45	9	41	369	-9	81	0729
28-36	5	32	160	-18	324	1620
19-27	3	23	69	-27	729	2187
10-18	2	14	28	-36	1296	2592
	50	-	2500	-	-	14256

$$\overline{X} = \frac{\Sigma fx}{Ef} = \frac{2500}{50}, \quad S = \sqrt{\frac{\Sigma fx^2}{\Sigma f}}$$

$$= \underline{\underline{50}} \qquad = \sqrt{\frac{14256}{50}} = \sqrt{285.12}$$

$$S = \approx \underline{\underline{16.89}}$$

a. i. 70%; $z = \frac{70-50}{16.89} = \frac{20}{16.89}$

$\underline{1.18}$

ii. 35%; $z = \frac{35-50}{16.89} = \frac{-15}{16.89} \approx \underline{\underline{-0.89}}$

b. $z = \frac{X - \overline{X}}{S} \Rightarrow Z \times X = X - \overline{X}$

$X = zS + \overline{X}$

. i. z = 2; X=2(16.89)+50 = 33.78 + 50

X=83.78% 84%

ii. $z =- 1.5; X= (-1.5)(16.89) + 50$

$\approx -25.34 +50$

$\approx 25\%$

T - SCORES

As you have seen, z - scores are usually small in magnitude. Generally, z - scores lies within the range -4 to +4, (-4 < z <4). A z - score may be zero or negative while the raw score that is converted is certainly non-zero and positive. These characteristics of z - scores make the use of z - scores inconvenient and difficult to appreciate. T - score overcomes these limitations of the z - score.

T- score is a linear transformation of z - score. Thus,

T = 10z + 50. If generally, $-4.0 \leq z \leq 4.0$ then:

$10(-4.0) + 50 \leq T \leq 10(4.0) + 50.$

That is to say that generally **$10 \leq T \leq 90$.**

ILLUSTRATION

Given that a distribution has a mean of 45 and a standard deviation of 12; convert (i) 79 (ii) 15 that belong to the distribution to T - scores.

SOLUTION

A distribution of percentage scores has a mean 50 and a standard deviation of 15. If two percentage scores from the distribution were converted to T - scores as (i) 20 and (ii) 80, find the raw scores.

SOLUTIONS

You must go through z - scores.

$$T = 10z + 50 \Rightarrow z = \frac{T-50}{10}$$

i. $z = \frac{20-50}{10} = \frac{-30}{10} = \underline{-3}$

Raw score $R = zS + \overline{X}$

$R = -3(15) + 50 = -45 +50$

$R = 5$

ii. 80 as T - score; $z = \frac{80-50}{10}$

$$z = \frac{30}{10} = \underline{\underline{3.0}}$$

z - score of 3. converts to

$R = 3(15) + 50$

$= 45 + 50$

$R = 95$

CHARACTERISTICS/PROPERTIES OF Z - SCORE AND T - SCORE; AND COMPILATION AND COMPARISON OF OVERALL SCORES IN RAW FORMS AND T - SCORES.

A - z - score:

i. range; Generally 4.0 $\leq$ z $\leq$ 4.0;

ii. mean; $\bar{z} = 0$ and

iii. standard deviation $S_z = 1.00$.

B - T - Scores:

i. range - generally $10 \leq T \leq 90$

ii. mean $\bar{T} = 50$ and

iii. standard deviation $S_T = 10$.

ACTIVITY

Given the distribution of scores below.

Group	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45
Freq.	2	4	7	12	15	12	7	4	2

1. Convert (i) 38 and (ii) 7 to Z - scores.
2. Convert (i) 41 and (ii) 18 to T - scores

SKEWNESS AND KURTOSIS THROUGH DIAGRAMS ONLY

Skewness of a Distribution

A. **Symmetry:**

A distribution is said to be symmetric if it is possible to cut its graph into two mirror image halves. Such distributions have bell-shaped graphs. This shape of the frequency curves are characterised by the fact that observations equidistant from the central maximum have the same frequency e.g. the normal curve.

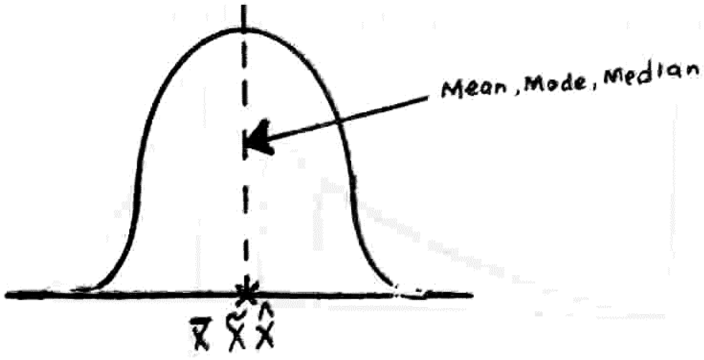


Fig. 4.3. *Symmetric Distribution*

In symmetric curves, the mean, mode and median all coincide at the middle. In this case, the highest frequency lies in the middle and decreases both ways down.

B. Skewness

Skewness is the degree of asymmetry (departure from symmetry) of a distribution. If the frequency curve (smoothed frequency polygon) of a distribution has the length of one of its tails (relative to the central section) disproportionate to the other, then the distribution is described as skewed.

Frequency distributions can be classified according to the general shape of their frequency curves.

(a) **Positive Skewness**: A distribution skewed to the right, that is with the longer tail to the right of the central maximum than to the left, is said to be positively skewed. Fig 4.4 shows a sketch of positively skewed data.

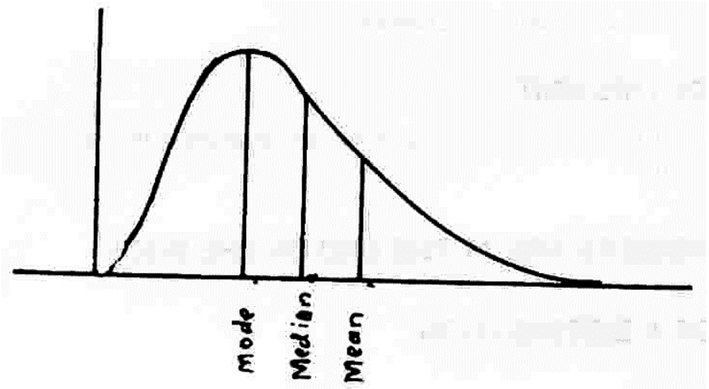


Fig. 4..4 **A** ***Positively Skewed Distribution***

In such a distribution the bulk of the cases fall to the lower part of the range of values and relatively few show extremely high values. In positively skewed distribution the mean has the highest value, followed by the median and the mode has the least value.

(b) **Negative Skewness:** A distribution skewed to the left is said to be negatively skewed.

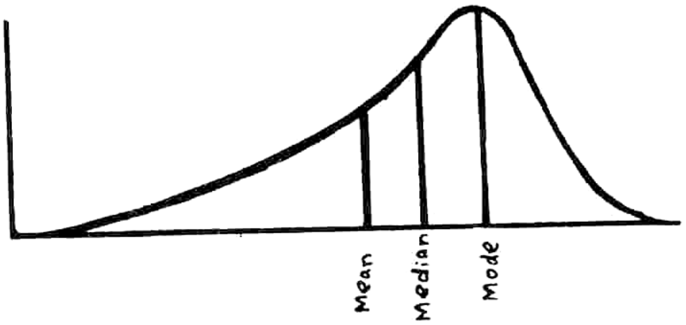
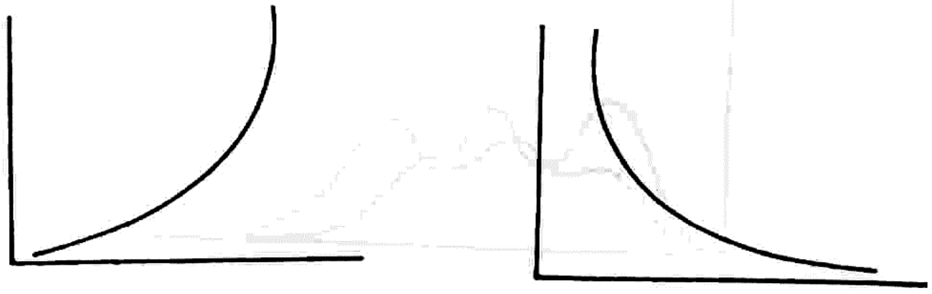


Fig. 4.5 Negative Skewed Distribution

Since the long tail of the distribution occurs among the low values of the variables, the bulk of the distribution shows relatively high values although there are few quite low values.

In a negatively skewed distribution, the mean is of the lowest value followed by the median and the mode has the highest value.

(c) In a **J-shaped or reverse J-shaped** curve, curves are extremely skewed in one direction with a maximum occurring at one end.



(d) **U-Shaped frequency Curve**

These frequency curves have their maximum (highest frequency) at both ends.

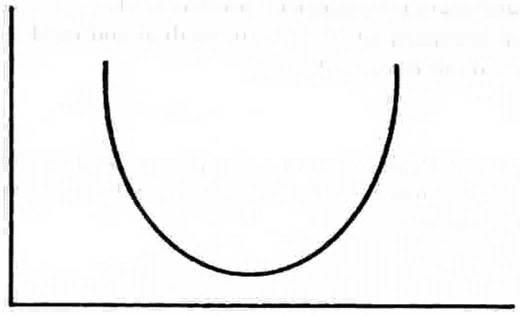


Fig. 4.7. U-Shaped

(e) A **bimodal frequency** curve has two equal maximum points.

Fig. 4.8a shows a bimodal frequency curve.

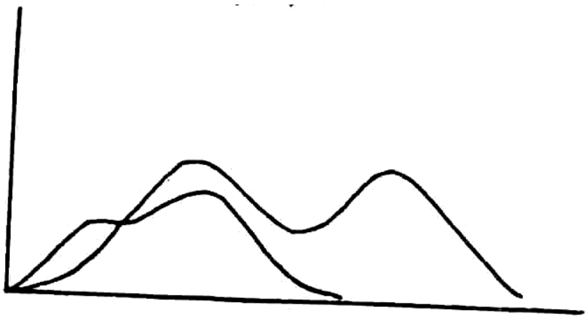


Fig. 4.8a . Bimodal Distribution

(f) A multimodal frequency curve has more than two maxima

Fig. 4.8b below

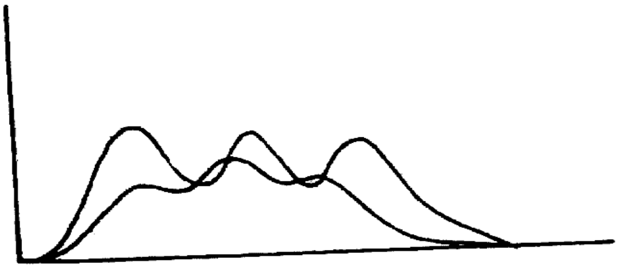
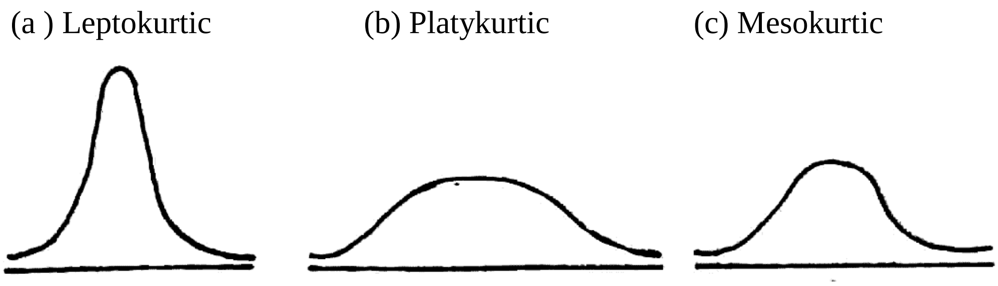


Fig. 4.8b A Multimodal Distribution

C. Kurtosis

Kurtosis is the degree of peakedness of a distribution. A distribution having a relatively high peak such as in (a) below is called leptokurtic, while the curve of (b) which is flat- topped is called platykurtic.



(a) Leptokurtic (b) Platykurtic (c) Mesokurtic

Fig. 4.7 Graphs of three types of kurtosis.

The normal distribution which is not very peaked or very flat-toped is called mesokurtic. The sketch (in Fig. 9.9c) above is a mesokurtic curve.

ASSIGNMENT

Consider the values of the statistics in the table below.

	Mean	Median	Mode
i.	80	75	70
ii.	74	74	74
iii.	60	72	80

In each case, how would you describe the distribution of the observations from which the statistics are obtained?

Use the terms: symmetric, positively skewed, negatively skewed etc).

REFERENCES

Avy, Donal et al (1979): *Introduction to Research in Education* U. S. A. Holt, Rineheat and Winston, Inc.

Best, J. W and Kahn, J. V. (1986): *Research in Education*. London: Practice Hall Inter.

Boyinbode I. R. (`1984*): Fundamental Statistical Methods in Education and Research*, Ile-Ife, DC SS Books.

Gay L. G. (1970) - *Education Research: Competencies for Analysis and Application*. Ohio. Charles E. Merill.

Guilford, J. P. and Fruchter, B. (1973): *Fundamental Statistics in Psychology and Education*.

McCall, R. B (1980): *Fundamental Statistics for Psychology*, U. A. A. Harcourt B. Jovanovich Inc.

UNIT 5: MEASURES OF RELATIONSHIP-CORRELATION AND REGRESSION

INTRODUCTION:

Correlation and Regression imply relationships between two or among more than two attributes'/traits values as possessed or exhibited by individuals that make up a set, sample or population. Attributes' or "trait" values may be quantitative or qualitative. Quantitative attribute values may be dichotomized or polychotomized. That is to say that they are broken into categories. These chotomizations or categorizations may be natural or forced (artificial). Whichever are the natures of the variables (attributes, traits or characteristics) concerned, correlation and/or regression are/is undertaken for a number of reasons. When one, two or more variables is/are very difficult to measure, then efforts and time are devoted to measuring it/then once together with measuring the other one or more variables that are supposed (as believed by researchers or as experts are convinced) to be closely related/associated with the variable(s) difficult to measure. Using the obtained values, a correlation or regression coefficient is established. Employing the coefficient of correlation or regression the values of the variable(s) very difficult to measure are estimated from obtained/given values of the one or more than one variable(s) that is/are easy to measure. Correlation or regression coefficient between two variables informs the researcher or student the extent of closeness or moving together of the values of the two variables. When the values of a variable are known with respect to units of time or events of the past, correlation or regression coefficients become handy tools for predicting the values of the variable in the future.

OBJECTIVES

There are many types of correlation and regression. In this unit, you will learn about; the Pearson *r*, the spearman *rho,* and the contingency coefficient. So by the end of this unit you should be able to:

i. state some common types of correlation or regression;

ii. illustrate and give examples of the five broad categories of linear correlation or regression;

iii. compute the Pearson product-moment correlation coefficient;

iv. compute the Spearman - Brown rank and other correlation coefficients;

v. compute the contingency coefficient;

vi. compute simple regression coefficient constant;

vii. establish the regression equation; and

viii. use the regression equation.

TYPES OF CORRELATION AND REGRESSION

Correlation is generally classified as;

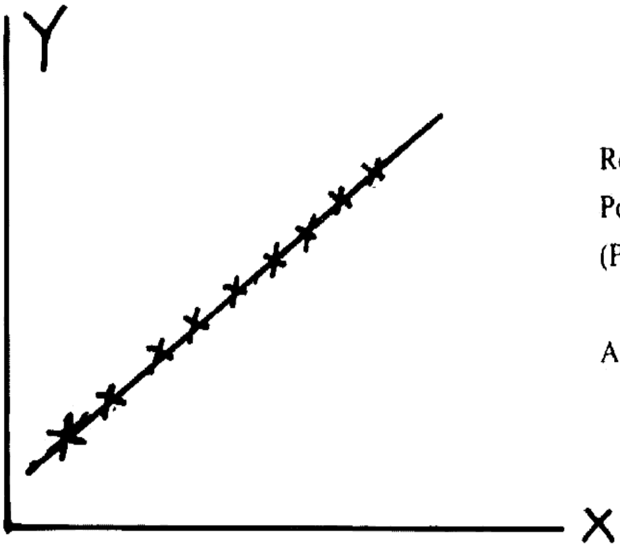
i. simple or

ii. multiple.

Correlation is **simple** if it concerns ***relationship between two variables*** while it is **multiple** if it concerns ***the relationship among three or more variables***.

Regression is generally classified into two broad types: linear and curvilinear regression. Linear regression is either ***simple linear regression*** or ***multiple linear regression***.. Curvilinear regression could be ***quadratic***, ***cubic***, ***exponential***, ***logrithmic***, etc. In this unit you are going to learn about simple correlation and simple linear regression.

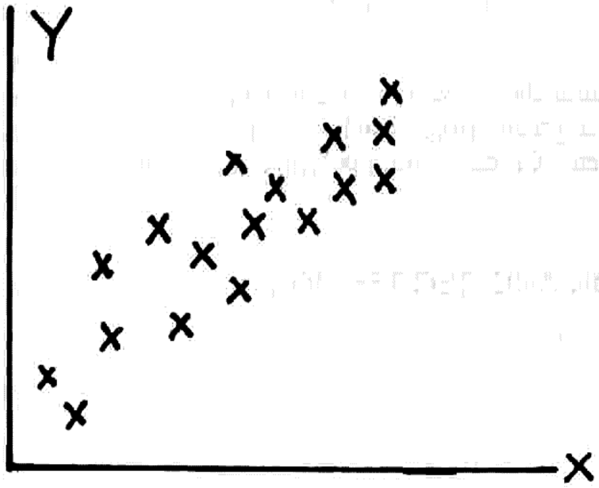
Generally, simple correlation and simple linear regression coefficient may be subdivided into five general types as regards the value and sign (or direction) of such coefficients. The five subdivisions/types may be sketched as shown below:



Relationship is
Positive/Direct and Perfect
(Perfect estimation/prediction)
$r = +1.00$
All plots are on the same line.

Fig. 5.1a

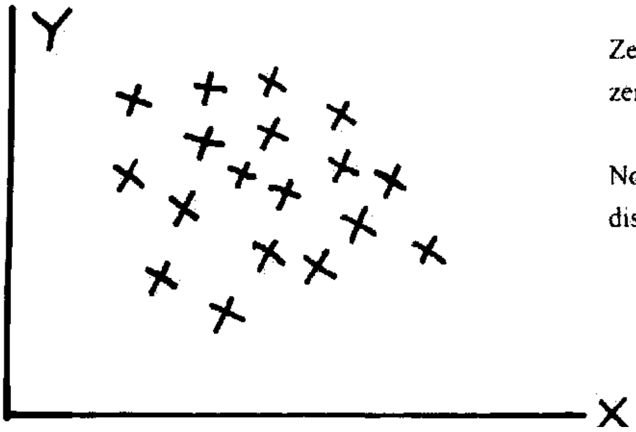
Fig. 3.2.1a



Relationship is Positive/
Direct but non perfect
$0 < r < +1.00$
All plots are not on the
same line but a discernable
trend

Fig. 5.1b

Fig. 3.2.1b

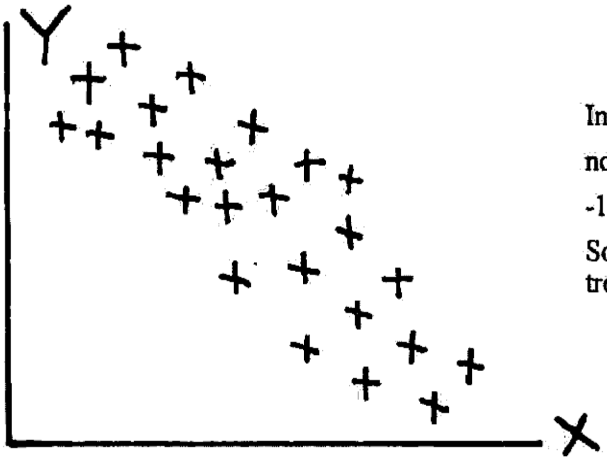


Zero or approximately zero relationship.

$r = 0.00$

No trend is readily discernable.

Fig. 5.1c

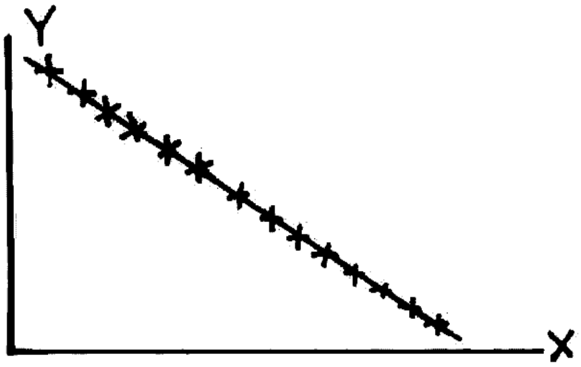


Indirect/Negative and non-perfect relationship

$-1.00 < r < 0.00$

Some left to right downward trend is discernable.

Fig. 5.1d



Relationship is Indirect/Negative but Perfect

$r = -1.00$

Fig. 5.1e

Fig. 3.2.1e

The above Figures 5.1 a - e are referred to as scattergrams. In Fig. 5.1a and 5.1e mere looking and tell you the type of relationship but mere gaze cannot ascertain the strength or index of relationship between X and Y values especially in Figs. 5.1b to 5.1d. If you also call a number of persons to fit in line that represent the plots in each of Figs. 5.1b to 5.1d, the persons will probably fit in different lines in each case. So mathematicians/statisticians have devised methods of ascertaining the strength/index of relationship between X and Y or fitting the line that is the best fit when a set of values of X and Y are given. The methods devised among others are (i) the Pearson *r*, (ii) the Spearman *rho* and (iii) the contingency coefficient and its associates.

Pearson *r* is employed when the distribution is bivariate, continuous and normal (or approximately so).The Spearman *rho* is employed when the distribution is bivariate, continuous and normal. However the scores of the individuals concerned in each variable are ranked in order of magnitude. The resulting ranks are used.

The contingency coefficient and its associates are employed when the data are frequency counts of individuals that belong to cross or joint categories of two contingency. Often the data are presented in contingency tables.

CORRELATION COEFFICIENTS - COMPUTATION.

A. **Pearson product moment correlation coefficient (Pearson *r*).**

Below are scores in a twenty itemed multiple choice test in each of a unit in Maths. and Chem. by l2 students (Scores are denoted ***X*** and ***Y*** respectively).

S/No.	1	2	3	4	5	6	7	8	9	10	11	12
X	15	9	12	13	6	10	5	11	8	7	14	10
Y	18	10	16	10	8	15	12	13	10	6	12	14

Calculate the Pearson ***r*** for the above.

$$\left.\begin{array}{l} x = X - X \\ y = Y - X \end{array}\right\} = \text{ deviation from the mean of X and Y scores.}$$

$$= \quad X = \frac{\sum X}{\pi}$$

$$= \quad \frac{120}{12}$$

$$= \quad 10$$

$$= \quad Y = \frac{\sum Y}{\pi}$$

$$= \quad \frac{144}{12}$$

$$= \quad 12$$

X	Y	x	y	xy	x^2	y^2
15	18	5	6	30	25	36
9	10	-1	-2	2	1	4
12	16	2	4	8	4	16
13	10	3	-2	-6	9	4
6	8	-4	-4	16	16	16
10	15	0	3	0	0	9
5	12	-5	0	0	25	0
11	13	1	1	1	1	1
8	10	-2	-2	4	4	4
7	6	-3	-6	18	9	36
14	12	4	0	0	16	0
10	14	0	2	0	0	4
120	144	00	00	73	110	130

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{73}{\sqrt{(110)\ (130)}}$$

$$\approx \frac{73}{119.58} \qquad \approx 0.61$$

In the above case the means of *X* and *Y* are whole numbers. Supposing the means of ***X*** and ***Y*** involved fractions or decimals then x and y will involve fractions. To remove the rigour involved in multiplying decimals by decimals, and premature approximations, the raw score or machine formula approach is adopted.

X	**Y**	**XY**	**X^2**	**Y^2**
15	18	270	225	324
9	10	90	81	100
12	16	192	144	256
13	10	130	169	100
6	8	48	36	64
10	15	150	100	225
5	12	60	25	144
11	13	143	121	169
8	10	80	64	100
7	6	42	49	36
14	12	168	196	144
10	14	140	100	196
120	144	1513	1310	1858

$$r = \frac{n\Sigma XY - \Sigma X \ \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2]\,[n\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$r = \frac{12(1513)\ 120(144)}{\sqrt{[12(1310) - (120)^2]\,[12(1858) - (144)^2]}}$$

$$r = \frac{18156 - 17280}{\sqrt{(15720 - 14400)\ (22296 - 20736)}}$$

$$r = \frac{876}{\sqrt{(1320)\ (1560))}}$$

$$r \approx \frac{876}{1434.99}$$

$$\approx \underline{0.61}$$

B. **Spearman - Brown Rank Order Correlation Coefficient (Spearman *rho*).**

The spearman *rho* is an approximation to or estimate of Pearson r. The computation of the coefficient is based on ranks of the scores used for the Pearson ***r.***

R_x denotes ranks of scores ***X*** R_Y denotes ranks of scores ***Y.*** Score 10 occurs twice in ***X*** column occupying the 6th and 7th positions. These positions are added and divided by 2 to give 6.5,

X	Y	R_x	R_Y	D	D^2
15	18	1	1	0.0	0.00
09	10	8	9	-1.0	1.00
12	16	4	2	2.0	4.00
13	10	3	9	-6.0	36.00
06	08	11	11	0.0	0.00
10	15	6.5	3	3.5	12.25
05	12	12	6.5	5.5	30.25
11	13	5	5	0.0	0.00
08	10	9	9	0.0	0.00
07	06	10	12	2.0	4.00
14	12	2	6.5	-4.5	20.25
10	14	6.5	4	2.5	6.25
				0.0	114.00

12s in ***Y*** are treated as 10 in ***X*** but because there are three tens (10s) in ***Y*** for the positions 8, 9 and 10 are shared for the three $\left[\frac{(8 + 9 + 10)}{3}\right]$ = 9 to give 9.

$$\rho = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6(114)}{12(144-1)} \quad = 1 - \frac{684}{12(143)}$$

$$\rho = 1 - \frac{684}{1716} \qquad = 1 - 0.3986$$

$$\rho = \underline{0.60}$$

ACTIVITY

S/No.	1	2	3	4	5	6	7	8	9	10
X	5	8	6	10	7	11	9	9	12	8
Y	10	8	9	7	9	6	8	8	6	9

1. Calculate the Pearson *r* between X and Y above using the **deviations from the means approach**.
2. Calculate through the **raw score approach** the ***P,*** Pearson ***r*** between X and Y given above.
3. Calculate the Spearman *rho* based on the data given above.

CONTINGENCY COEFFICIENT AND ITS ASSOCIATES

Below is a contingency table showing the frequencies of male and female adults who are of above average, average and below average height in a random sample of 250 adult workers in a university.

GENDER	HEIGHT			TOTAL
	Above Av.	About Av.	Below Av.	
Male	40	50	10	100
Female	30	60	60	150
TOTAL	**70**	**110**	**70**	**250**

What is the extent of relationship between gender and categories of heights of adults? We need to compute a contingency coefficient. We first calculate a chi-square statistics. We compute and draw up an expected contingency table from the totals of the observed contingency.

$$\text{Cell ij} = \frac{ith\ \ Raw\ Total\ \times\ jth\ Column\ Total}{Grand\ Total}$$

where *i* is the i^{th} raw and

j is the j^{th} column

GENDER	**HEIGHT**			**TOTAL**
	Above Av.	**About Av.**	**Below Av.**	
Male	28	44	28	100
Female	42	66	42	150
TOTAL	**70**	**110**	**70**	**250**

Each cell is filled by computing

$$\text{Cell } ij \quad = \quad \frac{ithRow\ Total\ X\ \ jth\ column}{Grandtotal}$$

$$\frac{100\ X\ 70}{250} = 28 \qquad \frac{\overset{3}{\cancel{160}} \times \overset{14}{\cancel{71}}}{\underset{\cancel{5}\ 1}{\cancel{250}}} \quad = 42 \text{ or } 70 - 28 = 42.$$

$$\frac{100\ X\ 110}{250} = 44$$

140 - 44 = 66

70 - 28 = 42

$$\chi^2 = \sum \frac{(0-E)^2}{E} \quad \text{where } x^2 = \text{chi square}$$

$\sum$ = summation

O = observed frequency

E = expected frequency

However, a table may help you do this faster.

0bserved	E xpected	0 - E	(O - E)^2	$\frac{(0-E)^2}{E} = x^2$
40	28	12	144	5.14
50	44	6	36	0.82
10	28	-18	324	11.57
30	42	-12	144	3.43
60	66	- 6	36	0.55
60	42	18	324	7.71
250	250	00	-	29.22

Contingency coefficient, C is given by

$$C \;=\; \sqrt{\frac{\chi^2}{N+\chi^2}}$$

Where C = contingency coefficient

X^2 = chi square

N = Grand total of subjects or cases.

Grand total, N in this case is 250.

x^2 = 29.22

N = 250

$$\therefore \quad C \;=\; \frac{\sqrt{29.22}}{250\ \ 29.22} \;=\; \frac{\sqrt{29.22}}{279.22} \;=\; \sqrt{0.1046}$$

$$C \;=\; 0.32$$

Correlation of Attributes

The presentation in a contingency table often concern attributes of human beings or objects. The degree to which one of the attributes depend upon, is associated with or related to the other attribute is referred to as correlation of attributes. In a *k x k* contingency, the correlation of attributes, *r* is given as:

$$r \;=\; \sqrt{\frac{\chi^2}{N\,(k-1)}}$$

For a 2 × 2 table, the correlation attribute is called **tetrachoric**.

Supposing the table below concerns the frequencies of height categories of people of various body forms.

BODY FORM	**HEIGHT CATEGORY**			**TOTAL**
	Tall	**Aver. H.**	**Short**	
Ectomorph	60	30	10	100
Mesomorph	50	200	50	300
Endomorph	20	30	50	100
TOTAL	**130**	**260**	**110**	**500**

Step 1 = Calculate the expected Contingency Table

Compute this, using the formula: $\frac{ith\ \ Raw\,Total\ \times\ jth\ column\ Total}{Grand\,Total}$

BODY FORM	HEIGHT CATEGORY			TOTAL
	Tall	**Aver. H.**	**Short**	
Ectomorph	a 26	b 52	c22	100
Mesomorph	d 78	e 156	f 66	300
Endomorph	g26	h 52	i22	100
TOTAL	**130**	**260**	**110**	**500**

Step II Compute the Chi square using the formula

$$\chi^2 = \sum \frac{(0-E)^2}{E}$$

You may now prepare the kind of table for computing χ^2 as done before.

$$a = \frac{100x130}{500} = 26$$

g = 26 also

$$d = \frac{300\ x\ 130}{500} = 78$$

$$b = h = \frac{100\ x\ 260}{500} = 52$$

$$e = \frac{300\ x\ 260}{500} = 156$$

0	E	0 - E	(0 - E)^2	$\frac{(0-E)^2}{E}$
60	26	34	1156	44.46
30	52	-22	484	9.31
10	22	-12	144	6.55
50	78	-28	784	10.05
200	156	44	1636	12.41
50	66	-16	256	3.88
20	26	- 6	36	1.38
30	52	-22	484	9.31
50	22	28	784	35.64
500	500	000	-	132.99

$\chi^2{}_{\text{call}} = 132.99$

$$r = \sqrt{\frac{132.99}{500(3-1)}}$$

$$r = \sqrt{\frac{132.99}{1000.00}}$$

$r = \sqrt{0.132.99}$

$r = \underline{0.36}$

ACTIVITY

Calculate r for the following

GENDER	INFECTION		TOTAL
	YES	NO	
Male	70	30	100
Female	40	110	150
TOTAL	110	140	250

REGRESSION EQUATIONS

Earlier in this unit you learnt that there is a line of best fit that represents any scatter diagram that shows a linear trend. Such a line must have the general form $y = a_1 x + a_o$.

In this equation $y = a_1 x + a_o$.

a_1 is the coefficient of regression and a_o is a constant.

If the regression equation is the line of best fit then

$$a_1 = \frac{\Sigma\, xy}{\Sigma\, x^2}\text{ ; where } x = X - \overline{X} = X - \overline{X}$$

$$y = Y - \overline{Y}$$

OR

$$a_1 = \frac{n\Sigma\, XY - \Sigma X\ \Sigma Y}{n\Sigma X^2 - (\Sigma\, X)^2}$$

$$a_o = Y - a_1\, X$$

ASSIGNMENTS

1. Compute the Pearson product moment correlation coefficient between ***X*** and ***Y*** as given below.

X	2	5	8	11	14	17	4	8	8	10	12
Y	1	7	13	20	24	30	6	14	13	17	20

2. Calculate the correlation of attributes as regards the data below.

ATTRIBUTE II	ATTRIBUTE I CATEGORIES			TOTAL
CATEGORIES	a	b	c	
k	70	130	200	400
l	50	150	100	300
m	180	90	30	300
TOTAL	**300**	**370**	**330**	**1000**

REFERENCES

Avy, Donal et al (1979): *Introduction to Research in Education* U. S. A. Holt, Rineheat and Winston, Inc.

Best, J. W and Kahn, J. V. (1986): *Research in Education*. London: Practice Hall Inter.

Boyinbode I. R. (`1984*): Fundamental Statistical Methods in Education and Research*, Ile-Ife, DC SS Books.

Gay L. G. (1970) - *Education Research: Competencies for Analysis and Application*. Ohio. Charles E. Merill.

Guilford, J. P. and Fruchter, B. (1973): *Fundamental Statistics in Psychology and Education*.

McCall, R. B (1980): *Fundamental Statistics for Psychology*, U. A. A. Harcourt B. Jovanovich Inc.

UNIT SIX PROBABILITY AND ITS LAWS

INTRODUCTION

In common language, the term probability must have been coined from the word probable. When applied to the occurrence of events or outcomes of experiments, it implies the likelihood, the chance or relative odds for an event to occur. In the general sense, if something is probable, it implies that the something may or may not happen.

Probability as an aspect of Applied Mathematics or a branch of statistics, may be referred to ***as the science of quantifying or gauging the odds for or chance of an event occurring***. The importance of the study of probability lies in the area of mathematical expectation and the estimation of risks. In testing null hypothesis in educational research, some risks are involved. It is therefore important that you know a little about probability theory.

OBJECTIVES

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

1. define the probability of an event;
2. calculate the probability of (simple) events when necessary and sufficient data are given;
3. state and explain (illustrate or show) the simple laws of probability; and
4. apply the laws of probability in computing or calculating the probability of events when necessary data are given.

MEANING OR DEFINITION OF THE PROBABILITY OF AN EVENT, E DENOTED by Pr {e}

If in an experiment, there are *n* possible outcomes that are equally likely and the event E can occur *k* times or in *k* ways, then the probability of E, Pr {E} = p = $\frac{k}{n}$ .. The probability that event E does not occur is

Pr {not E} = q = $\frac{n-k}{n}$

Thus, in the cast of a single die, what is the probability of getting a score greater than four (or in other words 5 or 6)? Supposing the outcomes 1,2,3,4,5 and 6 are equally likely (i.e. the die is fair), then n, the number of all possible outcomes that are equally likely is six. The event, getting 5 and 6 can occur in **two ways**. Therefore, the probability of getting 5 or 6 in a cast of a die is **two/six** Pr (E) = 2/6 = 1/3.

Note that:

(i) $P = k/n \Rightarrow k \leq n$

(ii) $q = \frac{n-k}{n} = q = 1 - p \Rightarrow p + q = 1$

(iii) $o \leq p \leq 1$

When an event is sure to occur, the probability is 1 and when an event is sure not to occur/happen the probability is O.

The above definition of probability is classical.

Relative Frequency Definition

The probability of an event is the relative frequency of occurrence of the event when the number of (trials of the experiment) observations is very large. The probability of an event in this sense is the limit of the relative frequency of the occurrence of the event as the number of (trials of the experiment) observations increase indefinitely.

If in 2500 tosses of the Nigerian one naira coin (obtained by 50 students each tossing the coin independently 50 times, 1261 heads were observed, then the relative frequency of the occurrence of heads Pr(H) is $\frac{1261}{2500} = 0.5044 = 0.50$ to two decimal places. In the same way as above, if 2500 casts of dice were made and the table below shows the observations made/ obtained.

Score	1	2	3	4	5	6	Total
Freq.	407	420	425	413	430	405	2,500
Rel. freq		0.163	0.168	0.17	0.165	0.172	0.162 1.000

It is believed that the more the number of observations, the more the value obtained approached the limiting value of 0.50 for the probability of Head and 0.16 for the probability of each of the scores 1,2,3,4,5 and 6.

Calculation of Probabilities of Events

A. **Simple events**

1. What is the probability of the head turning up when a fair coin is tossed?

   Solution: Possible outcomes are H and T

   Relative frequency of H = ½

   $\therefore \; Pr(H) = \frac{1}{2}$

2. What is the probability of obtaining a multiple of 3 from the cast of die? Possible outcomes

   1,2,3,4,5,6,. Multiples of 3 are 3 and 6. They are two number out of six.

   ∴ Probability required is 2/6 = 1/3.

3. In a bag, there are ten identical balls in terms of size; five are black, three are white and two are yellow. What is the probability that a ball drawn at random is (i) white and (ii) yellow?

   (i) Pr(white ball) = 3/10 = 0.3

   (ii) Pr(yellow ball) = 2/10 = 0.2.

B. **Compound Events**

1. What is the probability of:

   (i) obtaining two heads

   (ii) at least two heads turning up and

   (iii) at most a head turns up;

   (iv) in three tosses of a fair coin?

Solution

Generate all possible outcomes

(HHH, (HHT), (HTH), (THH), (THT), (THT), (TTH), (TTT)

Total number of possible outcomes is 8.

(i) Outcomes in which two heads are obtained are three in number. Pr(two heads in three tosses) = 3/8 = 0.375

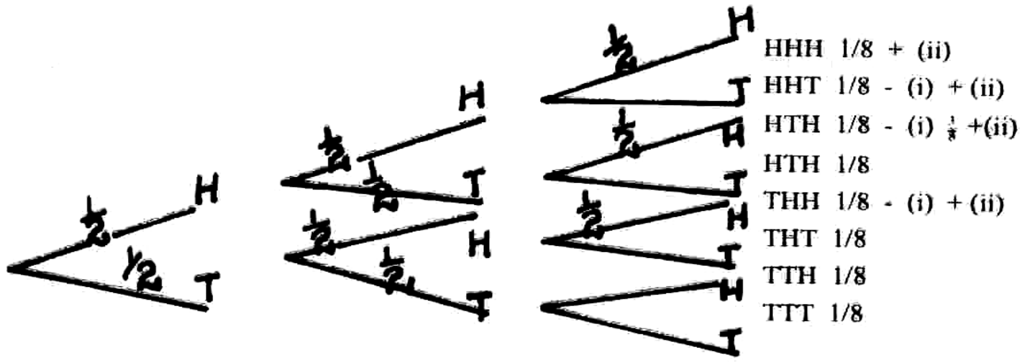
(ii) At least two heads implies two heads or more i.e. two heads or three heads. Outcomes satisfying these are four Pr(at least two heads in three tosses) = 4/8 = 0.5

(iii) At most a head means one head or no head.

Outcomes that are included are also four in number

Pr(at most one head) = 4/8 = 0.5.

A stokastic diagram or tree diagram maybe used as shown in the following diagram



2. When a die is cast twice or when two dice are cast once, what is the probability that the sum of scores is:

(i) a multiple of four;

(ii) at most 6 and

(iii) at least 10.

Solution

Generate the possible outcomes of casting a die twice or casting two dice one; a total of 36 outcomes.

	Second die/cast scores						
1st die cast scores	1	2	3	4	5	6	
1	1,1	1,2	1,3	1,4	1,5	1,6	7
2	2,1	2,2	2,3,	2,4,	2,5,	2,6	8
3	3,1,	3,2,	3,3	3,4	3,5	3,6	9
4	4,1	4,2	4,3	4,4	4,5	4,6	10
5	5,1	5,2	5,3	5,4	5,5	5,6	11
6	6,1	6,2	6,3	6,4	6,5	6,6	12

Note: For the sums, check the upward left to right diagonals. Then check downward left to the right diagonals for the sums similar to the located sum

(i) a multiple of four within range of the sum of scores is 4, or 8 or 12.

three outcomes sum up to 4

five outcomes sum up to 8

and one outcome sum up to 12.

In all nine outcomes.

Probability of multiple of 4 = 9/36 = 0.25

(ii) At most 6 implies a sum of:

6, 5, 4, 3, or 2. Five, four, three, two outcomes and one outcome results in a sum of 6, 5, 4, 3 and 2 respectively. So Pr {at most 6} = 5+4+3+2+1)/36

= 15/36 $\approx$ 0.417.

(iii) at least 10 implies 10, 11 or 12.

(iv) Outcomes that sum up to 10 are three

Outcomes that sum up to 11 are two

and outcome that sum up to 12 is one.

Total number of outcomes is six.

Pr(at least 10) = 6/36 = 1/6 = 0.16.

LAWS OF PROBABILITY AND APPLICATIONS

If two events E_1 and E_2 are **Mutually Exclusive**,

then

(i) Pr {E_1 E_2}= 0 where Pr {E_1 E_2} is the probability of E1, and E2 occurring;

(ii) Pr {E1+E2}=

Pr {E1} + Pr {E2} where Pr {E1+ E2} is the probability of {E1 ,} or {E2 } occurring.

If two dice are cast once, what is the probability that the sum of scores is

(i) 5 and 6 and

(ii) 5 or 6?

Solution

(i) Sum being 5 is mutually exclusive of sum being 6.

Pr {5 and 6} = 0

(ii) Sum can be 5 or 6 so Pr {5 or 6} = 9/36 i.e.

Pr {5} + Pr {6} = 4/36 + 5/36 = 9/36 = 0.25

By extension if E_1, E_2 ...., E_n are n mutually exclusive events: then Pr {E1 E2 ... En} = 0 while, Pr {E1 + E2 + ... + En} = Pr {E1} + Pr {E2} +... + Pr {En}.

Independent Events

If E_1 and E_2 are events (that are not necessarily mutually

exclusive) then: Pr $\{E_1 E_2\}$ = Pr $\{E_1\}$Pr$\{E_2\}$ provided E_1 and E_2 are independent.

If E_1, E_2 . . ., E_n are independent events by extension.

Pr$\{E_1 E_2 \ldots E_n\}$ = Pr $\{E_1\}$ Pr E_2 . . . Pr$\{E_n\}$

Note when you randomly draw for instance balls from a bag (i) with replacement, then probability of previous draw(s) does/do not affect the probability of successive draw(s), therefore successive draws are independent of previous draws; while (ii) without replacement successive draw/s is/are dependent upon previous draw/s

Examples

(a) In a bag, there are 6 yellow, 5 white and 4 black balls of identical size. If you draw three balls one at a time with replacement, what is the probability of (i) getting one ball of each colour (ii) two yellow and one black being drawn?

Solution

(i) Pr Y,W,B = Pr Y Pr W Pr B

= (6/15) (5/15)(4/15)

= 8/225 ≈ 0.036.

(ii) Pr YYB= Pr Y Pr Y Pr W

= (6/15)(6/15)(4/15)

= 16/375

≈ 0.043

(b) Given that a coin and a die are thrown up, what is the probability that a tail and one turned up?

Solution

Obtaining a tail is independent of obtaining the score of one, so Pr{T1} = Pr {T} Pr {1}

= $(\frac{1}{2})$ (1/6)

= 1/12

≈ $\underline{\underline{0.08}}$

Conditional Probability

Given that E_1, and E_2 are events that are not mutually exclusive, then the probability that E_2 occurs when E_1 has occurred is call the conditional probability of E_2 given E_1. This conditional probability is denoted

Pr $\{E_2/E_1\}$ or Pr$\{E_2$ given $E_1\}$

Pr $\{E_2/E_1\}$= $\frac{\Pr E_1\ E_2}{\Pr E_1}$ ``------------ Division Law"

If E_1 and E_2 are dependent events then Pr $\{E_1\ E_2\}$

= Pr $\{E_1\}$ Pr $\{E_2/E_1\}$

Illustration

Supposing μ = {1,2,3,4,5,6,7,8,9,10}

E_1 = event that the number drawn randomly from μ is even;

E_2 = event that the number randomly drawn from μ is odd;

and E_3 event that the number randomly drawn from μ is prime.

(i) Find Pr $\{E_3/E_2\}$ and

(ii) Find Pr$\{E_3/E_1\}$.

Solution

(i) Pr E_1 = 5/10 = $\frac{1}{2}$

Pr $\{E_2\}$ = 5/10 = $\frac{1}{2}$

Pr $\{E_3\}$ = 4/10 = 2/5 provided $1 \neq E_3$.

Pr $\{E_3/E_2\}$ = $\frac{\Pr E_2\,E_3}{\Pr E_2} = \frac{3/10}{\frac{1}{2}}$

= 3/5

Note:

$\{E_2\ E_3\}$ = E_2 and E_3 = $E_2 \cap E_3$ = {3,5,7}

So Pr $\{E_1\ E\}$ = 3/10

(ii) Pr $\{E_3/E_1\}$ = $\frac{\Pr E_1\ E_3}{\Pr E_1}$

$\{E_1\ E_3\}$ = E_1 and E_3 =$E_1 \cap E_3$ = {2}.

pr (E3/E1 = $\frac{\Pr E_1\ E_3}{\Pr E_2} = \frac{1/10}{\frac{1}{2}} = 1/5.$

ASSIGNMENT

The faces of an octahedron are marked 1, 2 , 3 through to 8. If the octahedron is fair and is cast once, what is the probability of: (i) at least 6, (ii) at most 4 and (iii) a multiple of three, turning up?

REFERENCES

Avy, Donal et al (1979): *Introduction to Research in Education* U. S. A. Holt, Rineheat and Winston, Inc.

Best, J. W and Kahn, J. V. (1986): *Research in Education*. London: Practice Hall Inter.

Boyinbode I. R. (`1984*): Fundamental Statistical Methods in Education and Research*, Ile-Ife, DC SS Books.

Gay L. G. (1970) - *Education Research: Competencies for Analysis and Application*. Ohio. Charles E. Merill.

Guilford, J. P. and Fruchter, B. (1973): *Fundamental Statistics in Psychology and Education*.

McCall, R. B (1980): *Fundamental Statistics for Psychology*, U. A. A. Harcourt B. Jovanovich Inc.

UNIT SEVEN: DISTRIBUTION FUNCTIONS OF A RANDOM VARIABLE

INTRODUCTION

The importance of the use of functions as short hand notations, models for ascertaining, specifying or estimating values, and prediction models can never be over valued or appreciated. When we write that $y = f(x) = 2x - 1$ (or in short $y = 2x - 1$) we save a whole lot of time and space which will be required to list, though unsuccessfully, all the pairs, (x,*y*), which $y = 2x - 1$ stand for. A function as a mathematical model may be established through a large number of observed values or measurements. Once established, the model/function enables us to find quickly the missing values within the set of pairs or triples of values. If for instance $y = 2x - 1$ then we can quickly complete the following (x,y) = (3,-?) or (6,-? Or (-?, 9).

OBJECTIVES

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

(i) explain what is meant by a distribution function of random variables;

(ii) list and explain some types of distribution functions;

(iii) apply some of the distribution functions in solving some related and appropriate problems/exercises.

(iv) define (or explain the meaning of) distributions

(v) define or state specific discrete distributions

(vi) generate or compute the values of specific discrete distributions

(vi) explain the meaning of continuous distributions

(vii) define or state specific continuous distributions and

(viii) give examples of continuous distributions including some features.

MEANING OF DISTRIBUTION FUNCTION OF A RANDOM VARIABLE

A variable is said to be random if the occurrences of the values of the variable are haphazard, do not follow any unique pattern or are not certainly predictable. A distribution function is a function or more so a mathematical rule or model that assigns each element of a distribution to a unique element of a set of values.

A distribution function of a random variable is a mathematical rule/model that assigns each element of a distribution (which is an orderly arranged possible values of a random variable) to a unique set.

TYPES OF DISTRIBUTION FUNCTIONS

(a) **Probability functions**

(i) If x is a random variable then the function probability of x_i denoted $P\ (x_i)$ maps x_i to its probability (or relative frequency of x_i)

Example:

Let x be the sum of the values on the faces of a die that turn up in the two casts of the die.

x	2	3	4	5	6	7	8	9	10	11	12
P(x)	$^1/_{36}$	$^1/_{18}$	$^1/_{12}$	$^1/_{9}$	$^1/_{36}$	$^1/_{6}$	$^1/_{36}$	$^1/_{9}$	$^1/_{12}$	$^1/_{18}$	$^1/_{36}$

Note: if $x_i = x_1,\ x_2 \ldots .x_n$

Then $\sum P\ (x_i) = 1$.

(b) **Factorial Permutation and Combination Functions**

Let n stand for a number of objects and r the number of objects taken from n for serial arrangement ($r \leq n$). Then the number of ordered arrangements of the r objects collected from n objects is a function $f(n, r) = nPr$

$$^nP_r = {}^{n!}/_{(n-r)!}$$

$$^8P_3 = {}^{8!}/_{(8-3)!} = 8\times7\times6\times5\times4\times3\times2\times{}^1/_{5!}$$

$$= 8\times7\times6\times5\times{}^{5!}/_{5!}$$

$$^6P_4 = \frac{6!}{(6-4)!} = \frac{6!}{2!} = \frac{6\times5\times4\times3\times2\times1}{2\times1} = 360$$

(c) **Expectation Function**

This distribution function maps all the products of the elements of a distribution and their corresponding probabilities into a value.

Expectation of x denoted ($E(x)$ is given as

$$E(x) = \sum x_i\, P(x_i)$$

$$E(x) = x_i\ \ P(x_i) + x_2\ P(x_2) + \ldots . +\ x_n\ P(x_n)$$

Supposing x is the number of successes in hitting a target in 6 trials and the probability of success is 0.5.

Find the *E(x).*

X	0	1	2	3	4	5	6	
P(x)	0.016	0.094	0.234	0.313	0.234	0.094	0.016	0.99
xP(x)	0.00	0.094	0.468	0.939	0.936	0.470	0.096	3.0

So $E(x) \quad = \quad xP(x) \quad = \quad \underline{\underline{3.0}}$

Generally the expectation function maps the entire distribution on the mean of the distribution.

$$\sum xp(x) = \bar{x}$$

(d) Z- and T- as distribution functions.

(a) The standard score Z- is a function since it maps every member of a distribution *x* to a unique Z- score.

Thus $f(x) = Z = \dfrac{x-\bar{x}}{S}$

Given that $x = 53$ and $S = 12$ and $f(x) = Z = \dfrac{x-\bar{x}}{S}$

Find $f(83)$ and $f(35)$

$$f(83) = Z = \frac{83-53}{S} = \frac{30}{12} = 2.5$$

$$f(35) = Z = \frac{35-53}{12} = -18 = \underline{\underline{-1.5}}$$

(b) T-score is derived from the T- function

$$f(x) = \frac{10(x-\bar{x})}{S} + 50 = T(Z) = 10Z + 50$$

Beside the examples given above there are some other distribution functions. These include; the multinomial, the Poisson, and the normal distribution functions, etc. The Poisson and normal distribution functions will be treated in subsequent units.

SOME APPLICATIONS OF DISTRIBUTION FUNCTIONS

COMPUTING THE PROBABILITY OF SIMPLE OR COMPOUND EVENTS

Example

What is the probability of a marks-man hitting a target three times out of six trials given that the relative frequency of hitting the target is $^2/_5 = 0.4$? The exercise concerns Bernoulli processes thus the probability function *P(x)* is given as

$P(x) = {}^nC_xp^x\, q^{n-x}$ where $n = 6$; $x = 3$; $p = 0.4$; $q = 0.6$

$= {}^6C_3\, (0.4)^3\, (0.6)^3 = 20\, (0.064)\, (.6)^3$

$= 20\, (0.064)\, (0.216) = 0.27652$

What is the probability of getting at least 4 heads in 6 tosses of a fair coin?

$$P(x = 4, 5, 6) = \sum_{i=4}^{6} 6Cx_i\, (0.5)^6$$

$$= [{}^6C_4 + {}^6C_5 + {}^6C_6]\, (0.5)^6$$

$$= (15 + 6 + 1)\, (0.015625)$$

$$= (22)\, (0.015625)$$

$$= 0.34375$$

In terms of relative frequency P(x = 4, 5, 6)

$$= \frac{\sum_{i=4}^{6} 6Cx_i}{\sum_{i=0}^{6} 6Cx_i}$$

$$= \frac{15+6+1}{26}$$

$$= \frac{22}{64} = \frac{11}{32}$$

$$= \underline{\underline{0.34375}}$$

DISTRIBUTION FUNCTIONS INVOLVING COMBINATORIAL ANALYSIS/COMPUTATIONS

Example

In how many ways can a committee of 5 members be formed from 9 persons? This question involves the combination function.

$f(x) = nCx$ Here $n = 9$, $x = 5$.

$$f(5) = 9C5 = \frac{9!}{(9-5)!5!}$$

$$= \frac{9\times 8\times 7\times 6\times 5!}{4!\times 5!}$$

$$= \frac{9\times8\times7\times6}{4\times3\times2\times1} = 3\times7\times6 = \underline{\underline{126}}$$

EXPECTATION FUNCTION

What is the average score of the sum of values obtainable from casting an Octahedron twice if the faces bear the values 1, 2, 3 … 8?

The above involves the expectation function *E(x)* where *E(x)* = $\sum xp(x)$.

X	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
F	1	2	3	4	5	6	7	8	7	6	5	4	3	2	1
p(x)	$\frac{1}{64}$	$\frac{2}{64}$	$\frac{3}{64}$	$\frac{4}{64}$	$\frac{5}{64}$	$\frac{6}{64}$	$\frac{7}{64}$	$\frac{8}{64}$	$\frac{7}{64}$	$\frac{6}{64}$	$\frac{5}{64}$	$\frac{4}{64}$	$\frac{3}{64}$	$\frac{2}{64}$	$\frac{1}{64}$
Xp(x)	$\frac{2}{64}$	$\frac{6}{64}$	$\frac{12}{64}$	$\frac{40}{64}$	$\frac{30}{64}$	$\frac{42}{64}$	$\frac{56}{64}$	$\frac{72}{64}$	$\frac{70}{64}$	$\frac{66}{64}$	$\frac{60}{64}$	$\frac{52}{64}$	$\frac{42}{64}$	$\frac{30}{64}$	$\frac{16}{64}$

$$\sum xp(x) = \frac{1}{64}(2 + 6 + 12 + 40 + 30 + 42 + 56 + 72 + 70 + 66 + 60 + 52 + 42 + 30 + 16)$$

$$= \frac{1}{64}(70+170+196+130) = \frac{1}{64}(576)$$

$$= \underline{\underline{9}}$$

ACTIVITY

1. How many four digit numbers can be formed from the figures 1, 2, 3, 4, 5, 6, 7, 8, 9? No figure is repeated in each number.
2. How many committees each made up of 3 men and 2 women can be composed from a group of 7 men and four women?

DISCRETE AND CONTINUOUS DISTRIBUTIONS

The way and manner in which many things are done in statistics depend on whether the data concerned are discrete or continuous. In data representation, discrete data are handled using pictograms, bar charts and line charts or their likes while continuous data are handled using histograms, frequency polygons and ogives or their associates. In the use of the measures of central tendency; for summarizing or precise description of set of data; as bench marks for intra and intergroup comparisons; and as estimators of population parameters; the mode and median and at best the mean may be employed for discrete distribution; while for continuous distributions the mode, the median and the mean are readily employed.

Therefore, the knowledge of what discrete and continuous distributions are and some of their illustrations become necessary.

DISCRETE DISTRIBUTIONS

A distribution consists of values of a variable and their corresponding frequencies as generated from a statistical experiment, a data collection activity or function. A discrete distribution therefore consists of values of a discrete variable and their corresponding frequencies. In other words, a discrete distribution consists of whole number values and their corresponding frequencies as generated from a statistical experiment, a data collection activity or a function.

Having known what: (i) distribution and (ii) discrete distribution means; you can now note some examples and illustrations of discrete distributions. The very first example is the binomial distribution concerning the number of successes (X) out of a number of trials (*n*) of Bernoulli process (experiment). You should recall that the distribution as defined here is a binomial distribution with mean *np* and variance *npq*.

ILLUSTRATION

Supposing in 7 tosses of a coin, you want X heads. What is the distribution of X? Recall the distribution is:

X	0	1	2	3	4	5	6	7
Freq.	$7C_0$	$7C_1$	$7C_2$	$7C_3$	$7C_4$	$7C_5$	$7C_6$	$7C_7$

In actual figures you have

X	0	1	2	3	4	5	6	7	$\overline{X} = 3.5$
Freq.	1	7	21	35	35	21	7	1	$S^2 = 1.75$

Continuous Distributions

A continuous distribution consists of a set of values of a continuous variable. A distribution is said to be continuous if the members of the distribution take on fractional and / or whole number values that are within a number line continue from "*a* to *b*".

Examples of Continuous distributions

1. The ages of human beings who belong to a given community at a given point in time.
2. The life spans of a batch of electric bulbs manufactured by a firm.
3. The weights of a random sample of yam tubers harvested from a farm.
4. Means of basic salaries of random samples of civil servants with replacement.
5. The length of diagonals of cover page of books, journals and magazines in a certain school library.

Indeed many more examples exist and you should define/state ten such distributions yourself.

Hypothetical illustrations

Age distribution in years of students in a school.

Group	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55
Freq.	10	30	20	20	15	10	5	5

Usually, a continuous distribution is described by stating its mean, variance/standard deviation and the nature of its curve. So you compute the mean and variance of the above distribution.

Middle point,	x	18	23	28	33	38	43	48	53
Freq.	f	10	30	20	20	15	10	5	5
	fx	180	690	560	660	570	430	240	265
	x^2	324	529	784	1089	1444	1849	2304	2809
	fx^2	3240	15870	15680	21780	21660	18490	11520	14045

$$\overline{X} = \frac{\sum fx}{\sum f} = \frac{3595}{115} = 31.26$$

$$S^2 = \frac{n\sum fx^2 - (\sum fx)^2}{n^2} = \frac{115(122285) - (3595)^2}{115^2}$$

$$= \frac{1138750}{13225} = 86.11$$

The distribution is positively skewed. So the hypothetical age distribution given in 1 above is said to be continuously distributed with a mean of 31.26 and a variance of 86.11.

ASSIGNMENT

Prepare a frequency table of the distribution *X* where *X* is the sum of the possible outcomes of casting a regular octabhedron twice. The faces of the octahedron are marked 1,2,3,4,5,6,7,8. Describe the distribution.

REFERENCES

Avy, Donal et al (1979): *Introduction to Research in Education* U. S. A. Holt, Rineheat and Winston, Inc.

Best, J. W and Kahn, J. V. (1986): *Research in Education*. London: Practice Hall Inter.

Boyinbode I. R. (`1984*): Fundamental Statistical Methods in Education and Research,* Ile-Ife, DC SS Books.

Gay L. G. (1970) - *Education Research: Competencies for Analysis and Application*. Ohio. Charles E. Merill.

Guilford, J. P. and Fruchter, B. (1973): *Fundamental Statistics in Psychology and Education*.

McCall, R. B (1980): *Fundamental Statistics for Psychology*, U. A. A. Harcourt B. Jovanovich Inc.