
UNIT 1 INTRODUCTION TO STATISTICS

- 1.0 Introduction
- 1.1 Objectives
- 1.2 Statistics: Definition and Uses
 - 1.2.1 Definition of Statistics
 - 1.2.2 Uses of Statistics
 - 1.2.3 Limitations of Statistics
- 1.3 Measurement Scales
- 1.4 Types of Data
- 1.5 Data Collection
- 1.6 Population and Sampling
- 1.7 Descriptive and Inferential Statistics
 - 1.7.1 Descriptive Statistics
 - 1.7.2 Inferential Statistics
 - 1.7.3 Probability Models
- 1.8 Summary
- 1.9 Solutions/Answers
- 1.10 References

1.0 INTRODUCTION

This is the first Unit of the course on Mathematical Foundations. This Unit introduces you to the basics of statistics, which have been discussed in detail in this course. The unit defines the term statistics and traces a brief history of the development of statistics. The uses of statistics are also highlighted briefly. The statistical analysis is performed on the data, which is obtained by means of data collection processes. This Unit first introduces the data measurement scales and then defines the types of data. It also briefly presents the data collection methods. In addition, the unit also provides a brief introduction to the terms population, population parameters and sample.

After collecting the data, you would like to use this data to perform an analysis. Therefore, a brief introduction has been provided about the terms descriptive analysis and inferential analysis. Since the analysis is based on a probability model, a brief account of probability models has been introduced in this unit.

This is an introductory unit, and some of the concepts discussed in this unit will be explained in the later units.

1.1 OBJECTIVES

After going through this unit, you should be able to:

- define statistics and briefly trace its evolution;
- list and illustrate the uses of statistics;
- explain different measurement scales;
- define different types of data;
- list various methods of data collection;
- differentiate between population and sampling;
- define the terms descriptive and inferential statistics.

1.2 STATISTICS: DEFINITION AND USES

Consider the following case:

An Open University has many regional offices. These regional offices maintain records of fees collected from the students and the expenses made by the office in different programmes using spreadsheet software. These records may be used to determine information such as:

- Fee collection of each programme.
- The percentage of students of a given programme who have not paid the fee.
- Which programme is most profitable?
- Total expenses of each programme.
- Which programmes are not profitable and may be changed?
- Which programme has the maximum number of students?
- Programme-wise distribution of the fee collected.
- Number of dropout students of a programme.
- Expected fee collection of future years.
- Expected number of students in different programmes during the next term.

How can you create such information from the raw data on fee records of the students? You will be using Statistics to do so.

Statistics is used by human society, 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 past experiences (i.e. based on statistical data he/she has actually experienced), and (ii) future events are also predicted by using/examining the past behaviour of that particular event.

The credit for 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. 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 a major role in addressing various problems of planning of national development and social welfare in the country. 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.

Statistics in the World

Without going into more details, we will concentrate on only some major discoveries in the area of Statistics at the international level. Some key developments in statistics are:

- John Graunt (1620-1674) 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 the father of vital statistics.
- Gauss (1777-1855) gave the principle of least squares and the normal law of errors.
- J. Bernoulli (1654-1705) was the first person who stated the law of large numbers.
- Statistical methods in the field of biometry were first introduced by Sir Francis Galton (1822-1911).
- Later, Professor Karl Pearson (1857-1936) created correlation coefficient theories. He was also the founder of the 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'.
- Credit of discovering a 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 the Father of Statistics.

1.2.1 Definition of Statistics

Statistics is a very old science, and it has developed through the ages. So, it is not surprising that through its long journey, its definitions given by different authors vary. We will present the most recent and broader definitions of Statistics:

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

1.2.2 Uses of Statistics

In present times, statistics is regarded as a science having sound techniques of handling huge data and providing valuable conclusions. Today, statistical methods find their 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. Let us briefly define some of the applications of statistics.

- **Statistics and Industry:** Statistics plays an important role in quality control and production engineering. For example, various control charts are used to maintain a certain quality level and different inspection plans are used in production engineering. You may find the average life of some products, such as an electric bulb, using Statistical methods.
- **Biology and Statistics:** Professor Karl Pearson has stated that the whole doctrine of heredity rests on statistical basis. This is generally said that the height of the child is associated with the height of the father. To test this type of hypothesis, statistics is the only science that provides the scientific methods. Vital statistics is totally devoted to the different aspects of human life, like the average life expectancy of men and women, birth and death rates, etc.

- **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 good planning, it has become necessary to analyse 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 analyse these types of data, one has to totally depend on the statistical techniques.
- **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 businessperson to know the coming trend of the 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 fertilisers are available in the market. For a good yield, it has become necessary to know 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.
- **Statistics and Insurance Sector:** Insurance sector totally depends on the statistical data and different concepts of probability theory. Life tables lie at the heart of human insurance. Future lifetime of a life is calculated using the concept of random variables and their expected values. 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 a very important aspect in every discipline. In many disciplines, such as psychology, tourism, education, management, engineering, etc., one has to collect data on the characteristics of interest under study. Some of the related questions here are: which measurement of scale to be used, and which tests to be used. This requires the knowledge of different types of measurement scales and, accordingly, a suitable statistical tool is needed.
- **Statistics and Economics:** In order to know about the development of a country, it has become necessary to obtain the data related to its economic growth. Again, statistical tools are needed to collect relevant data (such as related to agricultural, industrial, literacy, etc.) and for its analysis.

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

1.2.3 Limitations of Statistics

In the previous sub-section of this unit, you have seen a wide range of applications of statistics. 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 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) **Dealing with a Group:** The 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 not 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, if a statistical report says that 70% population of India lives in rural areas. It does not imply that if you visit a public place like a bus stand, a railway station, etc. and ask the people about their living place. Results may surprise you and may highly differ from the above figure. But you may note that as sample size increases, the result will also come nearer and nearer to exact figure 70%. But there are sciences like mathematics where exactness is maintained. For example, if a book seller gets 5% profit on selling a particular book. Then it is sure that if the sale price of that particular book is Rs 200, he/she will get Rs 10 as profit, and in case of the sale price of Rs 300, the 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 drawbacks of the science of statistics. There are many statistical tools of similar type. For example, to find average in a particular situation, which of the possible tools like mean, median, mode, geometric mean, harmonic mean, etc., is appropriate, needs the hands of experts. Similarly, to test a given statistical hypothesis, which of the possible tools like Z-test, t-test, χ^2 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.

After going through the basics of statistics, let us now discuss an important aspect of statistics – measurement in the next section.

1.3 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 notebook, 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 four 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:

Nominal Scale

The word nominal has come from this Latin word, i.e. ‘Nomen’. In Latin, ‘Nomen’ means name. 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 at least 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 (“=”, “ $\neq$ ”) 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 categorise the data based on a 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 three categories: male ‘M’, female ‘F’ and Third Gender ‘T’

Category	Name/Code
Male	M or 0
Female	F or 1
Third Gender	T or 2

Please note that the main thing is that we have to give a unique name to each category.

(ii) 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.

Note 1: We note that in nominal scale, we have just coded the objects. Signs of less than or greater than do not make any sense in a nominal scale. For example, Female > Male or Male > Female does not make any sense. That is, we cannot talk about the order between the three categories in the case of a nominal scale.

If in a measurement scale, order also makes sense, then this scale comes under the heading of an ordinal scale, discussed below.

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 designations in an institute can also be measured by using an 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.
- (ii) Based on the income of a family, generally, families of a society may be divided into three categories:

Higher income group

Middle-income group

Low-income group

The income of a person in the higher income group is better than that of a person in the middle income group, and the income of a person in the

middle income group is higher than that of a person in the lower income group. But the actual difference between them 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.

Interval Scale

The third level of measurement, i.e. interval scale, possesses the property of difference, which was not satisfied in the case of nominal and ordinal scales. The 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, etc. Mathematically, this scale includes $+$, $-$ in addition to $>$, $<$ and $=$, $\neq$, $\neq$.

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

- (i) Measurement of temperature in degrees Celsius ($^{\circ}\text{C}$ assumes 0°C when water starts freezing to ice. So, in degrees Celsius, the origin is arbitrary; that's why measurement of temperature in degrees Celsius comes in an interval scale. Because the degree Celsius origin is arbitrary, we cannot say that 30°C is twice as hot 15°C . Similarly, measurement of temperature in Fahrenheit comes in the interval scale.
- (ii) Mean sea level (MSL) also has an arbitrary origin because it is the 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 from place to place and time to time. So, measurement of sea level also comes in the interval scale.

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 a 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 $\times$, $\div$, $\times$, $\div$ in addition to $+$, $-$, $>$, $<$, $=$, $\neq$. But be careful never to take '0' in the denominator while finding ratios. For example, $4/0$ 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°C . This characteristic of origin allows us to make the statement that 50K ('50K' read as 50 degrees Kelvin) is 5 times hotter compared to 10K.

- (ii) Measurement of money also comes under the ratio scale because it satisfies all the requirements of an interval scale and has a natural zero. For example, suppose person A has Rs 100 and person B has Rs 500, then we can say that B has 5 times the amount of A. Also, a person may not have any money (i.e. absolute zero '0'). Here, the natural zero indicates the absence of money for a person. Thus, it satisfies all the requirements of a ratio scale.
- (iii) Both height (in cm.) and age (in days) of students 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 1.1 shows the list of permissible statistical tools in the case of nominal, ordinal, interval and ratio scales. Based on information provided by these scales, their levels from lowest to highest are nominal, ordinal, interval and ratio. 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.

Table 1.1: The Measurement Scale and Permissible Statistical Tools

MEASUREMENT SCALE	PERMISSIBLE STATISTICAL TOOLS	LOGIC/REASON
NOMINAL SCALE	Mode, chi-square 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.

Check Your Progress 1

1. Is machine learning or deep learning an application of statistics?

.....

.....

.....

.....

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 in which he/she was born.

Then, the data thus obtained comes under which measurement scale?
Give reasons in support of your answer.

.....
.....
.....
.....

3. Answer the following questions:

- (i) Which scale is at the lowest level?
- (ii) Which scale is at the highest level?
- (iii) Which scale has absolute zero?
- (iv) Which scale is used to find the mean sea level (MSL)?

.....
.....
.....
.....

4. Answer the following questions:

- (i) In which scale is the median not permissible?
- (ii) In which scale(s) mean is not permissible?
- (iii) In which scale(s) are geometric mean and harmonic mean not permissible?
- (iv) In which scale are geometric mean and harmonic mean permissible?

.....
.....
.....
.....

1.4 TYPES OF DATA

Data plays the role of raw material for any statistical investigation. Data can be 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 an 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
-
- TimeSeries data
 - Cross Sectional data
- } based on time component
-
- Primary data
 - Secondary data
- } based on the ways of obtaining the data
-
- 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
-
- TimeSeries 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 quantity. In fact, data are said to be quantitative data if a numerical quantity (which exactly measures the characteristic under study) is associated with each observation.

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

For example,

- (i) Weights in kilograms (say) of students in a class.
- (ii) Height in centimetres (say) of the candidates appearing in a physical examination.
- (iii) Length (in cm) of different tables in a furniture showroom.

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 basis of presence or absence, then the data thus obtained is known as qualitative data.

Generally, nominal and ordinal scales are used as measurement scales 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 marital status, then objects may be divided into four categories: married, unmarried, divorced, and widowed.
- (ii) If the characteristic under study is qualification (say) 'matriculation', then objects can be divided into two categories as 'Matriculation passed' and 'Matriculation not passed'.

Discrete Data

If the nature of the characteristic under study is such that values of observations may be at most countable between two limits, then corresponding data are known as discrete data.

For example,

- (i) The number of books on the shelf of an almirah in a library forms discrete data, as the number of books may be 0, 1, 2, 3 and so on. This is because the number of books cannot take fractional values.
- (ii) If there are 30 students in a class, then number of students present in a lecture forms discrete data.

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 limits.

For example,

- (i) Data obtained by measuring the heights of the students of a class of 30 students form continuous data, because if the minimum and maximum heights are 152cm and 175 cm, respectively, then the height of a student may take any possible value between 152 cm and 175 cm. For example, it may be 152.2 cm, 160.3 cm, etc.
- (ii) Data obtained by measuring weights of the students of a class also form continuous data because weights of students may be 48.235kg, 50.270kg, 42.330kg, etc.

You can refer to the scales of measurement in Section 1.3 for details on nominal, ordinal, ratio and interval types of data

Time Series Data

The collection of data is done to address a specific purpose. The purpose may have its connection with time, geographical location or both. If the purpose of data collection has a 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 is collected usually at regular intervals of time related to the characteristic(s) under study. The objective here is to show how characteristic(s) change over 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, as given in the following table, are nothing but time series data.

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	200000

Note 1: 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 2: If the purpose of the data collection has its connection with both time and geographical location, then it is known as **Spatio-Temporal Data**.

For example, data related to population of different states in India in 2001 and 2011 will be Spatio-Temporal 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:

- (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 the first to use these data, are called primary data. That is, these data are originally collected by these people and they are the first to use these data.

For example, suppose a research scholar wants to know the mean age of students in class 10 of a school. S/he collects data related to the age of each student in class 10 of that school by contacting each student personally, then the data 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 have been used by them at least once. And now, these data are going to be used at least second time

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

1.5 DATA COLLECTION

The data collection is performed to create primary data. There are a number of methods of collection of primary data depending upon many factors such as geographical area of the field, money available, time period, accuracy needed, literacy of the respondents/informants, etc. Here we will discuss briefly only the following commonly used methods.

- (1) Direct Personal Investigation Method
- (2) Telephone Method
- (3) Indirect Oral Interviews Method
- (4) Local Correspondents Method
- (5) Mailed Questionnaire Method

Let us discuss these methods briefly.

(1) Direct Personal Investigation Method

In this method, the investigator personally contacts the informants and collects the related data through face to face interviews of the informants. Due to face to face meeting of investigator and informants data collected under this method has maximum degree of accuracy. But the degree of accuracy depends upon the sincerity, honesty, unbiasedness and expertness of the investigator, because it is the investigator who ultimately gives the final shape to the information provided by the informants. This method of collection of primary data is recommended/suitable when field of enquiry is small, secrecy related to data is need, high accuracy is required and time as well as money is sufficient.

(2) Telephone Method

In the direct personal investigation method investigator has to personally contact the informants, but now-a-day's telephone is a very good communication tool. If the information of the interest is collected through telephone then data so obtained come under telephone method.

(3) Indirect Oral Interview Method

In this method, investigator does not meet the actual informants directly, but the related information is collected/obtained from other persons who are supposed to have the required type of information. The informants who provide the information about actual informants are known as 'witnesses'.

The success of this method mainly depends upon the skill and experience of the investigator. Because it depends upon the way, sequence and trick of questions prepared and asked by the investigator from the informants. It also depends on the behaviour and how much he/she is capable to create confidence in the informants.

For example, if a researcher wants to collect the data on the smoking habit of students of a particular class of a college, then it may happen that actual informants do not provide correct information. So, data can be collected with the help of class met or college met or from the neighbours.

(4) Local Correspondents Method

In this method, first of all some local correspondents or agents are appointed by the investigator or agency or institute to collect data. These local agents directly meet to the informants and collect data related to the required purpose in hand. Data collected by these local agents has a high degree of accuracy, because they are familiar with the local language, traditions, general behaviour of the people, etc., of that particular area.

This method of collection of primary data is suitable if

- Area of investigation is large, e.g. news channels have their reporters throughout India.
- Time period in which information is needed is very short. For example, news related to the happening of any special incident can be easily seen on the news channels in very short time period after its happening.
- Information is needed on a regular basis. For example, news are provided daily using this method.

(5) Mailed Questionnaire Method

In this method, first of all, a list of questions related to the information required by the investigator is prepared. At the time of preparing the list of questions, the following points should be kept in mind.

- Number of questions should not be too many.
- Each question should be related directly or indirectly to the objective(s) of the investigation.
- Each question should be clear.
- Generally, objective type questions should be used, but if necessary, multiple choice and open-ended questions can also be used.
- Language should be simple and effective.

After preparing the 'final list' of questions known as questionnaire, it is sent through mail to the informants. A covering letter requesting informants to fill the questionnaire and return it, and a brief introduction about the objective of the investigation, is also attached. This method is suitable when the informants are literate and area of investigation is large.

Check Your Progress 2

1. Characterise the following data as qualitative or quantitative data?

- (i) Cities of a country
- (ii) Temperature
- (iii) Attitude towards learning
- (iv) Fee details

.....

2. Identify whether the data are discrete or continuous in the following cases:

- (i) Number of people present at a party.
- (ii) Length of leaves 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 who visited a hospital on a particular day.

.....

3. Give reasons in each case whether the data are primary or secondary.

- (i) 2011 census data published by the Office of the Registrar General of India and to be used by itself.
- (ii) 2011 census data to be used by a demographer in his study.

.....

4. You have to collect data from about 5000 students of various colleges relating to their experiences in Learning. Which data collection method would you use? Justify your answer.

.....

.....

.....

.....

1.6 POPULATION AND SAMPLING

One of the important considerations while collecting primary data is identifying the population and collecting a representative sample from this population. In this section, we briefly discuss population and sampling.

Population: Population literally means a well-defined group (collection or bunch) of some objects (units or elements). Examples of populations and its units are: a city with some clear-cut territory having its dwellers as units (elements) or having houses as its units; a hospital with indoor patients as units or its doctors as units; a library with its employees as units or books as its units; a school with enrolled students as units; and so on. Thus, population is the collection or group of individuals /items/ units/ observations under study. The total number of elements/items/units/observations in a population is known as its size and generally denoted by N .

Sample: A finite subset of units of a population is called a “sample” and the number of units belonging to the sample is called the sample size. If n denotes the sample size, then necessarily we have the condition $n < N$ and then the sample is said to be a proper sample. However, in some situations, we have $n = N$, that is, all the units of the population are selected as the sample. In almost all kinds of statistical studies, a sample is selected for inferential purposes. However, even if the population is not large enough or it has a well-defined territory or it is not of destructive type; a sample is useful since besides it is always less costly, less time consuming, it helps us to get a handy result in very short time.

A sample of size n is not selected simultaneously, rather n units are selected one by one in n draws in order to constitute a sample. The process through which units of the population are selected in different draws to constitute a sample of specified size, is known as “sampling”. Thus, sampling is a draw-to-draw mechanism of selecting some units from a population. Sampling is mainly of two types, namely, (a) Probability Sampling or Random Sampling, and (b) Non-Probability Sampling. The technique of random sampling is of fundamental importance in the application of statistics. The Estimation theory is based on the assumption of random sampling. It is a scientific method of selecting samples accordingly to some laws of chance in which each unit in the population has some definite pre-assigned probability of being selected in the sample. Simple Random Sampling, Stratified Sampling, Systematic Random Sampling, etc., are examples of random sampling. In non-probability sampling, the sample is selected with definite purpose and, hence, the choice of sampling units depends entirely on the discretion and judgment of the investigator. It is, therefore, not a scientific method and attracts severe criticisms in many circumstances. Purposive Sampling, Judgement Sampling, etc., are the

examples of non-probability sampling. We know that if $n < N$, it is a sampling procedure. When $n = N$, the process is known as “complete enumeration or census”. An example of complete enumeration is the decennial census of India.

Need for Sampling in Statistical Inference: In many practical and real situations, the population under consideration is either infinitely large in size or it is unbounded in the sense that its boundaries are not well-defined, or it is of a destructive nature. For example, in order to determine average life of two hundred produced electric bulbs, it would be necessary to light these bulbs unless they all get fused, resulting in destruction of the entire lot of production. Similarly, in order to estimate the proportion of smokers in a large city, information has to be gathered from each and every person dwelling in the city which would be a very difficult task in terms of manpower, money and time to be required. Sometimes, the population size is not known if its boundaries are not well-defined, like number of fish in a pond or lake. In all such cases, it is not feasible to gather information on all the units, rather it becomes almost impossible. Keeping in view the difficulty in contacting each and every unit of the population due to these reasons and also to save the time, money and manpower to be required for this, generally a part of the population, popularly known as a “sample” is selected in some pre-assigned manner which in turn is used for drawing the inferences on the population itself or on the parameters of the population. The results obtained from the sample are projected in such a way that they are valid for concluding about the entire population. Therefore, the sample works like a “Vehicle” to reach (drawing) at valid conclusions about the population. In fact, a sample helps us to reach to the “whole”(population) from a “part” (sample) in all types of statistical studies. Thus, statistical inference can be defined as: “It is a process of concluding (projecting or inferring) something desired about a given population on the basis of sample results”.

Parameter and Statistic: A parameter is a function of population values that is used to represent certain characteristics of the population. For example, population size, population mean, population variance, population coefficient of variation, population proportion, population correlation coefficient, etc., are all parameters since their calculation involve all the population values. A statistic is a function of sample values only and does not contain any unknown population parameter in it. For example, if $X_1, X_2, \dots, X_n$ represent values of the variable X in a random sample of size n taken from a population, then sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is a statistic.

1.7 DESCRIPTIVE AND INFERENTIAL STATISTICS

People use statistics to summarise data and to make decisions about the problems on the basis of different types of information available. The science of statistics may be broadly studied under two headings: i) Descriptive Statistics, and (ii) Inferential Statistics

1.7.1 Descriptive Statistics

After collecting the information in terms of data, we may like to arrange the collected data in a proper manner because the collected data may be huge in volume. So, here the need for proper arrangement of data arises. In statistical terminology, the proper arrangement of data is known as the presentation of data. There are some frequently used tools available for representing data, and they may be classified into three basic forms: statistical tables or tabulations of data, diagrams and graphs. A pictorial presentation of the tabulated data may be done either with the help of different kinds of diagrams or by graphs. Some of these graphs include the graphs for frequency distributions, graphs for time series data, stem-and-leaf displays and box plots.

The next step is to proceed with data analysis to explore the properties of the data. The statistical tools, which describe the properties of the data, are known as descriptive statistics. The techniques of descriptive statistics are frequently used for analysing the data in various fields. The descriptive techniques involve three major properties that describe the data. These properties are:

- The numerical value of an observation (commonly referred as central value) around which most of the other numerical values show a tendency to concentrate or group, called central tendency.
- The degree to which numerical values are scattered or dispersed around the central value, called dispersion.
- The extent of departure of numerical values from symmetrical distribution (taken as normal distribution) around the central value, called skewness.

You will learn more about these in the subsequent units of this block.

1.7.2 Inferential Statistics

The description or distribution of observations is not the prime objective of any scientific investigation. Organising and summarising data is only one step in the process of analysing the data. In any scientific investigation, either the entire population or a sample is considered for the study. In most of the scientific investigations, a sample, a small portion of the population under investigation, is used for the study. On the basis of the information contained in the sample, we try to draw conclusions about the population. This process is known as statistical inference. Statistical inference is widely applicable in behavioural sciences. Inferential statistics deals with drawing conclusions about a large group of individuals (population) on the basis of observation of a few participants from among them or about the events which are yet to occur on the basis of past events. It provides tools to compute the probabilities of future behaviour of the subjects. Inferential statistics is the mathematics and logic of how this generalisation from sample to population can be made.

There are two types of inferential procedures:

Estimation: In estimation, inference is made about the population characteristics on the basis of what is discovered about the sample. There may be sampling variations because of chance fluctuations, variations in sampling techniques, and other sampling errors. A good statistical estimator should be unbiased, consistent and accurate.

Hypothesis Testing: In hypothesis testing, we hypothesise a value that characterises the population of observations. The question is whether that hypothesis is reasonable in the light of the evidence from the sample. Sometimes, hypothesis testing is referred to as a statistical decision-making process.

You will learn more about these in the Block 3 of this course.

1.7.3 Probability Models

As defined in the previous section, you use inferential statistics to make logical conclusions, as determined by data. However, these conclusions are probabilistic in nature. The probability models form the backbone of making such decisions. These probabilistic decisions are either in a certain approximate range or accuracy, say 95%. Thus, an underlying probability model must be considered while making the decisions.

What is a probability model? A probability model mathematically quantifies the likelihood of a prediction or a decision, given the uncertainty of data or processes that are used to make the prediction.

A probability model is defined by three elements: the sample space of a random process, an event of interest, which should be a subset of the sample space and a set of rules of a function that computes the probability for the event of interest. For example, consider that you perform a random experiment of tossing a coin and the outcome of this toss is to be predicted, then the following would define its probability model:

Sample Space is {Heads, Tails}, as a toss of a coin will result in one of the two states.

Assume that you are interested in the event that “Toss of the coin results in Heads.”

The probability function in this case would define the probability as $1/2$, if the coin is fair.

The probability models can be discrete if the integral outcomes are countable and finite, or continuous if the outcomes are represented using real numbers and are uncountable. You will learn more about probability in the Block 2 of this course.

Check Your Progress 3

1. What should you use, population or sample, for the following situations?

(i) Find the number of patients visiting a hospital in a day.

.....

(ii) Find the attitude of patients towards health.

.....

(iii) Average height of students of class 10.

.....

(iv) Average marks of the students of a School.

.....

2. A survey was performed to find the average age of a person who performs morning exercises. A sample of 2000 was taken for this. The survey showed that the average age of these people was 48 years. In addition, the researcher stated that her data shows that the national average is expected to be 45 to 51 years. Which of these is a Statistic, and which of the data is a parameter?

.....
.....
.....
.....

3. Which of the data in question 2 is obtained using descriptive statistics and which one is obtained through inferential statistics?

.....
.....
.....
.....

1.8 SUMMARY

This Unit defines the term statistics and discusses applications of statistics in different areas. It also highlighted the shortcomings of statistics. Next, the unit explains different measurement scales in detail with the help of examples. Further, the unit explains different kinds of data, such as qualitative, quantitative, time series data, cross-sectional data, and primary and secondary data. The data types will be useful to you, when you study the later units. Next, the unit discusses different ways of collecting primary data. Some of the popular methods are discussed very briefly in this section. As you know about the method of data collection, the unit also discusses about the terms population, population parameters and sampling.

Once the data has been collected, you need to analyse that data. Thus, the unit briefly defines the terms descriptive data analysis, which defines the process of data organisation and summarisation. This is followed by a brief discussion on inferential statistics that may be used to predict the future data or requirements. Finally, unit relates the process of analysis with probability models and explains in brief the concept of a probability model.

You should go through the references for more details on these topics. These references can be downloaded from eGyankosh.ac.in website.

1.9 SOLUTIONS/ANSWERS

Check Your Progress 1

1. Yes, Machine learning uses and builds statistical models, while topics like neural networks in Deep learning use statistical principles.
2. Nominal scale, because the characteristic 'name of the country' divides the tourists into different categories, each labelled with the name of his/her country.
3. (i) Nominal scale is at the lowest level, because it has only one permissible operation, counting.
 (ii) Ratio scale is at the highest level, because it has all four operations: counting, order, distance and absolute zero.
 (iii) Ratio scale is only one 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).
4. (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.

Check Your Progress 2

1

- | | |
|---------------------------------|--------------|
| (i) Cities of a country | Qualitative |
| (ii) Temperature | Quantitative |
| (iii) Attitude towards learning | Qualitative |
| (iv) Fee details | Quantitative |

2.

- (i) Number of people present at a party:

Discrete

(ii) Length of leaves of a plant:

Continuous

(iii) Lifetime in hours of an electrical bulb:

Continuous

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

Discrete

(v) Number of patients who visited a hospital on a particular day:

Discrete

3.

(i) Primary data, as data is used by the office itself

(ii) Secondary data, as data is used by another person, who is not the data creator

4. Mailed questionnaire method, as this method is suitable when the informants are literate and the area of investigation is large. Also, make a questionnaire that is not too large.

Check Your Progress 3

1.

(i) Find the number of patients visiting a hospital in a day: Population, as just one hospital, and limited quantitative data

(ii) Find the attitude of patients towards health: Sample, as very large data and qualitative question

(iii) Average height of students of class 10: Sample, as large data, through quantitative but expensive to use population

(iv) Average marks of the students of a School: Population, limited quantitative data.

2. The average age of these people was 48 years is a Statistic, and the national average is expected to be 45 to 51 years is a parameter.

3. The average age of these people was 48 years (a Statistic), obtained through descriptive analysis, whereas the national average is expected to be 45 to 51 years (a parameter), obtained through inferential statistics.

1.10 REFERENCES

IGNOU Course Material

1. Post-Graduate Diploma In Applied Statistics (PGDAST), MST 001: Matrices, Determinants and Collection of Data, UNIT 11: Introduction to Statistics (for Sections 1.2 to Section 1.4 and related and related portions in Sections 1.0, 1.1, 1.8, 1.9 and Check Your Progress questions)

Descriptive Statistics

2. Post-Graduate Diploma In Applied Statistics (PGDAST), MST 001: Matrices, Determinants and Collection of Data, UNIT 12: Collection and Scrutiny of Data (for Section 1.5 and related and related portions in Sections 1.0, 1.1, 1.8, 1.9 and Check Your Progress questions)
3. MPC-006 Block 1 Introduction to Statistics UNIT 2 Descriptive and Inferential Statistics (for Sections 1.7 and related and related portions in Sections 1.0, 1.1, 1.8, 1.9 and Check Your Progress questions)
4. MEV-019 Research Methodology for Environmental Science, Block 3: Statistical Analysis, Unit 9: Sampling Distributions (for Sections 1.6 and related and related portions in Sections 1.0, 1.1, 1.8, 1.9 and Check Your Progress questions)

Reference Book

5. “Introduction to Probability and Statistics for Engineers and Scientists”, 3rd Edition, Sheldon M. Ross, Elsevier Academic Press, 2004

