UNIT 11 INTRODUCTION TO STATISTICS

Structure

11.1 INTRODUCTION

As you know every subject has its origin, development stages, scope, uses and limitations.

In this unit, we will discuss origin and development, definition, scope and uses, and limitations of statistics. Different measurement scales and different types of data also have been discussed in this unit.

Objectives

After completing this unit, you should be able to:

- know origin and development stages of statistics;
- know definition, scope, uses and limitations of statistics;
- get an idea of different types of measurement scales; and
- get an idea of different types of data.

11.2 ORIGIN AND DEVELOPMENT OF STATISTICS

As we know that every subject, race, machine, etc have its own origin and development stages. Similarly, Statistics also have its own origin and development stages. In fact, in a single sentence it can be said that human society has been using Statistics knowingly or unknowingly right from the beginning of its existence. This is because (i) most of the decisions taken by a human being are based on the past experience (i.e. based on statistical data he/she has experienced actually), and (ii) future events are also predicted by using/examining the past behavior of that particular event.

The word "Statistics' seems to have been derived from Latin word 'Status' or Italian word 'Statista' or German word 'Statistik'. But according to the observations of great John Graunt (1620-1674), the word 'Statistics' is of Italian origin and it is derived from the word 'Stato' and statista means a person who deals with affairs of the state. That is, initially kings or monarchs or governments used it to collect the information related to the population, agricultural land, wealth, etc. of the state. Their aim behind it was just to get an

idea about the men power of the state, force needed for the purpose of a war and necessary taxes to be impose to meet the financial need of the state.

State Craft: The art of managing state affairs.

So, it indicates that initially it was used by kings or monarchs or governments for administrative requirements of the state. That is why its origin lies in the state craft.

On the basis of evidences form papyrus manuscripts and ancient monuments in pharaonic temples, it is assumed that first census in the world was carried out in Egypt in 3050 BC. Yet, China's census data around 2000 BC is considered as the oldest surviving census data in the world.

Statistics in India

Though the use of Statistics, knowingly or unknowingly, has been there in India from ancient times, yet if we talk about its origin on the basis of evidences, it takes us in 3rd century BC when "Arthashastra" came into existence written by one of the greatest geniuses of political administration, Kautilya. In it, he had described the details related to conduct of population, agriculture and economic census. An efficient system of collecting official and administrative statistics was in use during the reign of Chandra Gupta Maurya (324-300 BC) under the guidance of Kautilya. Many things like taxation policy of the state, governance and administration, public finance, duties of a king, etc. had also been discussed in this celebrated Arthashastra. Another evidence that statistics was in use during Emperor Akbar's empire (1556-1605) is in the form of "Ain-I-Akbari" written by Abul Fazl, one of the nine jems of Akbar. Raja Todar Mal, Akbar's finance minister and another one of the nine jems of Akbar, used to keep very good records of land and revenue and he developed a very systematic revenue collections system in the kingdom of Akbar by using his expertise and the recorded data. Revenue collection system developed by Raja Todar Mal was so systematic that it became a model for future Mughals and later on for British.

British Government, after transfer of the power from East India Company to it, started a publication entitled 'Statistical Abstract of British India' as a regular annual feature in 1868 in which all the useful statistical information related to local administrations to all the British Provinces was provided. In between some census reports were coming on based on a particular area, but not at the national level. The first attempt to get detailed information on the whole population of India was made between 1867 and 1872. First decennial census was undertaken on 17th February 1881 by W.W. Plowden, first census commissioner of India. After that a census has been carried out over a period of 10 years in India. 2011 census was the 15th census in India.

Credit of establishing Statistics as a discipline in India goes to Prasanta Chandra Mahalanobis (P.C. Mahalanobis). He was a professor of physics in the Presidency College in Kolkata. During his study at Cambridge he got a chance to go through the work of Karl Pearson and R. A. Fisher. Continuing his interest in Statistics, he established a Statistical laboratory in the Presidency College Kolkata. On 17 December 1931, this statistical laboratory was given the name Indian Statistical Institute (ISI) mainly to promote the study, dissemination of knowledge of Statistics, research in it and to develop statistical techniques which play major role in addressing various problems of planning of national development and social welfare in the country. P.C. Mahalanobis was the founder director of ISI. Some well known personalities other than P.C. Mahalanobis associated to ISI whose research work made the

institute unique internationally are Professor C.R. Rao, Professor R.C. Bose, Professor S.N. Roy, etc.

First post graduate course in Statistics was started by Kolkata University in 1941, while first under graduate course in Statistics was started by the Presidency College Kolkata. With the passage of time some more universities/institutes came up with courses in Statistics. Some of these are University of Mumbai, University of Pune, University of Madras, University of Mysore, University of Kerala and University of Lucknow. This list of institutions went on increasing with time and at present more than 1100 institutes are there in the country, which are offering under graduate or post graduate courses in statistics.

Statistics in World

Without going into more details, we will concentrate on only some major discoveries in the area of Statistics at international level. A lot of theoretical development in different areas of statistics took place in seventeenth and eighteenth centuries in many countries of the world. John Graunt (1620-1674 born in London), being a haberdasher by profession, has the credit of producing the first life table with probabilities of survival to each age. Due to this great achievement, he is known as father of vital statistics. This was the period when some other persons also did their contribution in the same area such as Edmund Haller (1656-1742) prepared a life table on the basis of the data collected by Casper Newman in 1691, relating to death records of Breslau. Sir William Petty (1623-1687) also prepared mortality tables and calculated expectation of life at different ages. G.F. Knapp (1842-1926) and W. lexis (1837-1914) also did valuable work on the statistics of mortality. Study of probability was also found to be very important in the area of Statistics, quantitative measure of which was given by Galileo (1564-1642), an Italian mathematician[For detail discussion on development of Probability Sec 1.1 of Unit 1 of MST-003 may be referred to]. Guass (1777-1855) gave the principal of least square and normal law of errors. J. Bernoulli (1654-1705) was the first person who states the law of large numbers in his great work Ars conjectandi published eight years after his death. Statistical methods in the field of biometry were first introduced of Sir Francis Galton (1822-1911). Later on Professor Karl Pearson (1857-1936) followed up the work of Galton and did significant contribution to Anthropology and correlation coefficient theories. Karl Pearson was also the founder of Statistical Research Laboratory in the university college, London in 1911. Credit of discovery of Chi-Square test also goes to Karl Pearson. Credit of discovery of ‘t test’ or ‘student t’ test goes to W.S. Gosset who wrote under the pseudonym of student’s ‘t’.

List of contributors in the area of Statistics did not end here but we conclude by throwing some light on the work done by Fisher. Credit of discovering of very powerful test known as Analysis of Variance (ANOVA) goes to Sir Ronald A. Fisher (1890-1962). Fisher also did a lot of work in the area of point estimation. Due to his remarkable contribution in the field of Statistics, he is known as Father of Statistics.

11.3 DEFINITION OF STATISTICS

In previous section of this unit, we have seen that statistics is a very old science and it has developed/grown up through ages. So, it is not surprising that through its long journey its definitions given by different authors time to time

may also vary. In fact, today statistics is quite different that of earlier times. Let us first see some definitions by different authors and then it will be clear that which definition is more broad and may be consider as a good definition of statistics.

- "Statistics is the science of counting". – **A.L. Bowly.**
- "Statistics is the science of average." – **A.L. Bowly.**
- Statistics is "The science of the measurement of the social organism, regarded as a whole, in all its manifestations." – **A.L. Bowly.**
- "Statistics are the numerical statements of facts in any department of enquiry placed in relation to each other." – **A.L. Bowly**.
- "By statistics we mean quantitative data affected to a marked extent by multiplicity of causes." – **Yule and Kendall.**
- "Science of estimates and probabilities." – **Boddington.**
- "The method of judging collective natural or social phenomena from the results obtained by the analysis of an enumeration or collection of estimaties." – **W.I. King.**
- "Statistics is the science which deals with collection classification and tabulation of numerical facts as the basis for explanation description and comparison of phenomenon". – **Lovitt.**
- "The science which deals with the collection, tabulation, analysis and interpretation of numerical data." – **Croxton and Cowden.**

From the list of above definitions given by different authors, and various other definitions the comprehensive definition of Statistics may be given as:

"Statistics is a branch of science which deals with collection, classification, tabulation, analysis and interpretation of data."

11.4 SCOPE AND USES OF STATISTICS

In present times, statistics is not known as only collection of data, but it is regarded as a science having sound techniques of even handling huge data and providing valuable conclusions. Today statistical methods are universally applied. That is, statistics find its application in almost every sphere of human activity such as economics, commerce, management, information technology, education, planning, banking, insurance sector, medical science, biology, industrial, agriculture, market research, etc. That is, scope of statistics is very wide and in single sentence we can say that statistics is the queen of all sciences.

In the above paragraph, we have listed many fields where statistics has its application. Let us see its uses in some of these fields.

Statistics and Industry

In industrial line, statistics plays an important role such as in quality control and production engineering various control charts are used to maintain a certain quality level and different inspection plans are used in production engineering. Even to find average life of some products such as electric bulb, sampling technique is used. Those learners who will opt Industrial Statistics specialisation will get these terms in detail in courses MSTE-001 and MSTE-002.

Biology and Statistics

Professor Karl Pearson has stated that the whole doctrine of heredity rests on statistical basis. This is generally said that height of the child is associated with the height of the father. To test this type of hypothesis, statistics is the only science which provides the scientific methods. Vital statistics is totally devoted to the different aspects of human life like average life of men and women, birth and death rates, etc. Those learners who will opt Bio-Statistics specialisation will get these terms in detail in courses MSTE-003 and MSTE-004.

Statistics and Medicine

Statistics also plays an important role in the field of medicine. The hypothesis of the types:

(i) Drug A is better than drug B.

(ii) Smoking and cancer are associated.

(iii) Smoking and TB are associated.

All are tested using t-test or χ^2-test as the case may be. Statistics also find its application in clinical trials.

Statistics and Planning

Every institution/organisation plans for its future targets. Now, a days for a good planning, it has became necessary to analysis the statistical data according to the field of interest such as availability of raw material, consumption, investment, resources available, income, expenditure, quality needed, etc. In order to analysis these types of data, one has to totally depend on the statistical techniques. Thus statistics is essential for planning.

Statistics and Commerce

In present times, there is a very tough competition in almost every business. Also fashions, likings/tastes, requirements, trends, levels of qualities, technologies, etc. are changing very fast. So for the success of the business, it has become necessary for a business man to know the coming trend of market in advance or as soon as possible. This can also be achieved only with the help of market survey, which requires statistical techniques.

Statistics and Agriculture

Presently there are a number of varieties of seeds for a particular crop. Also different types of fertilizers are available in the market. For a good yield, it has become necessary to know that which one is better. This job is again done by a very popular and widely used test known as Analysis of variance (ANOVA) discovered by Professor R.A. Fisher. You will learn about ANOVA in more detail in block 2 of course MST-005. Complete Block 2 of MST-005 is totally devoted to ANOVA.

Statistics and Insurance Sector

Whole insurance sector totally depends on the statistical data and different concepts of probability theory. Life tables lies in the heart of human insurances. Curtate future life time and complete future life time, of a life are calculated using concept of random variables and their expected values. (you will learn in detail about random variables and their expected values in block 2 of course MST-003). Due to the large use of statistics in insurance sector, a new branch of statistics known as Actuarial statistics has been started in some institutes throughout the world.

Statistics and Research

Research is very important aspect in every discipline. In many disciplines such as psychology, tourism, education, M.B.A., etc. one has to collect the data on the characteristics of interest under study. Now a very important question arises, related to the measurement of scale to be used and appropriate test to be used. This requires the knowledge of different types of measurement scales and accordingly suitable statistical tool need. (Different types of measurement scales-nominal, ordinal, interval and ratio have been discussed in detail in Sec 11.6 of this unit). Also appropriate statistical tool to be needed in a given situation have been listed in Table 11.1.

Statistics and Economics

In order to know about the development of a country, it has become necessary to obtain the data related to its economical growth. Again, statistical tools are needed to collect relevant data (such as related to agricultural, industrial, literacy, etc) and for its analysis.

Statistics and Common Man

Statistics also plays an important role in the welfare of common man. Common man of any country faces lot of problems in his routine life such as food shortage, hygienic drinking water, unemployment, poverty, medical, shortage of public transport, etc. Time to time statistical figures on these issues enables the government to think and sort out these problems.

Statistics helps the common man in their day to day life in another way also, e.g. in purchasing any good he/she used his/her past experience (actually based on the data he/she faced/experienced) and take the decision to buy or not buy a particular object. Similarly, a farmer decide about the crop to be yield based on his past experience (actually based on the data he has faced) and labourer choose one of the works which gives him more wages based on his past experienced (actually based on the data he has faced).

List of fields/areas where statistics is used does not end here. We have just touch some of the areas where statistics has its application. We close this section by saying that there is hardly any field where statistics cannot be used. Infact, statistics can be used in any field of interest.

11.5 LIMITATIONS OF STATISTICS

In previous section of this unit, you have seen wide range of application of statistics. Being the queen of all sciences, statistics also have its own limitations, some of them are described as follows:

(1) Indirect Approach Towards Qualitative Characteristic

Science of statistics basically deals with numerical data. Therefore statistical tools are applicable only for quantitative measures. But many a times characteristic under study is qualitative in nature such as honesty, beauty, intelligence, boldness, drinking, smoking, etc. So any statistical tool cannot be directly applied on these types of characteristics. However, study of these types of characteristics can be made possible by first converting the characteristic under study into numerical figures based on some uniform criteria. For example, intelligence can be converted into numerical figures with the help of the marks obtained by the individuals in a common test.

(2) Dealingness with a Group

Science of statistics deals with aggregates of objects not with individuals. The individual's figures, when taken separately do not come under the category of statistical data. So, applicability of any statistical tool becomes meaningless. For example, salary of one employee of an institute does give any message related to the salaries of the employees of that particular institution.

(3) Lack of Exactness

Statistical results are not exactly true, but they are true on an average.

For example,

(i) If a statistical report says that 70% population of India lived in rural area. It does not imply that if you visit at public place like bus stand, railway station, etc. and asked the people about their living place. Results may surprise you and may highly differ with the above figure. But you may note that as sample size increases, the result will also come nearer and nearer to exact figure 70%.

(ii) Consider another example, suppose past data show that 90% operations of a doctor are successful. It does not imply that out of the next 100 operations, exactly 90 will be successful. It may happen that figure that will obtain in future may be 90%, 80%, 87%, 95%, etc. But there are sciences like mathematics where exactness is maintained. For example, if a book seller get 5% profit on selling a particular book. Then it is sure that if sell of that particular book is of Rs 200 he/she will get Rs 10 as profit and in case of sale of Rs 300 profit will be Rs 15 and so on.

(4) Requirement of Experts Hands for Effective Use

Requirement of experts' hands for effective and appropriate use is one of the main draw backs of the science of statistics. There are many statistical tools of similar type.

For example,

(i) To find average in a particular situation, which of the possible tools likes mean, median, mode, geometric mean, harmonic mean, etc. is appropriate needs the hands of experts.

(ii) Similarly to test a given statistical hypothesis which of the possible tools like Z-test, t-test, χ^2-test, F-test, ANOVA, median test, run test, sign test, etc. is appropriate again needs the hands of experts. This limitation of the statistics limits the range of its effective users.

11.6 MEASUREMENT SCALES

Two words "counting" and "measurement" are very frequently used by everybody. For example, if you want to know the number of pages in a note book, you can easily count them. Also, if you want to know the height of a man, you can easily measure it. But, in Statistics, act of counting and measurement is divided into 4 levels of measurement scales known as

(1) Nominal Scale

(2) Ordinal Scale

(3) Interval Scale

(4) Ratio Scale

Let us discuss these scales of measurement one by one:

(1) Nominal Scale

In Latin, 'Nomen' means name. The word nominal has come from this Latin word, i.e. 'Nomen'. Therefore, under nominal scale we divide the objects under study into two or more categories by giving them unique names. The classification of objects into atleast two or more categories is done in such a way that

(a) Each object takes place only in one category, i.e. each object falls in a unique category, i.e. it either belongs to a category or not. Mathematically, we may use the symbol ("=", "≠") if an object falls in a category or not.

(b) Number of categories must be sufficient to include all objects, i.e. there should not be scope for missing even a single object which does not fall in any of the categories. That is, in statistical language categories must be mutually exclusive and exhaustive.

Generally nominal scale is used when we want to categories the data based on the characteristic such as gender, race, region, religion, etc.

To get more familiar with the idea of nominal scale, let us consider some examples:

(i) Classification into Different Categories Based on Gender

This can be done by dividing the population into two categories male 'M' and female 'F'

Category	Name/Code
Male	M
Female	F

Here we have named male as 'M' and female as 'F'. This is not the only way, we can also code male by '0' and female by '1' or we may use any other convenient symbols. So, we note that main thing is that we have to give a unique name to each category.

(ii) Classification into Different Categories Based on Caste

Different categories	Code alloted /Name given
General	Gen
Scheduled caste	SC
Scheduled tribes	ST
Backward class	BC
Others	'O'

Here also we can give a code to general, scheduled caste, scheduled tribes, backward class and other categories by '0', '1', '2', '3', '4' respectively.

(iii) Classification into Different Categories Based on Region

28 states and 7 union territories together classified India into 35 categories which can be coded by their usual names or may be coded by using some other symbols.

(iv) Classification into Different Categories Based on Religion

Population of India can be broadly categorised based on the following different religions:

Different categories	Codes allowed /Names given
Hindu	1
Muslim	2
Sikh	3
Isaiah	4
Others	5

(v) Classification into Different Categories Based on Number Allotted

In a sport event, the numbers allotted to the participants also come under nominal scale.

Note 1: We note that in nominal scale we have just coded the objects. Sign of less than or greater than does not make any sense in nominal scale. That is here we have coded Hindu, Muslim, by '1' and '2' respectively. But Hindu > Muslim or Muslim > Hindu does not make any sense.

Similarly, male > female or female > male does not make any sense. That is, we cannot talk about the order between two categories in case of nominal scale.

If in a measurement scale orders also make sense then, this scale comes under the heading ordinal scale discussed below.

(2) Ordinal Scale

We have seen that order does not make any sense in nominal scale. As the name ordinal itself suggests that other than the names or codes given to the different categories, it also provides the order among the categories. That is, we can place the objects in a series based on the orders or ranks given by using ordinal scale. But here we cannot find actual difference between the two categories.

Generally ordinal scale is used when we want to measure the attitude scores towards the level of liking, satisfaction, preference, etc. Different designation in an institute can also be measured by using ordinal scale.

To get more familiar with the concept of ordinal scale let us consider some examples:

(i) Opinion of persons about proposal of introducing co-education in a college comes under this scale. Suppose we assign '1'to strongly disagree, '2' to disagree '3' to indifferent (or neutral) '4' to agree and '5' to strongly agree. Here, the order also matters and mathematically we may use the symbols >, < in addition to those used for nominal scale, i.e. =, $\neq$ as here strongly agree opinion comes first in order as compared to agree and so on, i.e. $5 > 4 > 3 > 2 > 1$ or $1 < 2 < 3 < 4 < 5$. But actual difference between different categories in this Likert Scale is not possible. This is because, suppose "strongly agree" means he/she gives marks from 75% to 100% for the co-education to be introduced and suppose "agree" means the marks are given in the range say 50% to 75%. Now, the actual difference between "strongly agree" and "agree" is not feasible in this sense.

[∵ "strongly agree" may have the marks percentage as 80% and "agree" marks 74%. Similarly in other case these values may be 90% and 70% respectively.]

(ii) Suppose, a school boy is asked to list the name of three ice-cream flavours according to his preference. Suppose he lists them in the following order:

Vanilla
Straw berry
Tooty-frooty

This indicates that he likes vanilla more compared to straw berry and straw berry more as compared to tooty-frooty. But the actual difference between his liking between vanilla and straw berry cannot be measured.

(iii) In sixth pay commission, teachers of colleges and universities are designated as Assistant Professor, Associate Professor and Professor. The rank of Professor is higher than that of Associate Professor and designation of Associate Professor is higher than Assistant Professor. But you cannot find the actual difference between Professor and Associate Professor or Professor and Assistant Professor or Associate Professor and Assistant Professor. This is because, one teacher in a designation might have served certain number of years and have done a good quality of research work, etc. and other teacher in the same designation might have served for lesser number of years have done unsatisfactory research work, etc. So, the actual difference between one designation and other designation cannot be found. So one may be very near to his next higher designation and other may be very far from it depending on their quality of teaching/research.

(iv) Based on economic condition of a family, generally families of a society are divided into three categories:

Higher class family
Middle class family
Lower class family

Every body knows that economic condition of higher class family is better than middle class family and middle class family is in a better condition compare to lower class family. But the actual difference between the economic condition of a higher class family and middle class family or between middle class family and lower class family cannot be measured.

That is, we can only give order/rank to the three classes of the families but actual difference cannot be measured. In all the above examples, the actual difference is not possible because, all the ranks are on the ranges and not on fixed points.

Now, we are going to study the next higher level of measurement wherein the actual differences can be found. This scale is known as interval scale.

(3) Interval Scale

You have become familiar with the concept of interval and its length in Sec. 2.2 of Unit 2 of this course. If I = [4, 9] then length of this interval is $9-4=5$, i.e. difference between 4 and 9 is 5, i.e. we can find the difference between any two points of the interval. For example, $7, 7.3 \in I$ and difference

between 7 and 7.3 is 0.3. Thus we see that property of difference holds in case of intervals. Similarly, third level of measurement, i.e. interval scale possesses the property of difference which was not satisfied in case of nominal and ordinal scales.

Nominal scale gives only names to the different categories, ordinal scale moving one step further also provides the concept of order between the categories and interval scale moving one step ahead to ordinal scale also provides the characteristic of the difference between any two categories.

Interval scale is used when we want to measure years/historical time/calendar time, temperature (except in the Kelvin scale), sea level, marks in the tests where there is negative marking also, etc. Mathematically, this scale includes +, – in addition to >, < and $=, \neq$.

To get more familiar with the concept of interval scale, let us consider some examples:

(i) The measurement of time of an historical event comes under interval scale because there is no fixed origin of time (i.e. '0' year). As'0' year differ calendar to calendar or society/country to society/country e.g. Hindus, Muslim and Hebrew calendars have different origin of time, i.e. '0' year is not defined. In Indian history also, we may find BC (Before Christ).

(ii) Measurement of temperature in degree Celsius (^{0}C) assumes 0^0C when water starts freezing to ice and it becomes ice at -40°C. So, in degree Celsius origin is arbitrary that's why measurement of temperature in degree Celsius comes in interval scale. Because in degree Celsius origin is arbitrary, so we cannot say that 30^0C is twice as hot as 15^0C. Because if it is so then can we say that 4°C is -1 times -4°C? No it is meaningless. Similarly, measurement of temperature in Fahrenheit comes in the interval scale.

(iii) Mean sea level (MSL) also have arbitrary origin because it is mean of two means, mean high tide and mean low tide(and mean high tide and mean low tide vary according to high and low pressure zones). Further it also varies place to place and time to time. So measurement of sea level also comes in the interval scale.

(4) Ratio Scale

Ratio scale is the highest level of measurement because nominal scale gives only names to the different categories, ordinal scale provides orders between categories other than names, interval scale provides the facility of difference between categories other than names and orders but ratio scale other than names, orders and characteristic of difference also provides natural zero (absolute zero). In ratio measurement scale values of characteristic cannot be negative.

Ratio scale is used when we want to measure temperature in Kelvin, weight, height, length, age, mass, time, plane angle, etc. Ratio scale

includes $\times, \div$ in addition to $+, -, >, <, =, \neq$. But be careful never take '0' in the denominator while finding ratios. For example, $\frac{4}{0}$ is meaningless.

To get more familiar with the concept of ratio scale let us consider some examples, where ratio scale is used:

(i) Measurement of temperature in Kelvin scale comes under ratio scale because it has an absolute zero which is equivalent to -273.15^0C. This characteristic of origin allows us to make the statement like 50K ('50K' read as 50 degree Kelvin) is 5 time hot compare to 10K.

(ii) Measurement of money also comes under ratio scale because it satisfies all the requirement of interval scale and has a natural zero. For example, suppose there are 60 teachers in a particular school in Delhi. If we associate a unique number to each teacher related to the cash (in rupees) he/she has with him/her at the time of investigation. Then we have a fixed whole number corresponding to each teacher. Of course two or more teachers may have same cash (in rupees). These teachers will be allotted the same whole number and will fall in one category. Here we note that, the whole numbers allotted to the teachers can be ordered, have an actual difference and also have origin (i.e. absolute zero '0'). Here natural zero indicates the absence of money in the pocket of the teacher. If a teacher has Rs 500 and another teacher has Rs 100 then we can say that the teacher having Rs 500 has 5 times amount than a teacher having Rs 100. Thus it satisfies all the requirement of ratio scale.

(iii) Both height (in cm.) and age (in days) of students of M.Sc. Statistics of a particular university satisfy all the requirements of a ratio scale. Because height and age both cannot be negative (i.e have an absolute zero).

Permissible Statistical Tools

One of the advantages of measurement scale is that these help us to decide which statistical tool should be used in a given situation.

Table 11.1 shows the list of permissible statistical tools in case of nominal, ordinal, interval and ratio scales. Based on information provided by these scales, their levels from lowest to height are nominal, ordinal, interval and ratio (see Fig 11.1). That is why all the Statistical tools applicable on the lower scale will automatically be applicable on the next level scale. So, we will not repeat the permissible statistical tools used in lower level scale. It is understood that statistical tools which are permissible for nominal will be permissible in case of ordinal and so on.

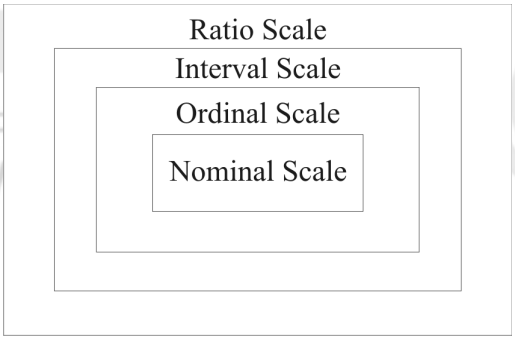

Fig. 11.1

Table 11.1

MEASUREMENT SCALE	PERMISSIBLE STATISTICAL TOOLS	LOGIC/REASON
NOMINAL SCALE	Mode, chi-square test and run test	Here counting is only permissible operation.
ORDINAL SCALE	Median all positional averages like quartile, Decile, percentile, Spearman's Rank correlation.	Here other than counting, order relation (less than or greater than) also exists.
INTERVAL SCALE	Mean , S.D., t-test, F-test, ANOVA, sample multiple and moment correlations, regression.	Here counting, order and difference operations hold.
RATIO SCALE	Geometric mean (G.M.), Harmonic mean (H.M.), Coefficient of variation.	Here counting, order, difference and natural zero exist.

Before closing this section let us consider some situations and appropriate measurement scale that can be used with the help of some examples followed by some exercises.

Example 1: If you want to collect the data based on the characteristic of literacy then which scale will be used? Explain with reasons.

Solution: Appropriate scale is nominal scale because population can be categorised in two categories literate (L) and illiterate (I). The symbols for literate and illiterate can be used according to our choice like 0, 1 or A, B or X, Y, etc.

Example 2: At a picnic spot in India, 1000 tourists visit over a period of 7 days. Each tourist is asked the name of the country of his/her birth. Then the data thus obtained come under which measurement scale.

Solution: Nominal scale, because the characteristic 'name of the country' divides the tourists into different categories each labels with the name of his/her country.

Example 3: Answer the following questions:

(i) Which scale is at lowest level?

(ii) Which scale is at highest level?

(iii) Which scale has absolute zero?

(iv) Which scale is used to find the mean sea level (MSL)?

Solution:

(i) Nominal scale is at lowest level, because it has only one permissible operation counting.

(ii) Ratio scale is at highest level, because it has all the four operations counting, order, distance and absolute zero.

(iii) Ratio scale is only scale out of the four measurement scales nominal, ordinal, interval and ratio scales which has absolute zero.

(iv) Because sea level has no absolute zero, so interval scale is used to find the mean sea level (MSL).

Example 4: Answer the following questions:

(i) In which scale median is not permissible?

(ii) In which scale(s) mean is not permissible?

(iii) In which scale(s) geometric mean and harmonic mean are not permissible?

(iv) In which scale geometric mean and harmonic mean are permissible?

Solution:

(i) In order to find median, we have to arrange the data in ascending or descending order of magnitude. But in nominal scale order operation is not present. So, in case of nominal scale data, median is not permissible.

(ii) In order to find mean, each observation of the data must be associated with a numerical quantity (which exactly measure the quantity of the characteristic). But, this requirement is not fulfilled by nominal and ordinal scales data. So, mean is not permissible in case of nominal and ordinal data.

(iii) In order to find geometric mean (G.M.) and harmonic mean (H.M.) absolute zero must be defined so that one can talk of quotient/ratio of two numbers. But as absolute zero is defined only in ratio scale, so G.M. and H.M. are not permissible in nominal, ordinal and interval scales data.

(iv) As discussed in (iii), G.M. and H.M. are defined only in ratio scale.

Now, here are some exercises for you.

E 1) Answer the following questions:

(i) Which scale is considered as the best scale of measurement so far as criteria of information provide is concerned?

(ii) Which scale is used in case of measurement of height, weight and age?

(iii) Allotment of license plates to different car comes under which scale of measurement?

(iv) Characteristic of equal distance between any two observations is maintained by which scale(s) of measurement?

E 2) Measurement of blood group comes under which scale of measurement?

11.7 TYPES OF DATA

Data play the role of raw material for any statistical investigation and defined in a single sentence as

"The values of different objects collected in a survey or recorded values of an experiment over a time period taken together constitute what we call data in Statistics"

Each value in the data is known as observation. Statistical data based on the characteristic, nature of the characteristic, level of measurement, time and ways of obtaining it may be classified as follows:

- Quantitative data
- Qualitative data

} based on the characteristic

- Discrete data
- Continuous data

} based on nature of the characteristic

- Nominal data
- Ordinal data
- Interval data
- Ratio data

} based on level of measurement

- Time Series data
- Cross Sectional data

} based on time component

- Primary data
- Secondary data

} based on the ways of obtaining the data

Let us discuss different types of data one by one:

Quantitative Data

As the name quantitative itself suggests that it is related to the quantity. In fact, data are said to be quantitative data if a numerical quantity (which exactly measure the characteristic under study) is associated with each observation.

Generally, interval or ratio scales are used as a measurement of scale in case of quantitative data. Data based on the following characteristics generally gives quantitative type of data. Such as weight, height, ages, length, area, volume, money, temperature, humidity, size, etc.

For example,

(i) Weights in kilogram (say) of students of a class.

(ii) Height in centimeter (say) of the candidates appearing in a direct recruitment of Indian army organised by a particular cantonment.

(iii) Age of the females at the time of marriage celebrated over a period of week in Delhi.

(iv) Length (in cm) of different tables in a showroom of furniture.

Here, is an exercise for you

E 3 Provide an example based on each of the following characteristic:
(i) Area (ii) Volume (iii) Money (iv) Temperature (v) Humidity (vi) Size

---

Qualitative Data

As the name qualitative itself suggests that it is related to the quality of an object/thing. It is obvious that quality cannot be measured numerically in exact terms. Thus, if the characteristic/attribute under study is such that it is measured only on the bases of presence or absence then the data thus obtained is known as qualitative data.

Generally nominal and ordinal scales are used as a measurement of scale in case of qualitative data. Data based on the following characteristics generally gives qualitative data. Such as gender, marital status, qualification, colour, religion, satisfaction, types of trees, beauty, honesty, etc.

For example,

(i) If the characteristic under study is gender then objects can be divided into two categories, male and female.

(ii) If the characteristic under study is marital status then objects can be divided into four categories married, unmarried, divorcee, widower.

(iii) If the characteristic under study is qualification (say) 'matriculation' then objects can be divided into two categories as 'Matriculation passed' and 'not passed'.

(iv) If the characteristic under study is 'colour' then the objects can be divided into a number of categories Violet, Indigo, Blue, Green, Yellow, Orange and Red.

Here, is an exercise for you.

E 4 Give an example based on the following characteristic:

(i) Religion (ii) Satisfaction

Discrete Data

If the nature of the characteristic under study is such that values of observations may be at most countable between two certain limits then corresponding data are known as discrete data (concept of countability have already been discussed in Sec 2.6 of Unit 2 of this course).

For example,

(i) Number of books on the self of an elmira in a library form discrete data. Because number of books may be 0 or 1 or 2 or 3,…. But number of books cannot take any real values such as 0.8, 1.32, 1.53245, etc.

(ii) If there are 30 students in a class, then number of students presents in a lecture forms discrete data. Because number of present students may be 1 or 2 or 3 or 4 or…or 30. But number of present students cannot take any real values between 0 and 30 such as 1.8675, 22.56, 29.95, etc.

(iii) Number of children in a family in a locality forms discrete data. Because number of children in a family may be 0 or 1 or 2 or 3 or 4 or…. But number of children cannot take any real values such as 2.3, 3.75, etc.

(iv) Number of mistakes on a particular page of a book. Obviously number of mistakes may be 0 or 1 or 2 or 3…. But cannot be 6.74, 3.9832, etc.

Continuous Data

Data are said to be continuous if the measurement of the observations of a characteristic under study may be any real value between two certain limits.

For example,

(i) Data obtained by measuring the heights of the students of a class of say 30 students form continuous data, because if minimum and maximum heights are 152cm and 175 cm then heights of the students may take any possible values between 152 cm and 175 cm. For example, it may be 152.2375 cm, 160.31326… cm, etc.

(ii) Data obtained by measuring weights of the students of a class also form continuous data because weights of students may be 48.25796…kg, 50.275kg, 42.314314314…kg, etc.

Here is an exercise for you.

E 5) Identify whether the data are discrete or continuous in the following cases:

(i) Number of people present in a party.

(ii) Length of leafs of a plant.

(iii) Lifetime in hours of an electrical bulb.

(iv) Number of cars standing in a showroom over a period of 7 days.

(v) Number of patients visited to a hospital on a particular day.

Nominal Data

Data collected using nominal scale is called nominal data.

Similarly, data collected using ordinal scale, interval scale and ratio scale are called **ordinal data, interval data** and **ratio data** respectively. These scales of measurement have already been discussed in detail in Sec. 11.6.

Time Series Data

Collection of data is done to solve a purpose in hand. The purpose may have its connection with time, geographical location or both. If the purpose of data collection has its connection with time then it is known as time series data. That is, in time series data, time is one of the main variables and the data collected usually at regular interval of time related to the characteristic(s) under study show how characteristic(s) changes over the time.

For example, quarterly profit of a company for last eight quarters, yearly production of a crop in India for last six years, yearly expenditure of a family on different items for last five years, weekly rate of inflation for last ten weeks, etc. all form time series data.

Yearly expenditures (in Rs) for a family on different items from 2006 to 2010 are given in the following table.

Year	Food	Education	Rent	Miscellaneous	Total
2006	40000	10000	36000	20000	106000
2007	45000	12000	40000	28000	125000
2008	54000	15000	45000	32000	146000
2009	60000	20000	50000	40000	170000
2010	70000	30000	55000	45000	2000000

Data given in above table is nothing but time series data.

Note 2: If the purpose of the data collection has its connection with geographical location then it is known as **Spatial Data.**

For example,

(i) Price of petrol in Delhi, Haryana, Punjab, Chandigarh at a particular time.

(ii) Number of runs scored by a batsman in different matches in a one day series in different stadiums.

Note 3: If the purpose of the data collection has its connection with both time and geographical location then it is known as **Spacio Temporal Data.**

For example, data related to population of different states in India in 2001 and 2011 will be Spacio Temporal Data.

Note 4: In time series data, spatial data and spacio temporal data we see that concept of frequency have no significance and hence known as **non-frequency**

data. For instance, in the example discussed in case of time series data, expenditure of Rs 40000 on food in 2006 is itself important, here its frequency say 3 (repeated three times) does not make any sense.

Note 5: Now consider the case of marks of 40 students in a class out of 10 (say). Here we note that there may be more than one student who score same marks in the test. Suppose out of 40 students 5 score 10 out of 10, it means marks 10 have frequency 5. This type of data where frequency is meaningful is known as **frequency data.**

Cross Sectional Data

Sometimes we are interested to know that how a characteristic (such as income or expenditure, population, votes in an election, etc.) under study at one point in time is distributed over different subjects (such as families, countries, political parties, etc.). This type of data which is collected at one point in time is known as cross sectional data.

For example, annual income of different families of a locality, survey of consumer's expenditure conducted by a research scholar, opinion polls conducted by an agency, salaries of all employees of an institute, etc.

Note 6:

(i) If you are interested to know the changes in a characteristic say expenditure of a family over a period of time then you have to use time series data.

(ii) If you are interested to know the changes in a characteristic say expenditure of different families at single point in time you have to use cross sectional data.

Primary Data

Data which are collected by an investigator or agency or institution for a specific purpose and these people are first to use these data, are called primary data. That is, these data are originally collected by these people and they are first to use these data. Primary data have been discussed in Sec. 12.2 of next unit (i.e. UNIT 12) of this course in detail.

For example, suppose a research scholar wants to know the mean age of students of M.Sc. Chemistry of a particular university. If he collects data related to the age of each student of M.Sc. Chemistry of that particular university by contacting each student personally then data so obtained by the research scholar is an example of primary data for the same research scholar.

Secondary Data

Data obtained/gathered by an investigator or agency or institution from a source which already exists, are called secondary data. That is, these data were originally collected by an investigator or agency or institution and has been used by them at least once. And now, these data are going to be used at least second time. Secondary data have been discussed in Sec. 12.3 of next unit (i.e. UNIT 12) of this course in detail.

For example, consider the same example as discussed in case of primary data. If the research scholar collects the ages of the students from the record of that particular university, then the data thus obtained is an example of secondary data. Note that, in both the cases data remain the same, only way of collecting the data differs.

11.8 SUMMARY

In this unit, we covered following topics:

1) Origin and development of statistics.
2) Definitions of statistics by different authors.
3) Scope and uses of statistics.
4) Limitations of statistics.
5) Different measurement scales and types of data.
6) Frequency and non frequency data.

11.9 SOLUTIONS/ANSWERS

E 1 (i) Ratio scale is considered as the best measurement scale so far as the criteria of information provide are concerned because all the four operations counting, order, distance and absolute zero are defined on the observations.

(ii) Measurement of height, weight, age requires absolute zero and only ratio scale has absolute zero. So, appropriate scale of measurement for height, weight, age is ratio scale.

(iii) Allotment of license plates to the different cars comes under nominal scale measurement, because license plates categories the cars or license plates only provide unique names to the cars. Further, the car remains the same if some other registration number is provided to it.

(iv) Characteristic of equal distance between any two observations is maintained by two scales of measurements interval and ratio scales. For example, distance between temperatures of 18K and 13K is same as distance between 100K and 105K.

E 2) Blood group just divides the objects/things into four categories named as A, B, AB, O. So it comes under nominal scale.

E 3) Answers are not unique. There are a number of examples for each part, here one answer is provided for each part.

(i) Area of each state (in km^2) of India.

(ii) Volume of different buckets available at a particular shop.

(iii) Income of each family over a period of one year in a particular locality.

(iv) Highest or lowest temperature of a place over a period of 50 days.

(v) Level of humidity of a particular place at each hour of a particular day.

(vi) Size of different shoes present at a particular showroom on a specified day.

E 4) (i) If the characteristic under study is 'religion' then the objects can be divided into five categories Hindu, Muslim, Sikh, Isaiah, and others.

(ii) If the characteristic under study is 'satisfaction' then the objects can be divided into five categories (Likert scale) as shown on the next page:

Highly satisfied	Satisfied	Neutral	Dissatisfied	Highly dissatisfied
5	4	3	2	1
		OR		
2	1	0	–1	–2

E 5 (i) Number of people present in a party may be 2 or 3 or 4 or 5 or 6 and so on, but cannot be 2.3, 4.375, 9.62875, etc.
$\therefore$ it is an example of discrete of data.

(ii) Lengths of leafs of a plant form continuous data because lengths of leafs may be any real number, e.g. 3.75 cm, 2.959595… cm, etc.

(iii) It is an example of continuous data because life time of an electrical bulb may be any possible fraction of time. For example, 8 hours 8.76 hours, 100.25796 hours, 0.25 hours, etc.

(iv) It is an example of discrete data because number of cars may be 1 or 2 or 3 or 4 or 5 or 6 or 7 and so on, but cannot be 2.87, 5.687, etc.

(v) It is an example of discrete data because number of patients visited to a hospital on a particular day may be 0 or 1 or 2 or 3 or 4 and so on, but cannot be 2.8, 10.357, 7.856, etc.