UNIT 9: PROCESSING OF DATA

UNIT STRUCTURE

9.1 Learning Objectives

9.2 Introduction

9.3 Data and Data Types

9.4 Data Measurement Techniques

9.5 Editing of data

9.6 Classification of data

9.7 Tabulation of data

9.8 Let us sum up

9.9 Answers to check your progress

9.10 Further Readings

9.11 Model Questions

9.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- define data and various types of data
- explain various measurement issues related to data and data types
- describe data editing
- classify the data
- tabulate the data in uni-variate and multi-variate distributions.

9.2 INTRODUCTION

The basic aim of any research is to make generalisations and build theory upon them. Note that all scientific methods of research require careful observations at some level or other. Speaking plainly, observation is the act of acquiring knowledge and facts through our five sense organs. That is why observation is also known as sensory perceptions. The act of gathering sensory perception is known as empirical research.

Clearly, therefore, in empirical research we try to observe the attributes and whatever is obtained as a result of observation is called data. We will take up the issue of data in some detail a little later. The data help us in validating our research hypothesis as to whether we can accept or reject it. The act of validation makes generalisations possible and when we validate our hypothesis it turns into a theory. The theory is particularly useful for explanation and prediction about the phenomenon or similar phenomena.

It is, therefore, seen that data plays an important role in empirical research. It is based on data that sometimes distinction between quantitative and qualitative research is also made. In brief quantitative data deals with quantities or numbers. While qualitative data deal with attributes. In this chapter we will discuss various aspects and issues related to data, particularly quantitative data.

9.3 DATA AND DATA TYPES

Data is the plural form of the term datum meaning given or taken for granted. This simply implies when responses of the act of observation are taken as they are i.e. at the face value, we obtain data. Obviously, responses would vary according to enquiry and observations emanating there from. They can be numbers, string or texts and even pictures. When data are represented by numbers we get quantitative data, else for others we get qualitative data.

A number is a mathematical object used to count and measure. In mathematics, the definition of number has been extended over the years to include such numbers as zero, negative numbers, rational numbers, irrational numbers, and complex numbers.

Mathematical operations are certain procedures that take one or more numbers as input and produce a number as output. Unary operations take a single input number and produce a single output number. For example, the successor operation adds one to an integer, thus the successor of 4 is 5. Binary operations take two input numbers and produce a single output

number. Examples of binary operations include addition, subtraction, multiplication, division, and exponentiation. The study of numerical operations is called arithmetic.

A notational symbol that represents a number is called a numeral. In addition to their use in counting and measuring, numerals are often used for labels (e.g. telephone numbers), for ordering (e.g. serial numbers), and for codes (e.g. ISBNs).

The most familiar numbers are the natural numbers or counting numbers: 1, 2, 3 and so on. Traditionally, the sequence of natural numbers started with 1 (0 was not even considered a number by the Ancient Greeks.) However, in the 19th century mathematicians started including 0 within natural number system. In the decimal system (i.e. 10 base system), the symbols for natural numbers are written using ten digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. In this base ten system, the rightmost digit of a natural number has a place value of one, and every other digit has a place value ten times that of the place value of the digit to its right.

The negative of a positive number is defined as a number that produces zero when it is added to the corresponding positive number. Negative numbers are usually written with a negative sign (a minus sign). As an example, the negative of 7 is written ?7, so that 7 + (?7) = 0. When the set of negative numbers is combined with the set of natural numbers (which includes zero), the result is defined as the set of integer numbers, also called integers.

On the other hand, a rational number is one that can be expressed as a fraction or ratio with an integer numerator and a non-zero natural number denominator. Fractions are written as two numbers, the numerator and the denominator, with a dividing bar between them e.g. .

The real numbers include all of the measuring numbers. Real numbers are usually written using decimal numerals, in which a decimal point is placed to the right of the digit with place value one. Every rational number is also a real number. It is not the case, however, that every real number is rational. If a real number cannot be written as a fraction of two

integers, it is called irrational number. For example, the real number ? (pi), or 22/7 neither ends nor forever repeats. It, therefore, cannot be written as a fraction, and is an example of an irrational number.

Moving to a greater level of abstraction, the real numbers can be extended to the complex numbers. Discussion on complex number is, however, beyond the scope of this lesson.

After the various types of numbers, let us briefly discuss the types of data in general. As has been mentioned earlier data can be quantitative or qualitative. Depending on unit of collection, data can also be classified as cross sectional, time series or panel. The data is called cross sectional when it is collected from different sections at the same point of time. Usually it is denoted by xi; where "i" stands for individual units. When data is collected for the same unit at different points of time then it is called time series data. For instance we can record yearly income of a person for say 25 years. Usual notation for time series data is xt, where "t" is the time period referred to. Panel data, on the other hand, are those which are collected for the same unit over a period of time. Panel data are represented as xit, where i i.e. cross section units can be 1, 2, ……n and t i.e. time period can be 1, 2, 3, …. k. Most importantly, each of these data types are treated and analysed differently. Researchers, therefore, must have clear idea of data which may be of use in studying stated objectives.

Besides, data can also be primary and secondary. The primary data are collected by the researchers themselves while secondary data are collected by other than the researchers. Researchers must, therefore, know the mechanics of primary data collection as well as sources of secondary data. In chapter 7 we have dealt with mechanics of primary data collection in some detail.

It is very important to realise that dealing with numbers or more generally quantitative data, at the very first step, requires a distinct measurement scale. In fact, data measurement is a very important issue in research. Below we will discuss issues related to data measurement.



CHECK YOUR PROGRESS

Q.1: Differentiate between unary and binary operations.

.....

.....

Q.2: What is an irrational number? Give an example.

.....

.....

Q.3: What is time series data?

.....

.....

9.4 DATA MEASUREMENT TECHNIQUES

Let us begin by understanding what is meant by a measurement scale. For a moment, consider that you are standing in a room somewhere and you want to locate your position. Now, your location can be measured as a distance from any of the four walls of the room. It is possible that depending on the wall you choose, your distance can vary. The point here is that for measuring a distance one needs to refer to a particular point from where the measurement is actually to be taken. Mathematically, this point of reference is known as the origin of the measurement.

Next important issue involved here is the unit of measurement. You know that distance can be measured in different units - metre, centimetre, millimetre etc. Note that these units of measurement are nothing but intervals that divide the total scale into number of constituent parts. A careful examination will reveal that these units are, in fact, indicative of number of intervals created within the scale - 'centi' is indicative of hundred, 'mili' is indicative of thousand and so on. Important point here is to note that these intervals can either be equal or unequal. Very common example of an equal interval scale is decimal scale and that of an unequal scale is logarithmic scale. In the case of the first, each interval is equal to the preceding and succeeding intervals. In the case of the second, each interval succeeds by

a larger interval. While dealing with quantitative data this equality or inequality of interval becomes very crucial because some arithmetic operations like addition and subtraction are meaningful only when intervals are equal.

It is, therefore, evident that a measurement scale essentially relates to two important notions - origin and interval. Depending on the nature of origin and interval we can have different type of scales, and following that different tools of numerical analysis. Let us spend some time in briefly discussing different type of scales.

Depending on the type of origin and interval used in measurement scales there can be four types of scales. First and the weakest of all is the nominal or categorical scale. In social sciences, it is quite common to denote or nominate different categories or options with numbers. For instance, categories like 'male' and 'female' are often represented by say 1 and 2 respectively. Similarly options like 'yes' and 'no' can be represented by 1 and 0 and so on. Note that here the numbers are simply used to nominate or denote various categories and options. There is no origin or/and interval involved in this designation. For instance, while denoting male as 1 and female as 2, no origin is used to measure the type of sexes. As a result, denoting male by 2 and female 1 instead does not make any difference. Similarly, there is no interval involved either since one cannot divide the difference of sexes into number of intervals.

Next one is called the ordinal scale. An ordinal scale is based on an arbitrary origin. Let us take one example. Suppose a person is judging a beauty pageant. At best he can rank (i.e. place them in some order) the contestants. There is no way he can measure beauty of the contestants in the sense that he can find the exact amount of difference in beauty of any two contestants. Further, if same pageant is judged by another person there is no guarantee that he would give the same ranking. This is because there is no absolute origin involved in this measurement. Judging a beauty pageant is a mental exercise and is based on individuals' value judgement. Therefore, the concept of origin may vary from person to person. The ordinal scale is therefore, one where only relative ranking or ordering is possible, not the exact measurement.

Third type of scale is called interval scale. When values are measured with respect to a fixed origin with equal intervals then it is called interval scale. In most of the cases zero is treated as the fixed origin. When this is so, scale is referred to as ratio scale.

It is important to note the differences ordinal, interval and ratio scale. The ordinal scale involves an arbitrary origin meaning that concept of origin can very from person to person. Interval scale involves a fixed origin, need not necessarily be zero, but in case of ratio scale the origin has to be zero.

In quantitative research idea and application of measurement scale occupies a crucial place since various statistical and mathematical tools and operations are limited by the type of scale used in measurement. For instance, mean is applicable only for values measured in interval or ratio scale; if the values are ordinal appropriate measure will be median rather than mean. For nominal values neither mean nor median will be applicable and one has to use mode as a substitute and so on. Point is before embarking upon any quantitative analysis it absolutely necessary to know the quantities that we propose to analyse for not all methods and tools are permissible for all the types of quantities.

9.5 EDITING OF DATA

After collection of data, before the researcher analyses them it is essential to edit the data so that reliable and meaningful results are obtained. Note however, that editing is not manipulating. Editing is done for specific purposes without changing the information contained in data. Editing is carried out to in the first place to check whether there is any invalid code or value in the data. Invalid code or value is one which is not defined for the study. For example if male and female are coded as 1 and 2 value of 3 is not defined. While editing the researchers has to make sure that all data are in the values of 1 and 2. Second, while editing consistency amongst data are to be checked. For instance, while asking about reproductive health related questions it is expected that respondent should be married female. Therefore, such responses are to be checked thoroughly whether they have been provided by married females or not. Third, editing also involves classification of data. We will discuss classification below in detail.

9.6 CLASSIFICATION OF DATA

Sometimes, we may need to classify the frequency for making sense. Consider the following ages of 40 people:

12	23	21	28	31	46	49	51	69	72
84	16	77	43	37	33	26	19	54	42
11	06	17	33	45	30	40	55	73	38
14	22	26	33	36	76	80	67	62	59

Note the mere frequency will not summarise the above data effectively, for values are not repeated much. One therefore, can devise a classification scheme so that the data can be summarised properly. Let us make the age-classes as "up to 18", "18 to 30", "30 to 45", "45 to 60" and "more than 60". For classification tally marks are used for each class. The classified frequency then would be more convenient for summarising the distribution as shown below:

Age classes	Frequency
up to 18	7
18 to 30	6
30 to 45	11
45 to 60	7
more than 60	9

Note that classes can either be discrete or continuous. For instance, two classes 20-25 and 26-30 are discrete while the classes 45.5 - 50.5 and 50.5 -55.5 are continuous. It is to be noted that whether classes will be discrete or continuous depend on data type as well as kind of analysis targets. If data values are discrete in nature classes need to be discrete and for continuous data classes have to be continuous. Also certain analysis is carried out only for continuous classes. We will know more about the differences of discrete and continuous in the next section of the chapter.

It is, however, possible to make discrete classes continuous. It can be done simply by subtracting 0.5 from the lower class limits and adding 0.5 to the upper class limits. For example two discrete classes 20-25 and

26-30 can be made continuous as 19.5 - 25.5 and 25.5 - 30.5. It may be noted that the value exactly equal to 25.5 will be recorded in the second class and not in the first class.

9.7 TABULATION OF DATA

In quantitative research we deal with quantitative variables. A variable implies an attribute, values of which can vary. For example, if one wishes to study weight then weight will be one's variable of interest.

Fundamentally, variables (i.e. values of variable) can be either discrete or continuous. Discrete variables are whole numbers used primarily for counting, e.g. number of rooms, number of students, number of households etc. On the other hand continuous variables can be fractions, however small, apart from being whole numbers, e.g. distance, income, expenditure, height, age, weight etc. Continuous variables are primarily used for measurement rather than counting. Note that fraction does not carry meaning in case of discrete variables.

In short, a variable comprises of a set of responses and quantitative variables, in particular, comprises of numbers measured in terms of a particular scale. Let us suppose that one is interested in the study of the attribute age and he records ages of 10 persons as given below:

Age : 12, 23, 32, 18, 25, 29, 36, 41, 37, 26

It is evident from above that the variable age is given by a set of responses or values. The set of values (or responses) of a given variable, thus obtained, is called the distribution of the variable. In case, we are interested in only one variable, the distribution is called univariate distribution. When two variables are considered together, the two distributions together are known as bivariate distribution. When number of variable is more than two, it is called multivariate distribution.

The values of a variable can appear only once in a distribution or several times repeatedly. The number of times a value appears in a distribution is called its frequency. In other words, frequency of a value indicates how frequently the value appears in the distribution. When values are arranged along with their respective number of appearance or frequency

the distribution is called frequency distribution. For example, consider the following values of a distribution

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

The frequency distribution of the distribution is given as below

Value (X)	:	1	2	3	4
Frequency (f)	:	6	8	8	2

Speaking mathematically, frequency distribution f represents the number of times the value x is repeated in a particular distribution or series of a variable.

After collecting relevant data, researchers need to present them in an effective way to put across the basic arguments. Frequency tables are the first step towards this. A frequency table describes the data in a condensed way and shows how the values are repeated in a series. For instance, suppose one survey 15 people about their gender and data obtained are given as below

Sl No	Gender (Male =1, Female=2)
1	2
2	2
3	2
4	1
5	1
6	1
7	2
8	1
9	2
10	2
11	1
12	1
13	2
14	1
15	1

The above data can be summarised by frequency distribution as

Gender	**Frequency**
Male (1)	8
Female (2)	7

The above frequency table tells us that there are 8 males and 7 females in the data. The frequency distribution, however, can summarise data involving only one variable. One can use tables for summarising more than one variable as well. Such tables are called cross tables. The following is a typical example of cross-table:

Age classes	**Sex**		**Total**
	Male	**Female**	
up to 18	4	3	7
18 to 30	2	4	6
30 to 45	10	1	11
45 to 60	5	2	7
more than 60	4	5	9
Total	**25**	**15**	**40**

It may be noted that in cross-table we have three different totals: total over rows, total over columns and total over the table. Accordingly we will have three different percentages for each cell value. One has to be careful about interpreting the results. Let us consider the value of male in the age group 30 to 45 viz. 10. Now one can calculate column percentage for the value as = 40 percent. Similarly row percentage for the value can be calculated as =91 percent. Also, total percentage for the value can be obtained as =25 percent. It is important to understand that all three percentages mean different things. The first case implies that out of total male 40 percent fall in the age group of 30- 45. The second means out of total sample in the age group 30-45 some 91 percent are male. While the third case indicates out of the total sample 25 percent are male and belong to age group 30-45.

It is thus, clear that for single variable distribution frequency table and for two variable distribution cross tables can be used. However, in most of the cases we have to deal with more than two variables. How do we accommodate a third variable in a table?

A third variable can be accommodated in a cross table by using two concepts - called nesting and layer. Let us accommodate another variable say place of residence (viz. rural and/or urban) in the age-sex cross of above to understand these two concepts. When the variable is accommodated within column it is called nesting as shown below:

Age classes	**Male**		**Female**		**Total**	
	Rural	**Urban**	**Rural**	**Urban**	**Rural**	**Urban**
up to 18	2	2	1	2	3	4
18 to 30	2	0	3	1	5	1
30 to 45	5	5	1	0	6	5
45 to 60	0	5	1	1	6	1
more than 60	1	3	3	2	4	5
Total	**10**	**15**	**9**	**6**	**24**	**16**

On the contrary the third variable 'place of residence' can be accommodated in rows using the concept of layers which is demonstrated

Age classes	**Sex**		**Total**
	Male	**Female**	
Rural			
up to 18	2	1	3
18 to 30	2	3	5
30 to 45	5	1	6
45 to 60	0	1	6
more than 60	1	3	4
Total (rural)	**10**	**9**	**24**
Urban			
up to 18	2	2	4
18 to 30	0	1	1
30 to 45	5	0	5
45 to 60	5	1	1
more than 60	3	2	5
Total (urban)	**15**	**6**	**16**



CHECK YOUR PROGRESS

Q.4: What is interval scale?

.....

.....

Q.5: What are discrete variables ?

.....

.....

.....

Q.6: Which is the weakest scale?

.....

.....

.....

Q.7: Where are continuous variables generally applied?

.....

.....

.....



9.8 LET US SUM UP

- In this chapter we have discussed the definition of data and also various types of data. We have discussed primary and secondary data, quantitative and qualitative data, cross-section and time series data. We have also discussed various measurement issues related to data and data types. We have introduced the idea of scale of measurement and elaborated nominal, ordinal, interval and ratio scales and their utilities. We have also briefly discussed about data editing and its requirement. Then we have gone onto describe classification of data. Next we have discussed about single variable frequency tables and two-variable cross tables. Finally we have introduced the idea of nesting and layer to accommodate more than two variables in a table.



9.9 FURTHER READING

1. "Fundamental of Statistics"; S C Gupta, S Chand and Company, New Delhi
2. "Fundamental of Statistics"; A M Gun, M K Gupta, B Dasgupta, Vol. I, World Press
3. "Basic Statistics"; A L Nagar, Oxford University Press
4. "Basic Statistics"; B L Agarwal, New Age International
5. "Statistics without Tears", D Rowntree, Penguin Books



9.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: Mathematical operations are certain procedures that take one or more numbers as input and produce a number as output. Unary operations take a single input number and produce a single output number. For example, the successor operation adds one to an integer, thus the successor of 4 is 5. On the other hand, binary operations take two input numbers and produce a single output number. Examples of binary operations include addition, subtraction, multiplication, division, and exponentiation.

Ans. to Q. No. 2: The real numbers include all of the measuring numbers. If a real number cannot be written as a fraction of two integers, it is called irrational number. For example, the real number ? (pi), or 22/7 neither ends nor forever repeats. It, therefore, cannot be written as a fraction, and is an example of an irrational number.

Ans. to Q. No. 3: When data is collected for the same unit at different points of time then it is called time series data.

Ans. to Q. No. 4: When values are measured with respect to a fixed origin with equal intervals then it is called interval scale.

Ans. to Q. No. 5: Discrete variables are whole numbers used primarily for

counting, e.g. number of rooms, number of students, number of households etc.

Ans. to Q. No. 6: Nominal or categorical scale is the weakest scale.

Ans. to Q. No. 7: Continuous variables are primarily used for measurement rather than counting. Continuous variables can be fractions, however small, apart from being whole numbers, e.g. distance, income, expenditure, height, age, weight etc.



9.11 MODEL QUESTIONS

A. Short Questions (Answer each question in about 150 words)

Q. 1: What is data?

Q. 2: What are the different types of data?

Q. 3: What do your understand by a measurement scale?

Q. 4: What are the basic elements of a scale?

Q. 5: Define different measurement scales.

B. Long Questions (Answer each question in about 300-500 words)

Q. 1: Distinguish between primary and secondary data.

Q. 2: What is data editing? Why is data editing necessary?

Q. 3: What is classification of data? How a discrete class can be made continuous?

Q. 4: What is a cross table? How can one accommodate a third variable in a cross-table?

*** ***** ***

UNIT 10: DATA ANALYSIS

UNIT STRUCTURE

10.1 Learning Objectives

10.2 Introduction

10.3 Data Presentation

10.4 Hypothesis testing and Various Tests and Interpretations

- 10.4.1 T-Test and its Uses
- 10.4.2 F-test and its Application
- 10.4.3 Chi-Square Test and Application

10.5 Let Us Sum Up

10.6 Further Readings

10.7 Answers to Check Your Progress

10.8 Model Questions

10.1 LEARNING OBJECTIVES

After reading this unit, you will be able to-

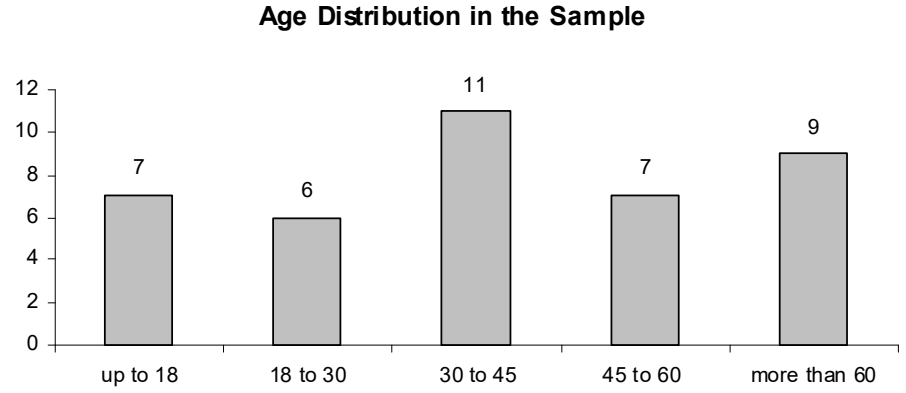
- outline the various diagram and graphic presentation methods
- describe the basics of hypothesis testing
- explain how to interpret data
- discuss how to report your analysis

10.2 INTRODUCTION

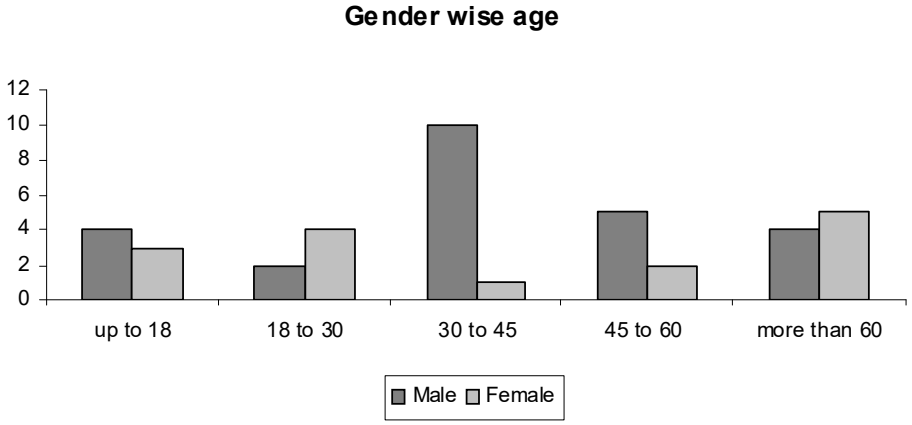
In the previous unit we have seen how data can be presented in tables. Data can also be presented through diagrams and graphs. But the chief purpose of research is to make generalisations. The process of generalisation involves formation of hypothesis and validation of hypothesis with the help of data. Based on the results of the test of hypothesis, report is prepared. In this unit, we will try to introduce key elements of the process of hypothesis testing together with various tests and their utilities. We will try to understand how results of various tests are interpreted. Finally, based on such interpretations, we will discuss how to prepare a report.

10.3 DATA PRESENTATION

Besides tables, data can also be represented by suitable diagrams. Most widely used diagrams are bar diagrams, histograms and pie diagrams. A bar diagram or bar graph is a chart with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. The frequency distribution of age classes can be represented by bar diagrams as shown below

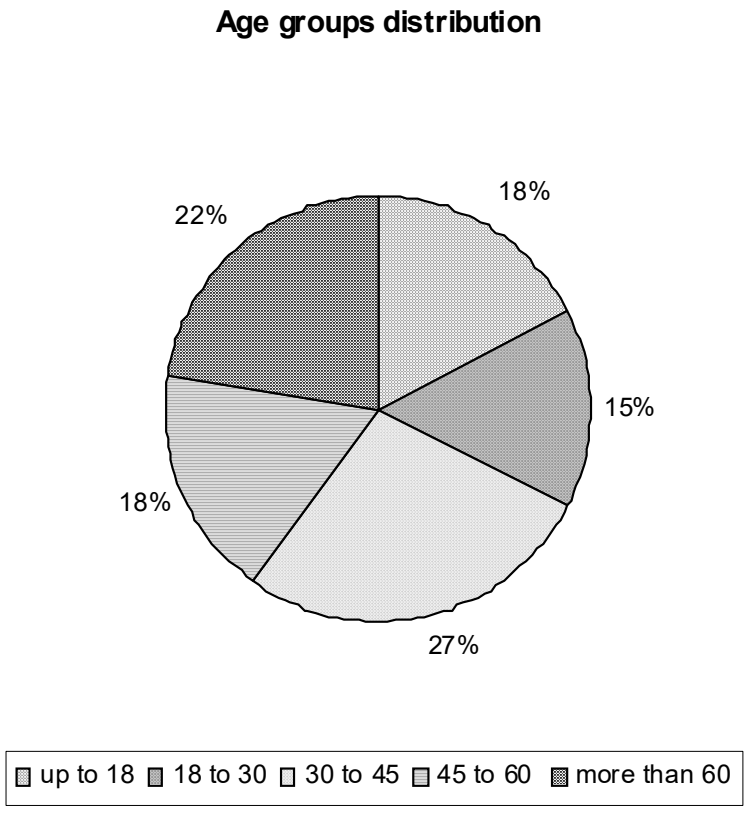


A bar diagram is very useful if you are trying to record certain information whether it is continuous or not continuous data. It is nevertheless more suitable for representing discontinuous data. There can be variations in the bar diagrams as well. For instance if one category has more than one value, for instance the male and female as shown in the cross table above, we use clustered bar diagrams. The appropriate bar diagram for the cross table given above will be as shown below



Bar charts also look a lot like a **histogram**. They are often mistaken for each other. A histogram is a graphical representation of data showing a visual impression of the *distribution* of data. It is normally used for continuous distributions.

Another common way of representing data is **pie diagram**. A **pie diagram** (or a **circle graph**) is a circular chart divided into **segments**, illustrating *proportion*. In a pie chart, the arc length of each segment (and consequently its central angle and area), is proportional to the quantity it represents. The angles are measured with 1 turn as unit. Together, the segments create a full disk. Pie charts can be an effective way of displaying information in some cases, in particular if the intent is to compare the size of a slice with the whole pie, rather than comparing the slices among them. Pie charts work particularly well when the slices represent 25 to 50 per cent of the data. We can see the pie generated for our data on age groups as below





CHECK YOUR PROGRESS

Q.1: What is a bar diagram?

.....

.....

Q.2: What do you mean by pie chart?

.....

.....

Q.3: Describe histogram.

.....

.....

10.4 HYPOTHESIS TESTING, TESTS & INTERPRETATIONS

Research starts with an initial statement known as **hypothesis**. In statistics, the initial hypothesis is called **null hypothesis**. The hypothesis or null hypothesis to be precise is validated through data and after validation, the hypothesis turns into a **theory** which is nothing but a *broad generalisation* for entire population. Clearly, therefore, statistical inference is a key element in theory building. In this section we will briefly discuss about important elements of the statistical inference.

Let us start with the clear articulation of our exact problem with the aid of an easily comprehensible example. Suppose that we have a class of students. The normal assumption of deciding about promoting a student to a higher class requires that he or she acquires knowledge adequately at that level. For knowing whether the student has acquired the required knowledge adequately, a test is held. Based on the marks obtained in the test we decide about his or her promotion to a higher class. Note that this essentially resembles to the process of statistical testing. The process can be illustrated as under:

The **null hypothesis** here is that "the student is to not be promoted". To accept or reject this hypothesis we need a *criterion*. This criterion is called **test**. In our case annual examination can be called as the much

needed test. Every test is supposed to generate a *value* or *score* known as **test value**. In our case student's mark in the examination is the test value. Every test also has a **critical value**. The critical value defines the *acceptance and rejection* of the null hypothesis and it is given by some probability level. In our case the pass mark is the critical value since it defines whether the student is to be promoted or not. Statistical testing requires that the test value is compared with the critical value for possible acceptance and rejection. If for instance, the student's mark is more than the required pass mark we *reject* the null hypothesis and vice versa.

Note here how we have articulated the null hypothesis. When we want to proof something null is usually articulated otherwise. The reason behind this is that burden of evidence is fairly limited for disproving a hypothesis. For example if suppose we want to test whether man is mortal. If we articulate null as "man is mortal" – to reject this we need to wait for evidences until we find an immortal man! On the other hand if we take our null as "man is immortal" our burden of evidence is fairly limited – new need only one mortal man to reject the null!

The point here is that usually null is formulated for rejection support. However, this does not imply that all statistical tests will be based on rejection support hypothesis. Nevertheless, the fact remains that burden of evidence is fairly limited in rejection support rather than acceptance support. Therefore, formulation of null is very elementary in statistical testing.

Also recall that since we try to make generalisations based on sample results there always remains a scope for making errors. These errors are known as **sampling errors** since they arise because of the very act of sampling itself. Very clearly, we have two specific types of sampling error as mentioned below

Type I we may accept a hypothesis which should not have been accepted

Type II we may reject a hypothesis which should not have been rejected

The type I error is also known as **false positive** and the type II error is known as **false negative**. Given this two types of errors, one can also

find their rates of occurrence. For instance, say out 100 how many times they occur. If any one them occurs 5 times, then, obviously the rate is given by 5 percent. The rate of type I error i.e. the rate of false positive is known as **significance or level of significance** (also denoted by α) while the rate of type II error i.e. the rate of false negative is known as **power** (denoted by β). Ideally, we want low levels of these two rates for safe generalisation. However, these two has a trade-off one can be lowered at the expense of the other.

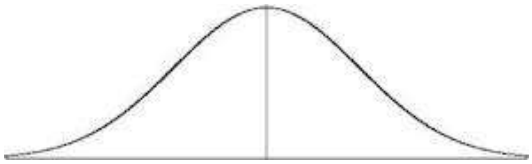
Further note that other way of expressing the level of significance is the **level of confidence**. If level of significance is say 5 percent the rate of type one error is 5 out of every 100. Speaking other wise we are correct in 95 out of 100, indicating thereby our level of confidence. Hence, 1-level of significance is the level of confidence.

For all statistical tests there is a predefined probability distribution of critical values for different levels of significances. These critical values provide us minimum evidence required to accept or reject the null hypothesis as the case may be. One has to compare one's own observed test value against the theoretical critical value for possible acceptance and rejection. The associated levels of significance will tell us the probability of making type I error while making decisions based on such results.

Recall our student and examination example. As there are different examinations for different occasion viz. annual examination for regular promotion, entrance examinations for getting admissions etc., there are different statistical tests as well suited for different occasions and situations. Below we will discuss very briefly some of the most popular and useful tests.

Most of the statistical analysis assumes **normality** on the part of the distributions under the study. Simply speaking, normality refers to an average situation of any phenomenon where "extremes are rare and average is more common". This means that in any situation, too high or too low values will be extremely rare, and most of the times, we will have cluster of values around the average, which as we will see shortly, nothing

but some central value. Following this rule, any distribution ideally should be normal and shape of a typical normal distribution is as given below:



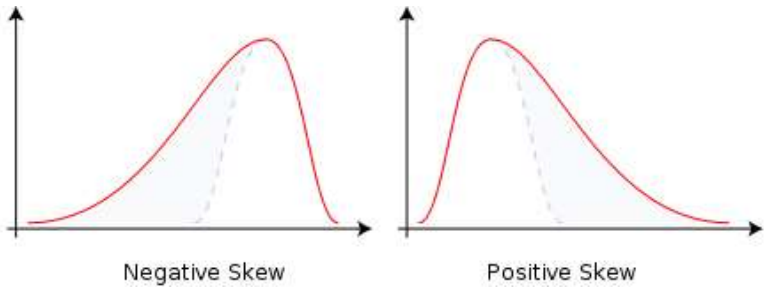
The above curve is popularly known as **normal curve**. We will discuss more of this curve and its property in some detail in the next unit. At this point, it is important to remember that majority of the statistical analysis requires fulfilment of the normality assumption for valid interpretation of results. The statistical analyses which require the normality assumptions to be fulfilled are known as **parametric statistics**. There are, of course, some analyses which do not need the assumption of normality, and they are called **non-parametric statistics**. Along with scales of measurement, assumption of normality also plays a critical role in quantitative analysis. For understanding process of testing hypothesis features of normal curve need careful observation.

Firstly, the normal curve is *symmetrical about mean*. This means that the mean is located at the middle and for same distances from it, both ways, the frequency is also same. For example say *h* unit more than the mean will have the same frequency as the *h* unit less than the mean. In other words, putting it simply, if we fold the curve about the mean both the tails should make a perfect overlap.

Noticeably, the normality would indicate that mean, median and mode are the same value. Evidently, mean divides the distribution into two equal parts since it is symmetrical about the mean and also it is the maximum value, hence, demoting the mode value as well.

If, however, there is a deviation from normality mean, median and mode values will differ and it will produce longer tails towards either side. When the tail will be elongated towards positive side we call it **positive skew** and when the tail elongates towards negative side we call it **negative skew**. Clearly deviation from normality is indicated by the presence of skews i.e. elongated tails. In case of skewed distribution, mode remains to be the

maximum as always, the mean moves towards the tails due to the impact of extreme value. The median stays in between the mode and the mean. Empirically, in such distributions the distance between the mean and median is about one-third of the distance between the mean and mode. Karl Pearson expressed this relationship as: Mode = Mean - 3 [Mean - Median]. This relationship between mean, median and mode with skewness is very important in analysing the data and making statistical interpretations.

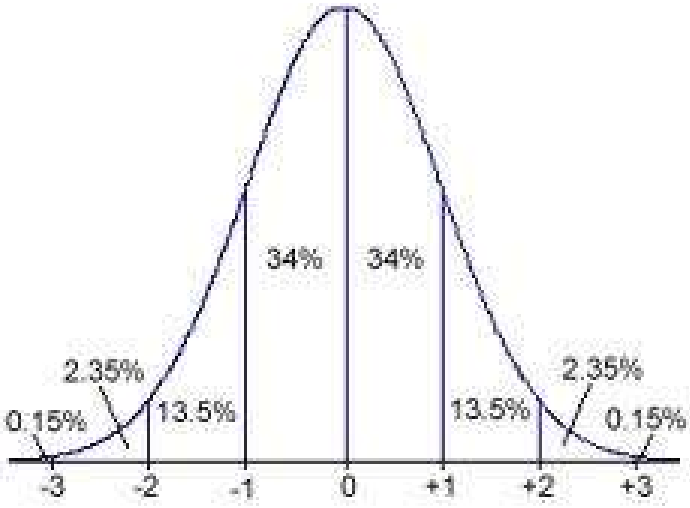


As has been pointed out the normal curve is symmetrical about the mean. Apart from the symmetry, there is another very useful property of the normal curve. The area under the normal curve representing probabilities of the values expressed in terms of the mean and standard deviations are standardised and already tabulated. Any value, say *x*, of a distribution can be expressed in terms of the mean and standard deviation of the distribution as

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

Where $\bar{x}$ and σ are the mean and standard deviation of the distribution. Recall that this simply indicates change of origin and scale. In this transformation we are re-scaling the values taking the mean as the origin and standard deviation as the interval. This scale is given by the horizontal axis of the normal curve where the mean can be treated as the origin and intervals are given by the standard deviations. The values of *Z* actually put the values of *x* in this new scale implying *how many standard deviations away is the value from mean* in the axis. For example, if mean is 5 and standard deviation is 3, then *Z* transform of 20 will be (20-5)/3= 5, which simply means that original value of 20 when put on the *Z* scale will lie 5 standard deviations away from the mean to its right side. This Z value is called **standard normal variable.**

The beauty of the Z is that for the entire range of values of Z, we have defined values of corresponding probability. The probabilities are given by the Z-table. The summary of the scheme is given below



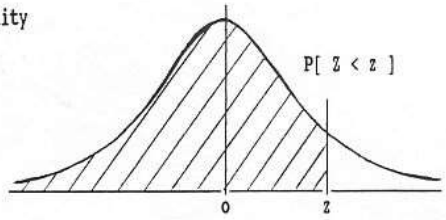
Note the symmetrical property of the curve. It is found from the above that about 68 percent of the values will be with the range of one standard deviation away from the mean both sides together. Similarly about 95 percent should lie in between two standard deviations away and 99 percent should be within three standard deviations away from the mean, if the given distribution is normal. Most importantly, not only probabilities are tabulated for 1, 2 and 3 but also in between 1 and 2, 2 and 3 as well and the results are summarised in a table known as Z table. A typical Z table is presented below. It may be seen from the table that probability associated with 1.05 is 85.10 percent, 1.31 is 90.49 percent and so on (try to figure out how). This has got enormous utility for practical purposes. Using this table, one can easily find out the proportion or probability of any value, or between any two values in the population from the sample data, provided the sample distribution is normal. We will try to see how this property can be used in real life situations with couple of examples. These examples will also indicate why the assumption of normality is so vital in statistical analysis and interpretations.

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $P[\,Z < z\,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Working Example 1

Suppose that legal speed limit of vehicle in a city is 40 km. A study was conducted to find out how many vehicles actually violate the legal norm. The study was based on a random sample of 200 vehicles. It was found that mean speed limit was 50 and standard deviation was 5. Also distribution of the sample speed is found to be normal. What can be concluded on the basis of the information?

Solution

Given that the distribution is normal and mean and standard deviation are 55 and 5 respectively. Hence, legal limit of 40 can be transformed as

$$Z = \frac{40 - 50}{5} = -2$$

This implies that the legal limit of 40 is located in the Z scale of the sample speed distribution 2 standard deviations away from the mean to the left. We also have the probability of the values in between -2 to 0 as 47.5 and 0 above 50.0. Therefore, total proportion above -2 is 47.5+50.0=97.5. This indicates that there is a *probability* that 97.5 percent of the vehicles are violating the legal speed limit in the city.

Note that this inference is possible only because of the normality in sample values.

Working example 2

If mean of a distribution is 24 and standard deviation is 2, find out the probable proportion of values lying in between 20 and 26 when the distribution is normal.

Solution

Z transformations of the values are given by

$$Z_1 = \frac{20 - 24}{2} = -2 \text{ and } Z_2 = \frac{26 - 24}{2} = 1$$

From the diagram the probable proportion of values between -2 and 1 is 81.5 percent.

This Z transformation is, therefore, a useful tool for statistical inference. Based on this from sample information one can safely conclude about the population provided that the sample distribution confirm to normality. Clearly it is a parametric measure.



CHECK YOUR PROGRESS

Q.4: What is hypothesis?

.....

.....

Q.5: What is test value?

.....

.....

Q.6: What is level of significance?

.....

.....

Q.7: What are positive skew and negative skew?

.....

.....

10.4.1 t-Test and its Uses

In quantitative research we very commonly compare means. For instance, we may be interested in studying income differences amongst working male and females. Or we may be interested in studying IQ levels of boys and girls etc. Here we, typically, want to compare *means* i.e. mean male income and mean female income or mean boy IQ and mean girl IQ etc. Obviously, since we are studying on sample basis there will be some differences. However, there is a limit to that. If differences are large enough which is not due to sampling and all then the difference is called **statistically significant**. When test result is found to be significant (at some level of significances) we simply mean that difference is genuine and there must be some reason to it. **Student t test** or popularly known as **t-test** is normally used to find out statistically significant differences amongst **two means**.

The t-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name). He published the test in Biometrika in 1908, but was forced to use a pen name by his employer, who regarded the fact that they were using statistics as a trade secret.

The t-test can be applied to various situations. Primarily, three situations emerge in research yielding three different types of t-tests. Imagine a situation we have conducted a sample survey and obtain our mean. Also we know that similar study was conducted earlier and we have that mean as well. Now we may be interested in studying whether our mean is different from the earlier mean. In this case we have one sample and the mean is tested against a pre-defined mean. This when applied is known as **one sample t-test**. The actual test procedure goes like this: let us assume that sample mean is $\bar{x}$ and theoretical mean is x_0 and also sample standard deviation is given by *s* then test statistic t is

$t = \frac{\bar{x} - x_0}{s / \sqrt{n}}$, where *n* is the sample size.

Null hypothesis is given by $\bar{x} = x_0$. The value of calculated t is to be compared with the tabulated value of t for the (n-1) degree of freedom. If the calculated t is found to be greater than tabulated t, the null is rejected and significant difference is concluded. Important here is to note that the above test is used *only when population variance or standard deviation is not known*. If the population variance is known as *V* then the above distribution will be tested by Z statistics where Z is given by:

$Z = \frac{x - x_0}{\sqrt{V / n}}$

Second situation is one when we have two *independent* or mutually exclusive samples and their means. Two samples will be independent when sample units of one do not have possibility of appearing in the other, for example, take any samples of male and female. Note that males will be excluded from the sample of females and vice versa. In other words the two are independent. When t-test is applied to two independent samples it is called **independent t test**.

Third situation emerges when we have joint distributions of values. For example we are interested in studying marks given by two examiners. Here we have a joint distribution of marks for each

student. We may be interested in finding whether the two examiners differ significantly in marking. When applied to such situations t-test is called **paired t test**.

Further, there are other two issues involved. The two samples can be of same size or different size. Also the two distributions can have equal variances or unequal variances. Depending on these factors actual formula of t-statistics can vary. These days, the testing is done on computers which handle these complexities quite easily.

Any case, one needs to understand the important procedural steps in conducting t-test. The test is applied to *compare two means* with the null hypothesis that there is no difference in them. When the test result is found significant we reject the null and say the there is significance difference in means compared. Then one must try to attribute reasons as what might have caused the difference.

10.4.2 F-Test and its Application

As has been pointed earlier the t-test can be applied only for comparing two means. There can be situations when researchers may have to compare more than two means. Then F-test becomes handy. The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the F-statistic as the variance ratio in the 1920s. It is, therefore, also applied to test the fit of model to the data, in explaining proportion of variance explained by a model out of the total variance observed.

It is important to observe that all the three tests – Z, t and F hold good only under certain assumptions. Most important among them is the assumption of normality. Unless distribution is normal these tests, although widely used, can not be applied to data. Two issues, therefore, merit some discussion: how do we know that our data confirm to normality and second if this is not so what is to be done.

First of all, let us confess that in real life situations, we hardly get data which are perfectly normal. In other words, in real life we are bound to have some degree of deviations from normality.

Therefore, one has to see whether deviation is significantly large enough to cause problems in analyses. There are two ways which can detect deviations from normality viz. graphical and formal test of normality. Most widely used graphical method is the **box-plot method**. The box-plot is a quick way of examining one or more sets of data graphically. Most of the statistical software produces box-blot just in a click. It provides a box within which most of the values lie with a line of median. If the median line lies exactly at the middle the box then normality is indicated. This method is very crude and rudimentary. One can also take help of formal test of normality like Kolmogorov-Smirnov test which indicates whether there is any significant deviation of normality. If there is any, one can adopt suitable transformation methods like change of scale i.e. use of logarithmic or exponential scales instead of decimal scale to improve normality. If however, no transformation works then one has to find equivalent distribution free tests for testing hypothesis and make generalisations.

10.4.3 Chi-Square (χ^2) Test and Application

One of the most important and most widely used distribution free statistical tests is the chi-square (χ^2) test. This means that we do not need normality for applying this test to our data. Moreover, this test can be applied to categorical variables as well. Logic of this test is discussed under. Let us consider the following data.

Hour:	1	2	3	4	5	6	7	8
Accident:	12	14	16	13	20	17	9	11

The data shows number of accidents occurring in different hours in a factory. Clearly, the 5th hour records the highest number of accidents. Can we then say that 5th hour is more prone to accidents?

The essential question is whether the distribution is fairly distributed or biased towards 5th hour. The concept of fairness, here, relates to equal probability. If the distribution is fair the each hour should have same

probability of having accidents, i.e. no hour should have bias for having more accidents. Probability being the average number of occurrence, one would see that a fair distribution should have 14 accidents per hour (note that the average accident is 14 in this case). So we expect that fair distribution should have been 14 per hour where as we have observed the actual as given by the series. Evidently, there is a deviation from fairness, which can be given as below:

1	*2*	*3*	*4*	*5*	*6*
Hour	**O**	**E**	**(O-E)**	**$(O-E)^2$**	**$(O-E)^2/E$**
1	12	14	-2	4	0.29
2	13	14	-1	1	0.07
3	16	14	2	4	0.29
4	13	14	-1	1	0.07
5	20	14	6	36	2.57
6	17	14	3	9	0.64
7	9	14	-5	25	1.79
8	12	14	-2	4	0.29
Total	**112**	**112**	**0**	**84**	**0.75**

The column 4 gives the deviations from fairness. Note that total deviation is 0. Recall that this is because of the property that sum of deviations from the mean is 0. To do away with the problem, we square the deviations and express that as proportion to the expected which is given in the column 6. The total 0.75 is known as the value of chi-square and it is formally given by

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i}$$

Like any other statistical test, the critical values for this test are also tabulated. The tabulated value for *n-1* degree for freedom i.e. 7 in this case at 5 percent level of significance is, for instance, 12.47. Since in our case the calculated value is found far less than what is tabulated, we can safely take the distribution as fair and hence there is no reason to believe that number of accidents is bias towards 5th hour. The high value observed in this distribution at 5th hour is only due to chance and not because of any external reason.

The second use of the chi-square lies in interpreting contingency tables. A contingency table classifies the distributions according to sub-categories. Simplest type of contingency table takes a 2 row 2 column form as given below:

	Class 1	Class 2	Total
Category 1	a	b	a+b
Category 2	c	d	c+d
Total	(a+c)	(b+d)	(a+b+c+d)

The chi-square statistics in this table is given by:

$$\chi^2 = \frac{(n-1)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

The null is taken as row and column are independent and tested for n-1 degrees of freedom. This test is very important for social sciences as contingency tables are widely used in almost all social science disciplines.

Idea of Degrees of Freedom (DF)

We have used the idea of degrees of freedom extensively above. This is a very important concept in statistical testing. Before giving a formal definition let us try to understand the logic of the whole concept first.

Let us consider an equation as under

$$x + y = 100$$

Note that we can choose any value of any *one* of the variables in absolute freedom (without any restriction). Moment we choose any value for any of these variables, the next is automatically get identified by the equation itself!

For instance, let us try to choose for *y*. We have absolute freedom in choosing values of *y* and we choose *y* as 20. Moment we choose y=20, x has to be 80 since the equation is to be satisfied. So our freedom of choice is only one. This degree of freedom can be seen as number of variables that we can choose without restriction. In general if there are *K* variables in total with one equation to solve we have *K-1* degree of freedom. Note that each equation stands for one restriction imposed over the system.



CHECK YOUR PROGRESS

Q.8: What is independent t-test?

.....

.....

Q.9: What is box plot method?

.....

.....

Q.10: What do you understand by chi-square method?

.....

.....



10.5 LET US SUM UP

- In this chapter we have discussed about various ways of diagrammatically representing data. We also have introduced the processes and principles of testing of hypothesis. We have introduced various tests like z test, t test, F test and chi-square test. We have also discussed their utility and applications. Further the chapter also discusses how data can be interpreted for reporting using the various test procedures.
- Bar charts also look a lot like a **histogram**. They are often mistaken for each other. A histogram is a graphical representation of data showing a visual impression of the *distribution* of data. Research starts with an initial statement known as **hypothesis**. In statistics, the initial hypothesis is called **null hypothesis**. The statistical analyses which require the normality assumptions to be fulfilled are known as **parametric statistics**. There are, of course, some analyses which do not need the assumption of normality, and they are called **non-parametric statistics**. The hypothesis or null hypothesis to be precise is validated through data and after validation, the hypothesis turns into a **theory** which is nothing but a *broad*

generalisation for entire population. One of the most important and most widely used distribution free statistical tests is the chi-square (χ^2) test.



10.6 FURTHER READING

1. *Fundamental of Statistics*; S C Gupta, S Chand and Company, New Delhi
2. *Fundamental of Statistics*; A M Gun, M K Gupta, B Dasgupta, Vol. II, World Press
3. *Basic Statistics*; A L Nagar, Oxford University Press
4. *Basic Statistics*; B L Agarwal, New Age International
5. *Statistics without Tears*, D Rowntree, Penguin Books



10.7 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: A bar diagram or bar graph is a chart with rectangular bars with lengths proportional to the values that they represent.

Ans. to Q. No. 2: A pie diagram (or a circle graph) is a circular chart divided into segments, illustrating *proportion*.

Ans. to Q. No. 3: A histogram is a graphical representation of data showing a visual impression of the *distribution* of data.

Ans. to Q. No. 4: A hypothesis is a proposed explanation for a phenomenon.

Ans. to Q. No. 5: To accept or reject this hypothesis we need a criterion. This criterion is called test. Every test is supposed to generate a value or score known as test value.

Ans. to Q. No. 6: The rate of type I error i.e. the rate of false positive is known as **significance or level of significance** (also denoted by α) while the rate of type II error i.e. the rate of false negative is known as **power** (denoted by β).

Ans. to Q. No. 7: If, however, there is a deviation from normality mean, median and mode values will differ and it will produce longer tails towards either side. When the tail will be elongated towards positive side we call it **positive skew** and when the tail elongates towards negative side we call it **negative skew**.

Ans. to Q. No. 8: When t-test is applied to two independent samples it is called independent t test.

Ans. to Q. No. 9: The box-plot is a quick way of examining one or more sets of data graphically.

Ans. to Q. No. 10: A chi-squared test is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true.



10.8 MODEL QUESTIONS

A. Short Questions

Q. 1: Draw a bar diagram for the following table.

Income Groups	Rural (%)	Urban (%)
Less than 10000	54	42
10000 to 40000	26	39
40000 to 80000	12	10
More than 80000	8	9

Q. 2: Draw a pie diagram for the following data.

Social Groups	Population (%)
General	62
Scheduled Caste	13
Scheduled Tribe	19
OBC	6

Q. 3: Clearly illustrate the process of statistical inference.

Q. 4: What do you understand by statistical significance and power?

B. Long Questions (Answer each question in about 300-500 words)

Q. 1: Explain the idea of skewness. Indicate the relation between mean, median and mode in relation to skewness of a distribution.

Q. 2: What is standard normal variable and what does it imply? Illustrate its utility with examples.

Q. 3: Distinguish between independent and paired sample t-tests. Distinguish between F and t test.

Q. 4: What is chi-square and how it is applied in quantitative research?

Q. 5: What is a contingency table and how chi-square is applied to it?

*** ***** ***

UNIT 11: SCALING TECHNIQUES IN RESEARCH

UNIT STRUCTURE

11.1 Learning Objectives
11.2 Introduction
11.3 Meaning of Scale and Scaling
11.4 Types of Scale
11.5 Scales to analyse Social distance
11.6 Importance of Scaling
11.7 Difficulties in Application of Scaling Techniques
11.8 Let us Sum up
11.9 Further Reading
11.10 Answers to Check Your Progress
11.11 Model Questions

11.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- explain the meaning of scale, scaling and types of scaling
- describe the importance of scaling and difficulties of using scaling teaching in sociological research.

11.2 INTRODUCTION

An important part of research is an understanding of the use of various scaling techniques while carrying out the research process. Such instruments are commonly referred to as as 'scales'. As time changes the instruments of measurement also changes. New techniques of measurement are being developed improve day by day. Scaling always hypothesizes the existence of a continuum of some kind. Its nature must be inferred from the character of the items selected to make up the scale. In this unit, we shall discuss about scale and scaling techniques, different types of scales and we shall also learn about social distance scale and sociometry scale and their application in Sociological research.

11.3 MEANING OF SCALE AND SCALING

Scaling is the branch of measurement that involves the construction of an instrument that associates qualitative constructs with quantitative metric unit. Scaling evolved out of efforts in psychology and education to measure "immeasurable" construct like authoritarianism and self-esteem. Scaling describes the procedures of assigning numbers to various degrees of opinion, attitude and other concepts. The process of measuring abstract concepts in concrete terms remains one of the most difficult tasks in empirical social science research. Scaling is a branch of measurement that involves the construction of measures by associating qualitative judgments about unobservable constructs with quantitative, measurable metric units. Stevens (1946) said, "Scaling is the assignment of objects to numbers according to a rule." In many ways, scaling remains one of the most arcane and misunderstood aspects of social research measurement, and it attempts to do one of the most difficult of research tasks- measure abstract concepts. The problem to which techniques are applied is that of ordering a series of items along some sort of continuum. In other words, they are methods of turning a series of qualitative facts into a quantitative series.

Originally the word 'scale' comes from the Latin word 'scala', meaning a ladder or flight of steps. A scale represents a series of ordered steps at fixed intervals used as a standard of measurement. Scales are used to rank people's judgments of objects, events, or other people from low to high or from poor to good. Scales can be unidimensional or multidimensional, based on whether the underlying construct is unidimensional (e.g., weight, wind speed, firm size) or multidimensional (e.g., academic aptitude, intelligence).The three most popular unidimensional scaling methods are: (1) Thurston's equal-appearing scaling, (2) Likert's summative scaling, and (3) Guttman's cumulative scaling. A scale is a continuum from the highest to the lowest points, and has intermediate points in between these two extremes. The scale point is higher than the third point and so on. A scale is a method of measurement. In social research; attitude, behaviour, and other qualitative characteristic are measured by means of different scales. The

scalability of the phenomena depends on its continuum. The various factors of the phenomena must be logically interrelated and should be capable of continued measurement. A scale must be reliable.

Scaling is the method of turning a series of qualitative facts (referred to as attributes) into a quantitative series (referred to as variable). The complexity of this undertaking, and hence it may have pitfalls, can best be seen by considering what is meant by ordering items along a continuum. For example, when one wishes to measure the size of a series of families, no problem arises. Families sare counted as having certain number of members; all families which have seven members also have at least six; all those with three also have at least two, and so on. This is obvious and in the nature of the ordinal character of measurement. It is clear that no man can save his second thousand dollars until the first is accumulated. This is because the amount of money in the bank has a quantitative character. Construction of scale and its application requires a high degree of consciousness of the logical problems surrounding the twin assumptions of (i) the existence of a continuum and (ii) the representativeness of that continuum achieved by the particular scale in mind. The question of representativeness is checked by tests of validity.

11.4 TYPES OF SCALES

Basically there are four types of scales found. They are discussed below:

- **Nominal scale**

  Nominal scales, also called categorical scales, measure categorical data. These scales are used for variables or indicators that have mutually exclusive attributes. For examples gender (two values: male or female), industry type (manufacturing, financial, agriculture, etc.), and religious affiliation (Christian, Muslim, Jew, etc.). Nominal scale is simply a system of assigning umber symbols to events in order to label them. Nominal scale provides convenient ways of keeping track of people, objects and events. One cannot do much with the numbers

involved. Nominal scales merely offer *names* or *labels* for different attribute values. The appropriate measure of central tendency of a nominal scale is mode, and neither the mean nor the median can be defined. Nominal terms have no corresponding reality, either in or outside our mind. These are mere words. Nominal scales are only in name. Numbers are assigned to them like labels. These provide identity. The identifying numbers can be switched around. They can be interchanged. This scale consists of two or more named categories into which individuals, objects or responses are classified.In a nominal scale, it is possible to distinguish two or more categories in relation to the specified attribute. Nominal data are marked by their equality. These cannot be placed in a meaningful order, or ranked as higher or lower with regard to each other, except according to some other yardstick. Nominal scale is the least powerful level of measurement. It indicates no order or distance relationship and has no arithmetic origin. It simply describe differences between things by assigning them to categories, for that reason nominal data is counted data. It is a simple method of classification rather than an arrangement along a continuum. The question of dimension is not important in this type of scale. If desired, the different groups may be numbered. Classification of individuals according to religion is an example of nominal scale. The categories of nominal scale must be related to each other. They must lie along the same dimension like landownership, gender, voting behaviour etc.

- **Ordinal Scale**

  Ordinal scale is the lowest level of the ordered scale that is commonly used. Ordinal scale places events in order, but there is no attempt to make the intervals of the scale equal in terms of some rule. Ordinal scales generally measures non-numeric concepts like happiness, satisfaction, discomfort etc. Ordinal scales are those that measure rank-order data, such as the ranking of students in a class as first, second, third, and so forth, based on their grade point average or test scores. However, the actual or relative values of attributes or

difference in attribute values cannot be assessed. Ordinal scale is easy to remember because it is hearing like order and that's the key to remember with ordinal scales- it is the order that matters. In addition to the qualities possessed by nominal scales, ordinal scales show transitivity. They possess asymmetrical order or succession. The use of an ordinal scale implies a statement of 'greater than' or 'less than' without our being able to state how much greater or less. The real difference between rank 1 and 2 may be more or less than the difference between rank 5 and 6. The central tendency measure of an ordinal scale can be its median or mode, and means are uninterruptable. Hence, statistical analyses may involve percentiles and non-parametric analysis, but more sophisticated techniques such as correlation, regression, and analysis of variance, are not appropriate. But, as like the nominal scale, ordinal scale does not measure the distance between the positions.

- **Interval Scale**

  In this type of scale the distance between two items is known. In interval scale the intervals are adjusted in terms of some rule that have been established as a basis for making the units equal. The units are equal only in so far as one accepts the assumptions on which the rule is based. The exact differences between two values are known in this scale. This scale has equal units of measurement. For that, it is possible to interpret not only the order of scale scores but also the distance between them.Interval scalesare those where the values measured are not only rank-ordered, but are also equidistant from adjacent attributes. For example, the temperature scale (in Fahrenheit or Celsius), where the difference between 30 and 40 degree Fahrenheit is the same as that between 80 and 90 degree Fahrenheit.. The intelligence quotient (IQ) scale is also an interval scale, because the scale is designed such that the difference between IQ scores 100 and 110 is supposed to be the same as between 110 and 120 (although we do not really know whether that is truly the case).Interval scale allows us to examine

"how much more" is one attribute when compared to another, which is not possible with nominal or ordinal scales. Allowed central tendency measures include mean, median, or mode, as are measures of dispersion, such as range and standard deviation. Interval scale provides more powerful measurement than ordinal scales because interval scale also incorporates the concept of equality of interval.

- **Ratio Scale**

  This scale incorporates the properties of an interval scale together with a fixed origin or zero point. Weight, length, and time are obvious examples. A ratio scale possesses a meaningful zero point.The zero point is not as precise as it was once believed to be. We can conceive of an absolute zero of length and similarly we can conceive of an absolute zero of time. This type of scale represents the actual amount of variables. Measures of physical dimensions such as weight, height, distance, etc. are examples. Researchers in physical science have the advantage to describing variables in ratio scale form but the behavioral sciences are generally limited to describe variables in interval scale form, a less precise type of measurement. This point exists at that point where the scale possesses no quality either this way or that. It is the top level of measurement and is not often available in social research. The factor which clearly defines a ratio scale is that it has a true zero point. A rational zero is a location on an interval scale deliberately chosen for reasons other than the current data. The distinctive feature of a ratio scale is that it has an origin defined by a dominating theory. These use of a ratio scale permits comparisons such as being twice as high, or one-half as much. Reaction time uses a ratio scale of measurement-time. Onthe basis of a ratio scale one can compare both differences in scores and the relative magnitude of scores. Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, and divided.

- **Attitude Scale:**

An attitude scale is a special type of questionnaire designed to produce scores indicating the intensity and direction of a person's feeling about an object or event. In this type of scale, the attitude of an individual towards a matter, thing, object or system can be known from the score of his responses given on a questionnaire. The score will place him/her on a scale. He simply expresses his like or dislike, agreement or disagreement with the issue involved, as given in the forms of questions. On the basis of his reply, he is assigned a score which indicates his position. Some relevant and indirect statements can also be used to reveal the attitude. In the attitude scale, most of the scales used are ordinal in nature, though there is attempt to treat the resulting data as internally scaled. This type of scales provide a quantitative measurement of attitudes, opinions or values by summarizing numerical scores given by researchers to people's responses to sets of statements exploring dimensions of an underlying theme.

There are various types of attitude scales. These are discussed below-

- **Point Scale:**

In this scale, at first, a crucial number of words of which the opinion are required is selected. The word may be prayer, dancing, birth control, smoking and so on. The respondent is to cross out every word that is more annoying than pleasing to him. One point is given to each agreement or disagreement, whichever is to be chosen. There is another variety of point scale. In such a scale two sets of words indicating both favourable and unfavourable opinions are given. The unfavourable items may be crossed or favourable item may be left uncrossed.

- **Thurstone Scale:**

Louis Thurstone, one of the earliest and most famous scaling theorists, published a method of equal-appearing intervals in 1925.

It is the most widely used method of scaling. Its procedure is used to develop scale items to measure attitudes. It is also known as equal-appearing scale, and judgment scale. It is a uni-dimensional scale, for it measures aggregate attitude towards a specific issue. A Thurston scale is an attitude scale consisting of items with which the respondent has either to agree or disagree. Only those items with which the respondent agrees are second. Each item has a value and the respondent's score on the scale corresponds to the median score of the items with which the respondent agree. In this scale a number of statements, whose position on the scale has been determined by judges, is used. The position is determined by the method of equal-appearing intervals.

Thurstone attempted to devise a method that would represent the attitudes of a group on specific issue in the form of a frequency distribution, the baseline indicating the whole range of attitude gradation from the most favourable at one end to the least favourable at the other with a neutral zero in between' (Thurstone L. L. and Chave E. J. 1929 cited in Young 1979: 352).

- **Summated (Likert) Scale:**

Likert scaling was developed by Rensis Likert in 1932. This is also referred to as the 'method of summated ratings' because a respondent's score is computed by summing the response values. It is like marks obtained by a student in a test. It is a uni-dimensional scaling method. It is the most widely used scaling technique to assess social attitudes. It is a very popular rating scale for measuring ordinal data in social science research. This scale includes Likert items that are simply-worded statements to which respondents can indicate their extent of agreement or disagreement on a five or seven-point scale ranging from "strongly disagree" to "strongly agree". Unlike the differential scale, the Likert scale uses only the definitely favourable and unfavourable statements. Likert scale excludes the intermediate opinions. This scale consists of a series of statements to which the respondent is to react. Likert scale needs a minimum of

two categories, such as 'agree' and 'disagree'. To know the degree of intensity it may further be divided into five or seven point response scale like; 'Strongly approve', 'approve', 'undecided', 'disapprove', 'strongly disapprove' etc. These five points will constitute the scale. Each point of the scale carries a score. "Strongly approve" is given the highest score (5 or +2) and "strongly disapprove" is given the least score (1 or -2). The score values are not indicated in the questionnaire. The following is a Likert type scale:

Statement 1: Prohibition should be made compulsory

1	2	3	4	5
Strongly approve	Approve	Undecided	Disapprove	strongly disapprove
(5 or +2)	(4 or +1)	(3 or 0)	(2 or -1)	(1 or -2)

A key characteristic of a Likert scale is that even though the statements vary in different items or indicators, the anchors ("strongly disagree" to "strongly agree") remain the same. Likert scales are ordinal scales because the anchors are not necessarily equidistant, even though sometimes we treat them like interval scales.Likert scale is considered more reliable because under it respondents answer each statement included in the instruments and it also provide more information and data. It takes less time to construct; it is frequently used by the students of opinion research.

- **Guttman Scale:**

  This scale is designed by Louis Guttman and it is also known as the cumulative scaling method. It is based on Emory Bogardus' social distance technique, which assumes that people's willingness to participate in social relations with other people vary in degrees of intensity. It measures that intensity by using a list of items arranged from "least intense" to "most intense". The idea is that people who agree with one item on this list also agree with all previous items. In practice, we seldom find a set of items that matches this cumulative

pattern perfectly. A scalogram analysis is used to examine how closely a set of items corresponds to the idea of cumulativeness. Guttman scale uses a series of items arranged in increasing order of intensity of the construct of interest, from least intense to most intense. The Guttman scale is based on the assumption that the various attitude statements in the scale belong to the same dimension. The attainment of a high degree of unidimensionality is the major concern of the Guttman scale. Guttman's scalogram collects response categories as yes/no, agree/disagree for each item. The affirmative responses are given 1 and 0 for negative response. The respondent's score is obtained by summing up the score for all the items. Guttman model is deterministic in nature. It assumes that a person, who responds positively/negatively to one item, must respond positively negatively to a series of others. The model can be made probabilistic rather than deterministic.The strength of the Guttman method is its capacity to identify more than one dimension in the scale. The coefficient of reproducibility is indicative of the extent to which the material relates to a single dimension. Construction of Guttman scale involves different phases. In the first phase, items in the form of questions are generated based on the concept under study. Each item must be given two options – in affirmative (agree) and negative (disagree). Items must cover the whole range of the underlying attitude continuum. In other words, items must move from simple agreement to overwhelming agreement. A group of judges rate the statements in terms of their favourableness or unfavourableness towards the concept. They then review the items and select final scale items.



CHECK YOUR PROGRESS

Q.1: a) From which language the word 'scale' originated?

b) Mention the different types of scale

.....

.....

.....

.....

.....

.....

11.5 SCALES TO ANALYSE SOCIAL DISTANCE

The concept of social distance refers to a continuum described by Robert E. Park as "the grades and degrees of understanding and intimacy which characterize pre-social and social relations generally." We can describe the continuum ranges from close, warm, and intimate contact on the one hand, through indifference, to active dislike, hostility, and rejection on the other hand. While applying the social distance scale two problems must be faced. First, it is necessary to define the remainder of the continuum, that is, to define the measurement as social distance toward something- a social group, a value, or an individual. The second problem, then, is to measure the social distance which does exist. This technique is quite flexible and can be adapted to variety of needs.

- **The Bogardus Social Distance Scale**

  The Bogardus Social Distance Scale is one of the oldest psychological attitude scales. This scale is used for measuring the social distance between nations, races, castes and so on. This scale is named after Emory S, Bogardus, who pioneered in this area. He and his students as well as other sociologists have applied this method to various types of social groups such ethnic minorities, social classes, and occupational types, as well as more abstract social values. A number of suggested relationships may be listed, to which members of an ethnic group may be admitted. The respondent is to indicate as to which racial group is to be admitted by him for each of the specified relationship. The attitude is measured by the closeness of relationship that a respondent is willing to accept or the social distance that he likes to maintain. The Social Distance Scale is an example of a Guttman scale in that it is unidimensional and cumulative. The unidimensional aspect means that the scale items

can be used to measure a single theoretical concept and only that concept. For example, in a scale composed of items that measure prejudice, items that measure a different concept would not be included. The items contained in a unidimensional scale can be placed on a continuum. Individual attitude is measured with the social distance indicated by them. The respondents are asked to agree or disagree with various items in it. The items are interconnected in such a way that if a person agrees with the fourth item, he is also likely to agree with the first three items in that regard. One who agrees with the eight items, he is also likely to agree with the first seven items.

In this Bogardus scale, the respondent has to indicate his first feeling. He has to give his reaction to each race as a group and he should not take into account any individual member of a group, best or worst, into account. The social distance can also be calculated mathematically. With the development of the Thurstone and Likert scaling methods, attention has been shifted away from cumulative scaling towards unidimensional scaling which seeks to avoid the influence of the external factors.

- **Sociometric Scale**

The term 'Sociometry' relates to its Latin etymology, 'socius' meaning companion, and 'metrum' meaning measure. Sociometry deals with social relations among groups and individuals. Sociometry is a "method for discovering, describing, and evaluating social status, structure and development through measuring the extent of acceptance or rejection between individuals in groups."

The term is closely associated with the work of the Austrian-born psychiatrist J.L. Moreno, who developed the method as a research and therapeutic technique. Sociometry has come to have several meanings; it is most commonly applied to the quantitative treatment of preferential interpersonal relations, but it is also used to mean the quantitative treatment of all kinds of interpersonal relations. Sociometric techniques have been widely utilized by sociologists, psychologists and psychiatrists in the study

of group structure, social status and personal traits. It helps to make more explicit and precise the configuration of group relationships, the characteristics and composition of cliques and other elements in a larger group. A Sociometric measure assesses the attractions within a given group. The basic technique involves asking all group members to identify specific persons within the group they would prefer (or would not prefer) to have as partners in a given activity. Many variations on this technique exist for studying different aspects of social preference. For example, a group's networking structure can be exposed through the Sociometric technique of recording all interactions between group members. Helen Jennings has defined Sociometry as "means of presenting simply and graphically the entire structure of relations existing at a given time among members of a given group." Generally Sociometry is concerned with the social interactions among social groups.

Sociometry is based on the principle of sociometric test. Moreno first applied "sociometric test" in his empirical study regarding the groupings in public school in USA. Sociometric method has been used quite extensively in social research by many writers like Moreno, Jennings, Festinger etc. In this test, every member of a group may be asked to select a specified number of people from within, who could be made his friend leader and colleague, and so on. Sociometry is fundamentally concerned with the social interactions among social groups. The interaction may be of any type. This method can provide useful data/information about the position of an individual in the group in relation to other groups and sub-groups. Interview and questionnaire are the two main methods for administering sociometric tests. In his test of girls training school, Moreno used this personal interview technique. In the interview each girl was questioned as to her attitude toward those choosing or rejecting her for each criterion. Did she like, dislike, or feel indifferent toward those indication attraction or repulsion toward her? Motivation for attraction and rejection was also investigated in the interview in order to obtain a clearer picture of each girl's position in the social structure. These are easy to administer and are suitable for different situations. Sociometry has been defined as "a method for discovering, describing, and

evaluating social status, structure, and development through measuring the extent of acceptance or rejection between individuals in groups." A careful sociometric study may prove very useful in revealing the social interrelations, the pattern of acceptance and rejection and the growth of leadership in a society. This method is appropriate for studying the areas having inter-disciplinary interests. The method can be utilized without a large scale investment of resources and hence, has special relevance in low-developed countries. In sociometric test it is difficult to secure reliable information. This problem is same as the problems occurring in case of attitude scales, case techniques and questionnaires. Sociometric techniques have to be used with discrimination and carefully adapted to each situation.



CHECK YOUR PROGRESS

Q.2: Who used the Social distance scale in Social science research? Discuss about social distance scale.

.....

.....

Q.3: Whose work is associated with the term sociometry? Discuss about sociometric test.

.....

.....

Q.4: Who developed the magic bullet theory?

.....

.....

Reliability of Scale:

There are various techniques available in social science research through which the reliability of scale is determined.

- *Retest method*. According to this method, the same scale is applied twice to the same population for the same objectives, and if the two results are similar, the scale is regarded as reliable.

- *Multiple form*. According to this technique, two types of scales are constructed, and they are applied to the same population. In case the two results are more or less similar, the scales may be regarded as reliable.
- *Split-Half Method*. In this method the scale may be divided into two equal parts. Then each part is taken as a whole and measurement is made separately. The correlation between the two scores obtained in this way is estimated and if the degree of correlation is high, the scale may be regarded as reliable.

11.6 IMPORTANCE OF SCALING

In our daily life we rely on different types of standardized tools of measurement. For example, a petrol dispensing machine in a petrol bunk measures the volume of petrol we ask for. A shop keeper uses a weighing machine to measure the required quantity. In the case of measuring temperature, a thermometer is used which is believed to be a standardized instrument. Similarly height, distance and a whole lot of other things which are part and parcel of our daily life are measured using standardized instruments. Scaling is useful in a number of ways. It improves objectivity. The matter under study can be expressed accurately. Scaling makes the matter concise. A lot of material is expressed with brief and to the point numbers.

Unlike the ones mentioned above, measuring social behaviour (attitudes, perceptions, preferences, or opinion) is difficult for its complexity. It is complex because social behaviour cannot be precisely explained or reduced to certain causal factors. For example, preferences of voters towards candidates in elections are conditioned by several factors like voters' awareness levels, socio-economic background of the voter, and profile of contesting candidates, parties and so on. Hence, it is difficult to single out a factor for preference of a particular candidate by a specific voter.

However, through an intensive and scientific exercise, researcher may be able to explain the attitude of voters in a particular constituency

towards contesting candidates. This is possible through a set of statements which is carefully worded and ordered in a scientific manner. Scaling, therefore, helps in assessing the attitudes or opinions of respondents with a higher degree of accuracy. Despite the use of Scaling techniques in psychological measurement and other areas of research; there also have some limitations of the application of scaling techniques.

11.7 DIFFICULTIES IN APPLICATION OF SCALING TECHNIQUES

Application of Scaling technique is not an easy task in social science research. As we know that human nature, behaviour, thinking is not same with each other. For that reason, it is not universally applicable. Human behaviour is heterogeneous and unpredictable. Social phenomena are complex and qualitative in nature. The concepts themselves are not defined properly. Thus quantitative measurement is a baffling affair. Social phenomena cannot be experimented in a controlled way. The phenomena may be individual, private or esoteric. Therefore, they cannot be properly watched and evaluated. Thus, scale cannot measure all the causative variables involved. We know that human behaviour is flexible and variations are there. Whatever attitude is expressed at a point of time may not be taken as reliable. Even a valid scale may not remain valid during the period of analysis.

In spite of these difficulties, the scaling techniques can be improved to measure the social phenomena. It is important for social science researches.



11.8 LET US SUM UP

- In this unit we first discussed the meaning of scale and scaling and also discussed some definitions of scaling.
- Scales are commonly used by people in their day to day activities.
- The word scale comes from the Latin word, ‘scala’, which means ladder or flight of steps.

- There are four types of scales namely; Nominal scale, Ordinal scale, Interval scale and Ratio scale.
- There are different attitude scales like; point scale, likert scale, Thurstone scale etc.



LET US KNOW

Scale: A scale represents a series of ordered steps at fixed intervals used as a standard of measurement. Scales are used to rank people's judgments of objects, events, or other people from low to high or from poor to good.

Sociometry: The term Sociometry relates to its Latin etymology, 'socius' means companion, and metrum meaning measure. Sociometry deals with social relations among groups and individuals. Sociometry is a "method for discovering, describing, and evaluating social status, structure and development through measuring the extent of acceptance or rejection between individuals in groups."

Social distance: This scale is made use for measuring the social distance between nations, races, castes and so on. This scale is named after Emory S, Bogardus, who pioneered in this area. He and his students as well as other sociologists have applied this method to various types of social groups such ethnic minorities, social classes, and occupational types, as well as more abstract social values.



11.9 FURTHER READING

1. Young, P. V. 1992. Scientific social surveys and research, New Delhi: Prentice Hall of India.
2. Goode, W. J and Hatt, P.K. 2006. Methods in Social Research. Delhi: Surajeet Publication.

3. Jahoda, Marie, Morton Deutsch, and Stuart W. Cook. 1951. Research Methods in Social Relations. New York: Dryden.
4. Thurstone, L. L, .and E. J. Chave. 1929. The Measurement of Attitudes Chicago: University of Chicago Press.



11.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: a) Latin word scala

Ans. to Q. No. 2: b) Point Scale, Thurstone scale, Likert scale, Guttman scale

Ans. to Q. No. 3: a)Emory S, Bogardus

Ans. to Q. No. 4: b) Moreno, Jennings, Festinger



11.11 MODEL QUESTIONS

A) Short Questions (Answer within 150 words)

Q. 1: Define scaling.

Q. 2: Write about the attitude scale.

B) Long Questions (Answer within 300-500 words)

Q. 1: Discuss the Bogardus and sociometric scale

Q. 2: Explain the importance of Scaling techniques? What are the limitations of using scaling method in Social Science Research?

*** ***** ***

UNIT 12: MEASURES OF CENTRAL TENDENCY

UNIT STRUCTURE

12.1 Learning Objective

12.2 Introduction and Meaning

12.3 Requisites of An Ideal Measure of Central Tendency

12.4 Arithmetic Mean

12.5 Median

12.6 Mode

12.7 Let Sum Up

12.8 Further Reading

12.9 Answer to Check Your Progress

12.10 Model Questions

12.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- describe the different measures of central tendency
- explain the different types of mean
- discuss the different methods of calculating mean
- learn what is mode and how to calculate mode
- define what is median and how to calculate median.
- explain the relation between mean, median and mode.

12.2 INTRODUCTION

One of the important objectives of statistical analysis is to determine various numerical measures which describe the inherent characteristics of a frequency distribution. One of such measures is the measures of central tendency which is used to explain data in brief. A measure of central tendency or measure of location is a single value which is used to represent an entire set of data.

12.3 REQUISITES OF AN IDEAL MEASURE OF CENTRAL TENDENCY

According to Prof. Yule, the following are the requisites of a good average or measure of central tendency.

- It should be rigidly defined, i.e, the definition should be clear and unambiguous so that it leads to one and only one interpretation by different persons.
- It should be clearly understandable and easy to calculate, such that it should be computed with sufficient ease and even a non-mathematical person should be able to understand it.
- It should be based on all the observations. In other words, while computing the average, the entire set of data should be used and there should not be any loss of information resulting from not using the available data.
- It should be suitable for further mathematical treatment. In other words, the average should possess some important and interesting mathematical properties so that its use in further statistical theory is enhanced.
- It should be affected as little as possible by fluctuations of sampling. This means that if we take independent random samples of the same size from a given population and compute the average for each of these samples the, for an ideal average, the value so obtained from different samples should not vary much from one another.
- It should not be affected much by extreme observations, i.e., very small or very large observations.

There are five averages or measure of central tendencies – mean, median, mode, geometric mean and harmonic mean. But the syllabus covers only the first three. Hence only mean, median and mode are discussed.

12.4 ARITHMATIC MEAN

In mathematics there are various kinds of mean among which the noted one is sample mean and population mean. For the data set, the

arithmetic mean is also called average. It is the sum of all scores divided by the number of the scores.

In general, if X1,X2,……..Xn are the given observations, then their arithmetic mean, usually denoted by

where “X is the sum of the observations and n is the total number of frequencies.

Calculation of arithmetic mean for Ungrouped data: the arithmetic mean for ungrouped data can be calculated in three ways – direct method, assumed mean method and step deviation method.

- Direct Method: In this method the sum of all the frequencies is divided by the total number of frequencies.

  Example 1: The mean of 5, 10, 15 and 20 is

  X (bar)=(5+10+15+20)/4

  = 50/4

  = 12.5

- Assumed mean method:

  This method involves the following steps:

  a. A number is assumed as mean from the given frequency distribution and denoted as A.

  b. Next the deviation from each frequency is calculated, which is denoted by d and the sum of this deviation, denoted by “d is calculated.

  c. Next the A.M. is calculated as follows:

  x(bar)=A+”d/N

  Where A= assumed mean; d= x- A, N= total no. of observations.

  c) Step Deviation Method:

In this method the deviation, d= x-A, is divided by a common factor ‘c’, such that d’= d/c = (x – A)/c is found out. The formula for calculating arithmetic mean by step deviation is:

x(bar)=A+(“d’/N) x c

where c = common factor

Where c= common factor

CALCULATION OF ARITHMETIC MEAN FOR GROUPED DATA:

Grouped data may be discrete or continuous. In each case, the methods of calculating mean are different.

- Discrete Series
- Direct Method:

In case of discrete series, frequency against each of the observation is multiplied by the value of the observation. Then the values, so obtained are summed up and divided by the total number of frequencies to get arithmetic mean. Symbolically,

x(bar)="fx/"f = "fx/N

" fx = sum of the product of the value of the variables and their respective frequencies

N = "f = sum of frequencies

- Assumed Mean Method:

In this method the following steps are involved:

a. Take any value from the series of variables X as assumed mean (A).

b. Take the deviation of the values of the variable from the assumed mean and denote it by d i.e. d=X-A

c. Multiply all these values of d by their respective frequencies.

The mean by assumed mean method may be found out by the following formula:

x(bar)=A+"fd/N

Where, A = Assumed mean

" fd = Sum of the product of deviation and frequency

N = "f = total frequency

- Step Deviation Method:

In this method the deviations are divided by a common factor, such that we have d'= d/c = (x – A)/c

One has to follow the following steps:

a. Assume any value of the variable as assumed mean.

b. Subtract the assumed mean from each value of the item and denote it by 'd'.

c. These deviations are to be divided by a common factor to get the step deviations, which is denoted by d'.

d. These deviations (d') are to be divided by their corresponding frequencies (f) such that we get fd' and then find out "fd'.

5. Apply the formula

x(bar)=A+("fd'/"f) x c

Where c=common factor

N= "f = total frequency

B. Continuous Series:

In case of continuous series, the process of calculating arithmetic mean is same as that in the case of discrete, the only difference being that one has to obtain the mid-points of the class interval. The class intervals may be exclusive, inclusive or of unequal sizes. In each case, the arithmetic mean is calculated in a similar way.

a) Direct Method:

The following steps may be followed in this method-

1. Calculate the mid value and denote it by m.
2. Multiply each of the mid value by their corresponding frequencies to get fm and then calculate "fm.
3. Calculate the arithmetic mean using the following formula–

x(bar)="fm/"f

Where, m= mid value " fm = sum of the product of mid values and their corresponding frequency " f= total frequency.

b) Assumed mean method:

The following steps are involved in this method –

1. Obtain the mid values (m) of each class interval.
2. Assume any value from the obtained set of mid values.

3. Calculate the deviations of the mid values from the assumed values and denote it by d.
4. Multiply each deviation (d) with their corresponding frequencies (f) to obtain fd, and also work out “fd.
5. Divide “fd by “f and add the dividend to the assumed mean to get the actual mean.

c) Step deviation method:

In this method deviations are divided by a common factor to simplify the calculations. The formula for calculating average is

x(bar)=A+(“fd’/”f)xc

where d’=d/c=(m-A)/c

A = assumed mean f=frequency

c = common factor N=”f=total frequency

c) Weighted Arithmetic Mean:

Till now the arithmetic mean that we have discussed has given equal importance/weightage to all the items. But it is not always that all items in a distribution are equally important such that equal weightage is given to them. For example, if we want to measure the change of in cost of living of a certain group of people, then simple mean of the prices of the commodities are not equally important. Food items such as wheat and rice are more important than other items. Again, in other items, medicines will be more important than cigarettes. Thus, we will need to assign weights to the items and hence calculate weighted arithmetic mean. Weighted arithmetic mean is the average when different items of a series are given different weights according to their relative importance weights are indicated by *w*.

Let $w_1, w_2, \ldots\ldots, w_n$ be the weights attached too various values $x_1, x_2, \ldots\ldots, x_n$ respectively. Then the weighted arithmetic mean, usually denoted by *xw*(bar) is given by

$$xw(bar) = (w_1x_1+w_2x_2+\ldots\ldots..w_nx_n)/w_1+w_2+\ldots\ldots..+w_n$$

$$= \text{“}wx/\text{”}w$$

Where, *w*=weight, x_w(bar) = weighted arithmetic mean

ARITHMETIC MEAN OF TWO OR MORE DISTRIBUTION:

If we know the sizes and means of two component series, then we can find the mean of the resultant series obtained on combining the given series.

If n1 and n2 are the sizes and x_1(bar),x_2(bar) are the respective means of two series then mean of the combined series of size n_1+n_2 will be given by–

$x(bar) = \{n_1x_1(bar) + n_2x_2(bar)/n_1+n_2\}/\{ n1+n2\}$

Combined Arithmetic Mean of three series will be

$x(bar) = \{n_1x_1(bar) + n_2x_2(bar) + n_3x_3(bar) /n_1+n_2\}/\{ n_1+n_2+n_3\}$

MERITS AND DEMERITS OF MEAN

Merits of Arithmetic mean:

1. It is rigidly defined.
2. It is easy to calculate and easy to understand.
3. It is based on all the observations.
4. It is suitable for further mathematical treatment. It possesses many important mathematical properties because of which it has very wide application in statistical theory.
5. Of all the averages, arithmetic mean is affected least by the fluctuations of sampling.

Demerits of Arithmetic Mean:

1. The Arithmetic Mean is very much affected by extreme values. Two or three vary large values of the variable may unduly affect the value of the arithmetic mean.
2. It cannot be calculated in case of qualitative data.
3. It cannot be determined in open end classes.
4. Sometimes arithmetic mean gives absurd and ridiculous results.
5. It cannot be determined by inspection nor can it be located graphically.

12.5 MEDIAN

In statistical language median is the number which specifically denotes the number which is placed in the middle of the data set. The

median value distinguishes the higher half value then the lower half value of a data sample. In statistics and in probability theory the median is the commonly used tools. Major merit associated with median set of data is that it is not compressed extremely by large or small values. In any statistical observation the median is the central observation figure and having a central breakdown figure of 50%.

CALCULATION OF MEDIAN FOR UNGROUPED DATA

In case of ungrouped data if the total number of items in the series is very less, then it is very easy to calculate median. First of all, values should be arranged in ascending or descending order. Then median can be found out simply by observing the data. Median is denoted by M.

For example, let us consider the values 25,38,32,46,35 arranging them ascending order we have 25,32,35,38,46. Here the median will be 35 because it is the middle most value that divides the series into two equal parts.

Now suppose there are two middle values as in the following series – 25,32,35,37,38,46. Here the two middle values will be 35 and 37. Here, the median will be calculated as follows:

M = (35+37)/2 = 72/2 =36.

On the basis of the above two examples we can summarize the formula for calculating median in the following points:

i. When N of the series is an odd number, median is estimated using the following formula after arranging the data in ascending or descending order.

M = size of $\{(N+1)/2\}^{th}$ item

ii. When N of the series is an even number, median is estimated using the following formula after arranging the data in ascending or descending order.

M = {size of $(N/2)^{th}$ item + size of $(N/2+1)^{th}$ item}/2

Median in case of discrete series

In case of discrete series we are to calculate cumulative frequencies to determine median. The following steps are involved in the calculation of median in discrete series.

i. Arrange the data in ascending or descending order.

ii. Find out the cumulative frequencies.

iii. Apply the formula and locate the Median at the size of items, in whose cumulative frequency the value of $(N/2+1)^{th}$ item lies.

Median in case of continuous series

While calculating median in the continuous series, we should first compute cumulative frequencies. After that median group is located by applying the formula, i.e., size of $(N/2)^{th}$ item. After ascertaining the group of the median the following formula will be applied –

Median = L+{(N/2 – c.f.)/f} x i

Here, L = lower limit of median class.

c.f. = cumulative frequency of the class just before the median class

f = frequency of the median class

I = the class interval of the median class

Steps involved in the calculation of median

i. The data are in ascending or descending order of their class interval.

ii. Cumulative frequencies are computed.

iii. Median class of the series is identified. It corresponds to the cumulative frequency which includes the $(N/2)^{th}$ item.

iv. Then the formula is applied to determine the actual median value.

MERITS AND DEMERITS OF MEDIAN

Merits:

1. It is rigidly defined.
2. Unlike arithmetic mean it is not affected by extreme values.
3. In case of open end classification, it is especially useful because only the position is to be ascertained.
4. Median can sometimes be located by simply inspection and can also be computed graphically.

5. Even if the value of the extremes is not known median can be calculated if the number of items is known.

Demerits:

1. Median indicates the value of the middle term, so, it is not based upon all observations.
2. Median cannot be put further algebraic treatment just as mean.
3. For the calculation of median, we are to arrange the data in ascending or descending order which makes calculation difficult.
4. Median is likely to be affected by the fluctuations of sampling.
5. Median may not be representative of a series in many cases.

12.6 MODE

In any set of numbers the most prevailing number is often referred as Mode. It is found by collecting and organizing data in order to count the frequency of each outcome. In statistical term it refers to the most occurring and frequent number in the set of numbers. It is generated by organizing data in an order that helps in counting the frequency of each result. The result with the highest number of occurrences is the mode of the set.

For example, if a car dealer says Etios is in fashion these days, it means that the sale of Etios is highest and maximum number of people drives Etios. As such Etios may be called the model car because its frequency is highest. Similarly eight students get the following marks in Economics – 38,42,38,38,54,46,42,35. Here the number 38 appears thrice. So the mode is 38.

Let us consider another set of marks – 42,55,68,36,52. Here all the marks appear only once. So mode cannot be determined. In case of discrete frequency distribution mode can be determined simply by observing the values.

Consider the following example

Variable 10	25	35	45	55	
Frequency	5	18	25	22	17

Here the highest frequency is 25 and the variable corresponding to 25 is 35. So, mode is 35. It may be pointed out that mode need not be a unique number like mean and median. Sometimes, one distribution may have more than two or more mode.

Determination of mode in case of continuous frequency distribution:

In case of continuous frequency distribution, the class with the highest frequency is identified as the model class. The formula for determining mode is as follows:

$M_0 = L + (f_1+f_2)/(2f_1 – f_0 – f_2) \times I$

Where,

L = Lower limit of model class

f_1 = frequency of the model class

f_2 = frequency of the class preceding the model class

f_0 = frequency of the class succeeding the model class

I = height of the model class

The following things need to be noted about mode:

i. If the frequency of each class is same, then the distribution will not have mode.

ii. In order to determine mode we need not determine mid value or prepare a table.

iii. If mid values are given we have to determine the classes first.

MERITS AND DEMERITS OF MODE:

Merits

i. Mode is easy to calculate as it does not need much mathematical calculation. In individual and discrete series it can be found out through inspection.

ii. It is not affected by extreme values.

iii. To calculate one need not know all the items and frequencies of the distribution. In simple series, it is enough to calculate mode if one knows the item with the highest frequency in the series.

iv. Since mode is the most common item in the series, it is not an isolated example like the median. It cannot be a value which is not found in the series.

v. Mode is the value which occurs most frequently in the series. So, mode is the best representative value of the series.

vi. Mode can be located graphically with the help of histogram.

vii. Mode can be used to describe quantitative as well as qualitative data.

Demerits

i. The value of mode is not based on each and every item of the series.

ii. The value of mode cannot always be determined. It is difficult to find modal class in case of bi – modal and multi modal distribution.

iii. Mode is affected by the fluctuations of sampling.

iv. It is not suitable for further mathematical treatment. From the modal values and the sizes of two or more series, the overall mode of the combined data cannot be determined.

v. Calculation f mode involves cumbersome procedure of grouping the data. Mode is not suitable when relative importance of items is desired.

Relation between Mean, Median and Mode:

i. Median always lies between arithmetic mean and mode.

That is, Arithmetic mean>Median>Mode

ii. In case of symmetrical frequency distribution, mean, median and mode will be identical.

iii. In case of asymmetrical frequency distribution, value of mean, median and mode will be different. Then mode can be determined from arithmetic average and median. In that case, the relationship between mean, median and mode is

Mode = 3 Median – 2Mean.

For example,

If in an asymmetrical distribution, median is 350 and mean is 400, then mode will be

Mode = 3x350 – 2x400

= 1050 – 800 = 250.



CHECK YOUR PROGRESS

Q.1: What are the different measure of cultral tendency?

.....

.....

Q.2: What are the different methods of calculating mean?

.....

.....

Q.3: What is mode?

.....

.....

Q.4: What is median?

.....

.....



12.7 LET US SUM UP

- A measure of central tendency or measure of location is a single value which is used to represent an entire set of data.
- Arithmetic of a given set of observations is their sum divided by the number of observations.
- Median is the middle value of the series when arranged either in ascending order or in descending order. It is the value in the series which divided the series into two equal parts; one part consists of the values equal to median or smaller than it and the other part having the value equal to the median or larger than it.

- Mode is the value which occur largest number of times in a series. According to Coxton and Cowden, “The mode of a distribution is the value at that point around which the items tend to be most heavily concentrated.”



12.8 FURTHER READING

1. Medhi, J. 2006. Statistical Methods : An Introductory text. New Age International (p) Ltd.



12.9 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: There are five averages or measure of central tendencies – men, median, mode, geometric mean and harmonic mean.

Ans. to Q. No. 2: (a) Direct method, (b) Assumed mean method (c) Step deviation method

Ans. to Q. No. 3: Another important measure of central tendency of a statistical series is the mode. It is the value which occur largest number of times in a series.

Ans. to Q. No. 4: Median is the middle value of the series when arranged either in ascending order or in descending order. It is the value in the series which divided the series into two equal parts; one part consists of the values equal to median or smaller than it and the other part having the value equal to the median or larger than it.



12.10 MODEL QUESTIONS

A) SHORT QUESTIONS: (Answer within 150 words)

Q. 1: What do mean by central tendency?

Q. 2: Write the formula for arithmetic mean.

Q. 3: Mention two advantages and two disadvantages of median.

Q. 4: Mention two advantages and two disadvantages of median.

Q. 5: Mention two advantages and two disadvantages of median.

B. LONG QUESTIONS: (Answer within 350-500 words)

Q. 1: What are the requisites of a good measure of central tendency?

Q. 2: What is an arithmetic mean? Write the different methods of calculating it.

Q. 3: What is an arithmetic median? Write the different methods of calculating it.

Q. 4: What is an arithmetic mode? Write the different methods of calculating it.

*** ***** ***

UNIT 13: MEASURES OF VARIABILITY

UNIT STRUCTURE

13.1 Learning Objectives

13.2 Introduction

13.3 Characteristics of a Good Measure of Dispersion

13.4 Objectives of Measuring Dispersion

13.5 Different Measures of Dispersion

- 13.5.1 Range
- 13.5.2 Quartile Deviation
- 13.5.3 Mean Deviation
- 13.5.4 Standard Deviation

13.6 Lets Sum Up

13.7 Further Reading

13.8 Answer To Check Your Progress

13.9 Model Questions

13.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- know about dispersion
- know about range
- know about quartile deviation
- know about mean deviation
- know about standard deviation
- recognise the difference between mean deviation and standard deviation
- know the different methods of calculating the various types of dispersion.

13.2 INTRODUCTION

Meaning and Concept

It is not always that measure of central tendency is able to describe a particular distribution completely. We require a further description of the series

and therefore the measures of central tendency must be supported and supplemented by some other measure and one such measure is measure of variability or dispersion. Dispersion or variability is a measure of the extent to which the individual items vary. In other words it gives an idea of compactness and scatteredness of the distribution. If we know the central tendency and dispersion of two or more distribution dispersion is found to be less than that in case of the other than the first distribution is said to be more consistent.

13.3 CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION

A good measure of dispersion should have the following characteristics –

1. It should be rigidly defined such that it is easily understandable.
2. It should be easy to calculate.
3. It should be based on all the observations.
4. It should not be affected by extreme values.
5. It should not be affected by the fluctuations in sampling.
6. It should be amenable to further mathematical treatment.

13.4 OBJECTIVES OF MEASURING DISPERSION

1. Dispersion helps to know the variation of different values of the items from average value of the series.
2. One can test the reliability of the average with dispersion.
3. The degree of variation between two or more series can be found out with dispersion.
4. Measures of variability are very useful for the calculation of skewness, kurtosis, correlation, regression, etc.

13.5 DIFFERENT MEASURES OF DISPERSION

The different measures of dispersion are:

1. Range
2. Average deviation

3. Standard deviation
4. Quartile deviation

13.5.1 Range

Range is the difference between the largest (L) and the smallest (S) values in a distribution. Thus

R=L-S

Where, L = largest value of the distribution

M = smallest value of the distribution

Dispersion will be higher with higher values of range.

Co –efficient of range

The co-efficient of range is obtained by dividing the range by the sum of extreme values.

Symbolically,

Co-efficient of range = (L-S)/(L+S)

Steps of calculating Range

1. Find the largest value of the series.
2. Find the smallest value of the series.
3. Apply the formula R=L-S

Calculation of range in individual series

Calculate range on the basis of the following data:

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks	48	52	35	76	81	39	67	45	32	93

Here, Largest value (L) = 93

Smallest value (S) = 32

Range = 93 – 32 =61

Calculation of range in discrete series

Marks	25	30	35	40	45
No. of students	7	3	9	11	13

Here, L = 13

S = 3

Therefore, R = 13 – 3 = 10

Co-efficient of range = (13 – 3) /(13+3)

= 10/26

= 0.26

Calculation of Range in continuous series

Class interval	Frequency
10 – 20	20
20 – 30	40
30 – 40	60
40 – 50	80
50 – 60	100

Here, L = Upper limit of the highest class = 60

S = Lower limit of the lowest class = 10

Therefore, R = 60 – 10 = 50

Merits of range:

1. It is easy to calculate and readily understandable.
2. It takes least time for computation.

Demerits of range:

1. Range considers only the extreme values, but tells nothing about the other values.
2. If in a set of observations, say the heights of children, the extreme values such as the tallest and the shortest remain unchanged and all the other values are changed, the range will remain unaffected.

3. Range is not based on all the observations.
4. Fluctuation in sampling greatly affects range. It varies from sample to sample.
5. Range cannot be used in case of open-end distribution. In open end distribution the lower limit of the lowest class and upper limit of the highest class are not known.

Uses of range:

1. In weather forecasting, where people are mostly interested in knowing the extreme values, range is used.
2. To know the variation in exchange rates, money value, gold prices and share values range is used.
3. Range is also used to study the variation in the quality of manufactured units.

13.5.2 Quartile Deviation

As range is affected by extreme values, a measure of distribution that is not affected by extreme values is required. Quartile deviation is a measure of distribution which divides the distribution into four equal parts, each containing 25% of the values.

With the help of upper and lower quartiles (i.e. Q3 and Q1, respectively), inter quartile range is calculated. Inter Quartile Range = Q3 – Q1.

Half of inter quartile range is Quartile Deviation. It is also known as Semi – Inter Quartile Range.

Quartile Deviation = Q.D. = (Q3 – Q1)/2

Example: Calculate the quartile deviation for the following data:

Roll No.	1	2	3	4	5	6	7
Marks	28	32	45	16	25	42	36

Solution:

First the marks are arranged in ascending order – 16,25,28,32, 36,42,45.

Now, we calculate Q3 and Q1.

Q1 = Size of $\{(N+1)/4\}^{th}$ item

= Size of $\{(7+1)/4\}^{th}$ item

= Size of 2^{nd} item

Therefore, Q1 = 25

Q3 = Size of 3{(N+1)/4}th item

= Size of $3\{(7+1)/4\}^{th}$ item

= Size of 6^{th} item

Therefore, Q3 = 42

Therefore, Q.D. = (42 – 25)/2

= 17/2

= 8.5

CALCULATION OF QUARTILE DEVIATION FOR DISCRETE SERIES:

EXAMPLE: Find quartile deviation for the following series:

Age	10	11	12	13	14	15	16
No. of participants	8	11	13	14	15	8	7

Solution:

In order to find Q.D. for discrete series we have to calculate cumulative frequency

Age	No. of participants	Cumulative frequency
10	8	8
11	9	17
12	11	28
13	12	40
14	13	53
15	6	59
16	1	60

Q1 = Size of {(N+1)/4}th item

= Size of {(60+1)/4}th item

= Size of 15.25th item

It will lie in the cumulative frequency 17. The value or age corresponding to the cumulative frequency 17 is 11

Therefore, Q1 = 17

Q3 = Size of 3{(N+1)/4}th item

= Size of 3{(60+1)/4}th item

= Size of 45.75th item

It will lie in the cumulative frequency 53. The value or age corresponding to this cumulative frequency is 14.

Therefore, Q3 =

Therefore, Q.D. = (14 – 11)/2

= 3/2

= 1.5

Calculation of Quartile Deviation for continuous series:

Class	10-20	20-30	30-40	40-50	50-60
Frequency	20	36	48	34	28

We first calculate the cumulative frequency:

Class	Frequency	Cumulative frequency
10-20	20	20
20-30	36	56
30-40	48	114
40-50	34	148
50-60	22	170

Q1 = Size of N^{th} /4 item

= Size of 170^{th} /4 item

= Size of 42.5^{th} item

The class containing 42.5^{th} item is 20-30. Now to calculate exact value of Q3 the following formula is used:

Q1 = L+{(N/4 – fc)/f} x i

Where, L = lower limit of the relevant quartile class

fc = value of the c.f. of the class preceding the quartile class

i = interval of the quartile class

f = frequency of the quartile class

Here, L = 20

N/4 = 42.5

fc = 20

i= 10

f=20

Therefore, Q1 = 20 +{(42.5 – 20)/20} x 10

= 20 + {(22.5)/20}x10

= 20 + 11.2

= 31.2

Q3 = size of $3N^{th}$/4 item

= Size of {3x170/4} item

= Size of 510/4 item

= Size of 127.2^{th} item

The class containing 127.2^{th} item is 40-50. Now the following formula is used for Q3.

Q3 = L + {3N/4 – fc}/f x i

Here, L, N/4,fc,f and I denote the same items as in Q1.

Therefore, Q3 = 40 + {127.2 – 114}/34 x 10

= 40 + 113.2/34 x 10

= 40 + 33.02

= 73.02

Therefore, Q.D. = ½(Q3 – Q1)

= ½(73.02 – 31.2)

= 20.91

Merits of quartile deviation:

i. Quartile deviation is based on middle 50% of the observations; hence it is a better measure than dispersion.

ii. It is easy to calculate.

iii. It is easy to understand.

iv. Quartile deviation can be calculated for both open end and close end distributions.

v. It is not unduly affected by extreme values.

Demerits of quartile deviation:

i. It ignores the first 25% and the last 25% of the observations.

ii. Fluctuations in sampling affect quartile deviation. A change in the value of a single item, may, in many cases, affect its value considerably.

iii. It is not capable for further algebraic treatment.

13.5.3 Mean Deviation

Mean deviation of a series is the arithmetic average of the deviations of various items from a measure of central tendency (mean, median or mode).

Theoretically, deviations can be taken from any of the three averages mentioned above but in actual practice mean deviation is calculated either from mean or median. Mean deviation is generally not calculated from Mode because its value is indeterminate and it gives erroneous conclusion. Between mean and median the deviations from median is less than deviations from mean and hence mean deviation from median is better than mean deviation from mean. The value of the mean deviation calculated from the median, is, therefore, always less than the value calculated from mean.

Mean deviation for ungrouped data:

The following steps are involved in the computation of mean deviation for ungrouped data:

i. Compute the mean or median of the series.

ii. Take deviations d=X-A from either mean or median ignoring ± sign.

iii. Obtain total of these deviations " |d | where, |d |= absolute value of the deviation

iv. Divide the above total i.e. in step (iii) by the number of items i.e. M.D. = " |d |/N. If the variable x takes the value $x_1,x_2,....x_n$ then

Mean deviation = $\{|x_1 - A|+|x_2 - A|+.....+|x_n - A|\}/n$

= " |d |/n

Where A = mean, median or mode of the x

d = x – A

Mean deviation for grouped data in discrete series

The following steps are involved in the computation of mean deviation for grouped data in discrete series:

i. Calculate the specific average from which mean deviation is to be found

ii. Obtain the absolute (positive) deviations (D) of each observation from the specific average.

iii. Absolute deviations (D) are multiplied with the respective frequencies (f) and the products are added up to obtain f |D| of each observation from specific average.

iv. Divide the total (f |D|) by the number of items to get mean deviation.

If the variable x takes the values x1,x2,……,xn with respective frequencies f1,f2,f3,…..fn.

Then,

Mean Deviation = $\{f_1|x_1 - A|+f2|x_2 - A|+....f_n|x_n|\}/N$

Where N = f

Mean deviation for grouped data in continuous series

In case of continuous, the formula for man deviation are same as that for a discrete series for the given continuous frequency distribution, the mid – points of class intervals have to be found out and they are taken as *m.* In this way the continuous series assumes the shape of discrete series. After that all the steps of discrete series are applied.

Symbolically,

Mean deviation from mean or median = { f|m - A|}/N

= { f|d|}/N

Merits of Mean Deviation:

i. It is rigidly defined.

ii. It is easy to understand and easy to calculate.

iii. It is based on all the observations of the series.

iv. As compared with the standard deviation, it is less affected by extreme values.

v. Mean deviation is based on the deviation from an average. So, it provides a better measure for comparison about formation of different distribution.

Demerits of Mean Deviation:

i. Mean deviation ignores the signs of deviations. As a result this method cannot be used for further algebraic treatment.

ii. Mean deviation cannot be computed for distribution with pen-end classes.

iii. Mean deviation when calculated from mode is not reliable because in many cases mode has no fixed value.

iv. It is difficult to compute when mean, median and mode are in fraction.

13.5.4 Standard Deviation

The most commonly used measure of dispersion is standard deviation and was first introduced by Karl Pearson in the year 1983.

One weakness of mean deviation is that in its calculation plus and minus signs are ignored. This drawback is removed in the calculation of standard deviation. One of the earliest way of doing away with the algebraic signs is to square the figures and this process is adopted in the calculation of standard deviation. “Standard Deviation is the square root of the arithmetic average of the squares of the deviation measured from the mean. This is generally denoted by sigma of the Greek language.” (Choudhury, et al,2013).

Standard deviation of ungrouped data:

In case of ungrouped data, standard deviation can be calculated by any of the four methods – Actual mean method, Assumed mean method, Direct method, Step Deviation method. The following are the formulas for each method.

i. Actual Mean Method ó= “(“[x-x(bar)]^2)/n)

= “(“d^2/n)

Where, n= number of items in the series

ó=standard deviation

ii. Assumed mean method

ó = “{(“d^2/n)-(“d/n)^2}

iii. Direct Method

ó=”{“x^2/n – (x bar)^2}

iv. Step Deviation Method

ó= “{(“d^2/n) × c}

Standard deviation of grouped data

In case of grouped data, standard deviation can be calculated by any of the two methods – Actual mean method and Assumed mean method. The following are the formulas for each method.

i. Actual Mean Method

In order to calculate standard deviation by this method we should proceed in the following way –

a. Calculate the mean of the distribution

   X bar = “fx/”f

   If it is a continuous series, then x bar =”fm/”f

b. Then the deviations of the values from the mean shuld be taken

   d=x-x bar

In case of continuous series, d = m – x bar should be determined

c. Then multiply the deviations with their corresponding frequencies to get fd values.

d. Calculate fd^2 by multiplying fd values with d values. Sum up them to get fd^2. Of course if we calculate d^2 first and multiply it with f then also we get the same value of fd^2

e. Apply the formula:

   Ó= fd^2/n

ii. Assumed mean method:

In order to compute standard deviation by assumed mean method we should proceed in the following way –

a. If it is a discrete series then calculate deviation of the values of the variable from the assumed mean, i.e., d=x-A should be determined.

   If it is a continuous series, calculate mid points f the class first. Then deviation of mid points from the assumed mean, i.e., d=m-A should be determined.

b. Multiple values of d with corresponding frequencies to get fd values and find fd.

c. Multiply fd by d to get fd^2 and find fd^2.

d. Then calculate standard deviation using the following formula:

   ó= $\{ fd^2/n – (fd/n)^2\}$

   where n= f

VARIANCE: It is the square of standard deviation. It is the measure of variability like standard deviation it is denoted by $ó^2$. Sometimes it is written as Var (x).

Merits of Standard Deviation:

i. It is rigidly defined.

ii. It depends on all the values of the variable.

iii. Fluctuations in sampling do not influence it much.

iv. Combined standard deviation of two or more groups can be computed.

v. It is capable of further algebraic treatment.

Demerits of Standard Deviation:

i. It is not easy to calculate standard deviation.

ii. It lays greater stress on extreme values and lesser stress n those values which are nearer to them.

Distinction between mean deviation and standard deviation:

i. While calculating mean deviation algebraic signs (i.e. ±) are not taken into account whereas while calculating standard deviation their signs are taken into account.

ii. Mean deviation can be computed from median, mode or mean and its value differs in these cases but the standard deviation is always calculated from the arithmetic mean.

iii. The mathematical properties possessed by standard deviations are far greater than those possessed by mean deviation.



CHECK YOUR PROGRESS

Q.1: What is dispension?

.....

.....

Q.2: How to calculate range?

.....

.....



13.6 LET US SUM UP

- Dispersion is a measure of the extent to which individual items vary.
- The different measures of dispersion are – range, standard deviation, mean deviation and quartile deviation.
- Range = L – S

  Where, L = largest value of the distribution

  S = smallest value of the distribution
- Quartile Deviation = Q.D. = (Q3 – Q1)/2
- Mean deviation = " |d |/n

  Where, d = x – A

  A = mean, median, mode
- Standard deviation can be calculated by different methods according to the type of data.



13.7 FURTHER READING

1. Gupta S.P. 2010. Statistical Methods. New Delhi: Sultan Chand & Sons.



13.8 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: Dispersion or variability is a measure of the extent to which the individual items vary. In other words it gives an idea of compactness and scatterendness of the distribution.

Ans. to Q. No. 2: Range is the different between the largest (L) and the smallest (S) values in a distribution. Thus

R = L - S

Where, L = largest value of the distribution

M = smallest value of the distribution.



13.9 MODEL QUESTIONS

A) SHORT QUESTIONS (Answer within 150 words)

Q. 1: What do you mean by dispersion?

Q. 2: What is range? How is it measured?

Q. 3: What is mean deviation?

Q. 4: What is standard deviation?

Q. 5: What is quartile deviation?

Q. 6: What are the characteristics of a good measure of dispersion?

Q. 7: What are the objectives of dispersion?

Q. 8: What is the difference between mean deviation and standard deviation?

B. LONG QUESTIONS (Answer within 350-500 words)

Q. 1: What do you mean by dispersion? What are the different measures of dispersion?

Q. 2: What is range? Discuss its merits and demerits.

Q. 3: What is mean deviation? Discuss its merits and demerits.

Q. 4: What is standard deviation? Discuss its merits and demerits.

Q. 5: What is quartile deviation? Discuss its merits and demerits.

*** ***** ***