B.Sc. I Semester – I
DSC-I: (Descriptive Statistics-I)

Unit 1
1.1. Introduction to Statistics

Definition of Statistics:

Statistics is the branch of science which deals with collection, presentation, analysis and interpretation of the data.

Importance of Statistics:

The importance of Statistics will be clear from the following points.

1. Statistical methods enable to condense the data.
2. Statistical methods give tools of comparison.
3. Estimation, Predication is also possible using statistical tools.
4. We can get idea about the shape, spread and symmetry of the distribution.
5. Inter-relation between two or more variable can be measured using statistical techniques.
6. Statistical methods help in planning, controlling, decision making etc.
7. The use of Statistical methods is important because considerable amount of time, money and manpower can be saved.
8. Statistical methods give systematic methods of data collection and investigation.

Various fields where Statistics is used:

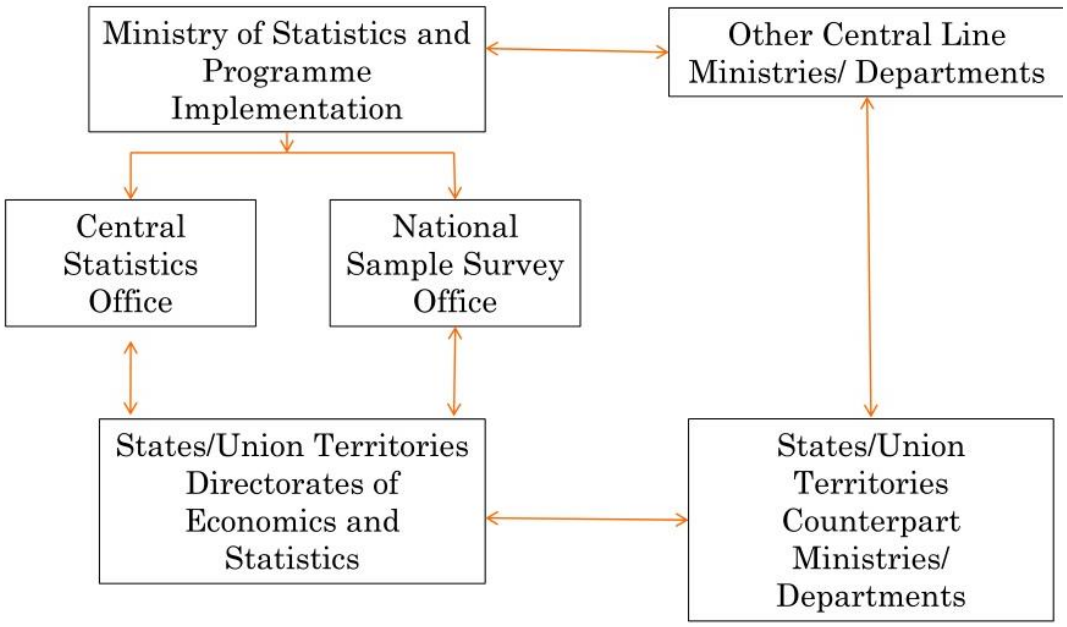
Now statistics holds a central position in almost every field, including industry, commerce, trade, physics, chemistry, economics, mathematics, biology, botany, psychology, astronomy, etc., so the application of statistics is very wide.

Names of various statistical organizations in India:

1) **Central Statistics Office (CSO),** is a governmental agency in India under the Ministry of Statistics and Programme Implementation responsible for co-ordination of statistical activities in India, and evolving and maintaining statistical standards. It has a Graphical Unit. The CSO is located in Delhi
2) **The National Sample Survey(NSS),** Acts as the nodal agency for planned development of the statistical system in the country, lays down and maintains norms and standards in the field of statistics, involving concepts and definitions, methodology of data collection, processing of data and dissemination of results. Organizes and conducts large scale all-India sample surveys for creating the database needed for studying the impact of specific problems for the benefit of different population groups in diverse socio economic areas, such as employment, consumer expenditure, housing conditions and environment, literacy levels, health, nutrition, family welfare, etc.
3) **Indian Statistical Institute** (**ISI**) is a higher education and research institute which is recognized as an Institute of National Importance by the 1959 act of the Indian parliament. It grew out of the Statistical Laboratory set up by Prasanta Chandra Mahalanobis in Presidency College, Kolkata. Established in 1931, this unique institution of India is one of the oldest institutions focused on statistics. Mahalanobis, the founder of ISI, was deeply influenced by the wisdom and guidance of Rabindranath Tagore and Brajendranath Seal. The institute is now considered one of the foremost centres in the

world for training and research in computer science, statistics, quantitative economics and related sciences.

4) The **Ministry of Statistics and Programme Implementation** (MoSPI) is a ministry of Government of India concerned with coverage and quality aspects of statistics released. The surveys conducted by the Ministry are based on scientific sampling methods. The Ministry of Statistics and Programme Implementation (MOSPI) came into existence as an Independent Ministry on 15 October 1999 after the merger of the Department of Statistics and the Department of Programme Implementation.
5) The **National Statistical Commission** (**NSC**) of India is an autonomous body which formed in June 2005 under the recommendation of Rangarajan commission. The objective of its constitution is to reduce the problems faced by statistical agencies in the country in relation to collection of data. Statistical agencies like the Central Statistics Office (CSO) and the National Sample Survey Organization (NSSO) face numerous problems in collecting data from State and Central government departments, but an autonomous body like the NSC is thought to be more able to coordinate things as a statutory status would lend it teeth. It would lay special emphasis on ensuring collection of unbiased data so as to restore public trust in the figures released by the Government.
6) **Registrar General and Census Commissioner of India**, founded in 1961 by Government of India Ministry of Home Affairs, for arranging, conducting and analyzing the results of the demographic surveys of India including Census of India and Linguistic Survey of India. The position of Registrar is usually held by a civil servant holding the rank of Joint Secretary. The Indian Census is the largest single source of a variety of statistical information on different characteristics of the people of India.



Population and Sample

Statistical Population:

An aggregate of objects or individuals under study is called population or universe. However many a times we record some quantitative or qualitative characteristic of each member in the population. These observations (or data) are collectively called as Statistical Population.

Population may contain finite or infinite elements. Accordingly it is called as finite or infinite population. e.g.

1) In the study of industrial development, all the industries under consideration are the population.
2) In the study of socio-economics conditions of a particular village, all families or houses in the village will be a population.
3) In titration experiment solution in beaker is a population.
4) In the study of bio-diversity all trees in the jungle is a population (infinite population)

Thus population may be a group of employees, collection of books, total industrial production, a group of persons suffering from a particular disease, group of students etc.

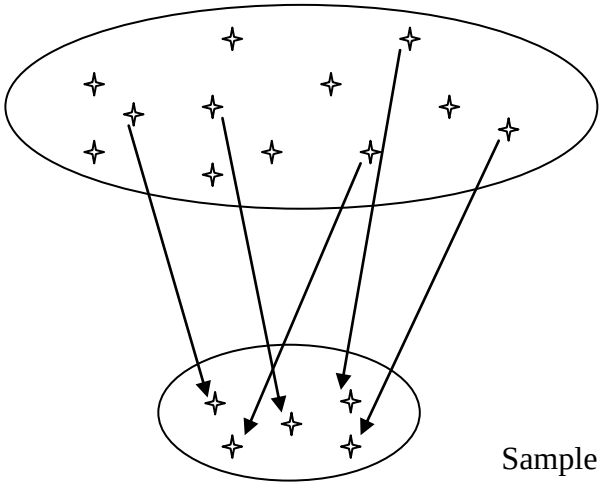
In order to study the population, one of the ways is to collect information about each and every element in the population. This method is called as census or complete enumeration.

After every ten years population census of India is conducted. In this census, information regarding every individual is collected.

Sample:

Any part of population under study is called Sample and method of selecting sample from population is called as sampling method. e.g.

1. While purchasing food grains, we inspect only a handful of grains and draw conclusions about the quality of the whole lot. In this case, handful of grains is a sample and the whole lot is population.
2. While examining blood of an individual a few drops are taken out of human body for diagnosis. These drops form a sample whereas entire blood in the body is a population etc.



Population

Sample

Census Method:

In order to study the population, one of the ways is to collect information about each and every element in the population. This method is called as census or complete enumeration. After every ten years population census of India is conducted. In this census, information regarding every individual is collected.

Sampling Method:

The manner or scheme of selecting sample from population is called sampling method. There are various methods used to select an unbiased sample from the population. Choice of the method depends upon the information available about the population, nature of data and the object of inquiry. Some of the important methods of sampling are Simple Random Sampling, Stratified Random Sampling, Systematic Sampling, Cluster Sampling, Two stage Sampling.

Advantage of sampling over census :

1) In sampling less number of elements and hence reduced processing time
2) In sampling method only a part is to be studied therefore expenses in collection of data and analysis are less than those in census. Thus it is economical.
3) In sampling method only limited number of elements is to be processed therefore accuracy can be increased.
4) If the population is too large and infinite then sampling is a greater scope.
5) If testing is destructive then sampling is only method.

Methods of Sampling:

There are various methods used to select an unbiased sample from the population. Choice of the method depends upon the information available about the population, nature of data and the object of inquiry. Some of the important methods of sampling are

a) Simple Random Sampling b) Stratified Random Sampling.
c) Systematic Sampling. d) Cluster Sampling.
e) Two stage Sampling.

a) Simple Random Sampling:

It is easiest and most commonly used method of sampling. In this method each element of population is given same chance of getting selected in the sample. If population consists of N elements then probability of selecting any element any draw is 1/N. The methods of random selection are Lottery method and Random Number method.

There are two types of Simple Random Sampling 1) Simple Random Sampling with replacement and 2) Simple Random Sampling without replacement.

1) Simple Random Sampling With Replacement(SRSWR):

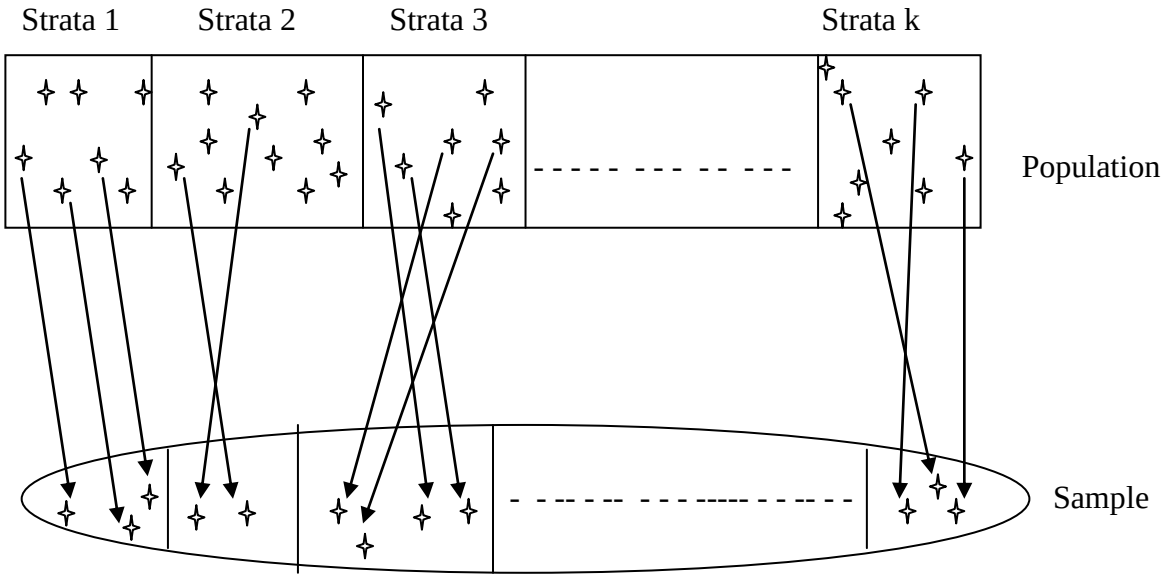
In this method, first element is selected at random from the population. It is recorded or studied completely and then replaced back in the population. Afterwards second element is selected similarly. This process is continued till a sample of required size is selected. In this method population size remains the same at every draw and the serious drawback is that the same element may be selected more than once in the sample.

2) Simple Random Sampling Without Replacement (SRSWOR):

In this method, elements are selected at random but those are not replaced back in the population therefore it called SRSWOR. In this method population size goes on decreasing at every draw and drawback of getting same element more than once in SRSWR is overcome in SRSWOR.

b) Stratified Random Sampling:

If the population is not homogeneous, SRS is not very effective. Therefore the entire population is divided into several homogeneous groups called as Strata. A simple random sample of a suitable size is selected from each stratum and then combining these sampled observations we can form a sample. The sample thus formed is called Stratified Random Sampling



C) Systematic sampling.

Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size. Systematic sampling can help researchers, including marketing and sales professionals, obtain representative findings on a huge group of people without having to reach out to each and every one of them.

Steps to Create a Systematic Sample

You can use the following steps to create a systematic sample:

1. **Define your population**: This is the group from which you are sampling.
2. **Settle on a sample size**: How many subjects do you want/need to sample from the population to get a reflective idea of it?
3. **Assign every member of the population a number**: If the group you're looking at consists of, say, 10,000 people, start lining them up and giving them numbers.
4. **Decide the sampling interval**: This can be achieved by dividing the population size by the desired sample size.
5. **Choose a starting point**: This can be done by selecting a random umber.
6. **Identify members of your sample**: If you have a starting point of 15 and a sample interval of 100, the first member of the sample would be 115, and so forth.

1.2. Nature of Data

Data:

The information collected for statistical investigation is called a data. It may be in the form of numbers, words, figures, symbols, question & answers etc.

There are two types of statistical data.

1) Primary data:

The data, which are originally collected by an investigator or agency for the first time for statistical investigation, are termed as primary data. Primary data are also called as raw data. Such data are original in character and more reliable.

e.g. Data obtained in a population census.

Primary data are collected by the following methods:

1) Direct personal observation or Interview.
2) Indirect Investigation. e.g. sleeping tablet-names–doctor, medical shopkeeper, STD
3) Questionnaires Method. – a) By post
   b) Through Enumerators
4) Through local correspondence. e.g. Newspaper agencies, Conditions of weather

2) Secondary data:

The data, which are not originally collected but rather obtained from published or unpublished sources, are know as secondary data.

e.g. Part of population census reproduced in another publications are secondary data.

Secondary data are collected from following sources:

1) Government publications
2) Publications of private organization
3) Publications of foreign Governments and International Bodies.
4) Journals, Magazines etc.

Difference between Primary and Secondary data:

- The main difference lies in the method of collection
- Primary data are original in nature. Hence those are more accurate than Secondary data
- Collection of Primary data is expensive as well as time consuming
- Primary data can be collected in accordance with objectives of study. Secondary data may fail in this regard.

Time Series Data:

Time-series data is a sequence of data points collected over time intervals, allowing us to *track changes over time*. Time-series data can track changes over milliseconds, days, or even years.

Having access to detailed, feature-rich time-series data has become one of the most valuable commodities in our information-hungry world. Businesses, governments, schools, and communities, large and small, are finding invaluable ways to mine value from analyzing time-series data.

Time Series Data Examples are Weather records, patient health metrics, economic indicators; these are all time series data.

Qualitative Data:

If data is collected according to such well-defined qualitative characteristic (Attribute), then such data is called qualitative data. Qualitative data cannot be measured.

e.g. Attributes such as Beauty, Honest, Religion, Grade in examination, etc. gives qualitative data .

There are two scales of measuring attributes such data:

1) Nominal Scale: -

Nominal Scale consists of two or more named categories into which the objects are classified. e.g. a) Classification of students in various divisions of the same standard

b) Classifications of individuals using blood groups.

c) Classification of individuals using Sex, Cast.

d) House numbers, Survey numbers and Pin code numbers

Remark:

1) In nominal scale, if numbers are used then those are allotted in purely arbitrary manner. Those numbers are just for identification purpose used in place of labels.
2) Those numbers are interchangeable.

2) Ordinal Scale:-

Ordinal scale of measurement gives number to groups of objects using some quantifiable characteristics. Therefore ordered arrangement of groups is possible in this type of scale. e.g. a) Groups of individuals according to income such as poor, middle class, rich.

b) Groups of students according to grades in exam. Such as Second class, first class, Distinction.

c) Groups using weight such as light, heavy.

d) Groups using height such as short, medium, tall.

Remark:

1) In ordinal scale, numbers given to groups as labels, serve the purpose of ranks. Hence those labels are not interchangeable.
2) In the ordinal scale, the groups are ordered according to some characteristic.
   e.g. A –Shortest – 1 , B – Medium – 2 , C – Tallest – 3
   Height of B is not double the height of A.

Quantitative data:-

If data is collected according to quantitative characteristic (which changes its value), then such data is called quantitative data. Qualitative data are measurable data.

e.g. Variables such as height, weight, age, income etc. gives Quantitative data.

There are two scales of measuring variables such data:

1) Interval Scale:-

Interval scale of measurement has equal units of measurement, however, the zero point is arbitrary. We measure the level of certain phenomenon and not he quantity.

e.g. Centigrade scale of temperature measurement, Marks of the candidates in elocution competition.

2) Ratio Scale:-

Ratio scale of measurement has well defined units starting with zero. We measure the quantity and not the level.

e.g. height, weight, time, age, etc. are measured in ratio scale.

Discrete Variable:-

A variable taking only particular values is called as discrete variable. Most of the discrete variables have integer values.

e.g. Number of students in a class, Population of a city, Number of workers in factory, Number of members in family etc.

Continuous Variable:-

A variable taking all possible value in a certain range (i.e. two appropriate numbers) is called as continuous variable.

e.g. Weight of person, height of person, temperature at a certain place, speed of a vehicle etc.

Presentation of data

Classification of data:-

The process of arranging data in different groups according to similarities is called as classification of the data. The groups so formed are called classes. classification can be used as a tool to condense the data.

Methods of classification:-

There are two methods of classification.

1) Inclusive Method :

In this method classes are so formed that a both the limits, upper limit and lower limit of the class are included in the same class. e.g.

Income in Rs.
200 – 299
300 – 399
400 – 499
500 – 599
600 – 699

In this example a person whose income is Rs. 299 is included in the class Rs. 200 – 299

2) Exclusive Method :-

In this method, the upper limit of the class is not included in that class but it is included in the next class. In this method the upper limit of a class is same as the lower limit of the next class. e.g.

Income in Rs.
200 – 300
300 – 400
400 – 500
500 – 600
600 – 700

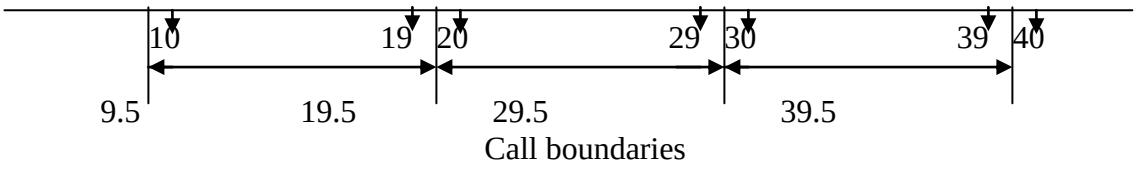
In this example a person whose income is Rs. 300 is not included in class 200 – 300. It is included in the class 300 – 400

Class Boundaries :-

The class boundaries are the numbers up to which the actual magnitude of the observation in the class can extend. The class boundaries are also called as actual limits or extended limits. e.g.

Class limits	Class boundaries
10 – 19	9.5 – 19.5
20 – 29	19.5 – 29.5
30 – 39	29.5 – 39.5

In case of exclusive method of classification, class limits and class boundaries are same.



Call boundaries

Class Mark or Mid values or Mid point :-

$$\text{Class mark} = \frac{\text{Upper limit} + \text{Lower limit}}{2}$$

Class Interval :

The range between the lower and upper limit of a class is called the class interval. e.g.

Weight in Kg.
40 – 50
50 – 60
60 – 70

In this example 40 – 50 is the first class interval, 50 – 60 is the second class interval and 60 – 70 is third class interval.

Class Width or Length of the class :-

It is the actual length of the class interval. We can find class width as follows

Class Width = Upper boundary – Lower boundary

Open –End Class :-

A class in which one of the limits is not specified is called an open-end class.

e.g.

Open-end classes

Daily Sale in Rs.
Below 3000
3000 – 4000
4000 – 5000
5000 & above

Class frequency :-

The number of items, which belong to the same class is called the class frequency. e.g.

Age group	No. of Student
18 – 20	200
20 – 22	380
22 – 24	190

In this example the number of students in the class 18 – 20 is 200. So the frequency of the class 18 – 20 is 200.

Frequency Distribution:-

The way in which the items are spread out or distributed into various classes is called the frequency distribution. The table in which these frequencies are shown is called as frequency distribution table or simply frequency distribution.

There are two types of frequency distribution

1) Discrete frequency distribution. 2) Continuous frequency distribution.

1) Discrete frequency distribution :-

If the variable considered is discrete then it is called discrete frequency distribution. In discrete frequency distribution how many times a particular value is repeated is counted and this number is the frequency of that value.

Some time discrete frequency distribution is called ungrouped frequency distribution.

2) Continuous frequency distribution :-

If the variable considered is continuous then it is called continuous frequency distribution. In continuous frequency distribution we count how many values fall into the same class and this number is the frequency of that class. (refers grouped frequency distribution)

Note:- In Continuous frequency distribution classes must be exclusive classes.

Cumulative frequency: -

If we added the frequencies up to the given value or above the given value, the added frequency is called cumulative frequency.There are two types of cumulative frequencies.

1) **Less than c. f. :-**
   If we add frequencies up to a given value or class then it is called less than c.f. L.c.f. of a class is the number of observations less than upper limit of the corresponding class.
2) **Greater (or more) than c.f. :-**
   If we add frequencies above a given value or class then it is called greater than c.f.
   g.c.f. of a class is the number of observations greater than or equal to lower limit of the corresponding class.

Cumulative frequency distribution :-

Table showing the cumulative frequencies corresponding to the class limits is known as cumulative frequency distribution.

e.g.

Marks	frequency	Less than c.f.	Greater than c.f.
10 – 20	8	08	42 + 08 = 50
20 –30	8	08 + 08 = 16	34 + 08 = 42
30 – 40	15	16 + 15 = 31	19 + 15 = 34
40 – 50	11	31 + 11 = 42	08 + 11 = 19
50 – 60	8	42 + 08 = 50	08
	50		

The above example shows that there are 8 students with marks less than 20, 16 students with marks less than 30 and so on .

Similarly 8 students with marks greater than 50, 19 students with marks greater than 60 and so on. this shown as follows

Marks	L. c. f.	Marks	G. c. f.
Less than 20	08	Greater than 10	50
Less than 30	16	Greater than 20	42
Less than 40	31	Greater than 30	34
Less than 50	42	Greater than 40	19
Less than 60	50	Greater than 50	08