Module I

Definition of Statistics

Statistics is a Science of decision making on the basis of sample observations drawn from a population under uncertainty. That is it is a mathematical discipline Concerned with the collection of data, summarization of data, analysis of data and interpretation of data toward a valid decision.

population

A set or group of observations relating to a phenomenon under statistical investigation is known as statistical population. The term "population" implies an aggregate or collection of measurements on a given variable. Ex → measurements of heights in your college

Sample :-

Sample is a subset of population or a true representation of population.

Ex :- measurements of heights of students of Statistics department in your college.

Types of Data :-

Quantitative and Qualitative data :-

By Quantitative statistical data we mean a sequence or a set of numerical measurements made on some of the objects in a specified population.

Ex the heights of 5 students of a college constitute ~~quantitative~~ quantitative data and the quantitative data are: 5'6", 5'5", 5'4", 5'8", 5'7".

By qualitative data we mean a set of observations in which each observation in the sample or population belongs to one of several mutually exclusive classes which are likely to be non-numerical.

The 'colour of flower' can be classified as red, blue, white and others. the colour of ten flowers in a garden are recorded as R, W, O, B, R, W, W, R, O, R. The data is qualitative data and the character "colour of flower" is a qualitative data.

Discrete and Continuous Data

When we study the data regarding quantitative characters, it is found that this may be of two types

a) Discrete variable, the character may take only some isolated values, like the number of members in a family.

b) Continuous variable, the character can take any value within its ~~range~~ range of variation. The height, weight of man are continuous variable.

Time Series and Cross-sectional data.

Time series data

When data are arranged according to the order of time, the data is known as time series data.

Ex:

Time (in years)	Production of Rice
1990	10
1991	11
1992	11.5
1993	12
1994	10.5
1995	12
1996	13

Cross-sectional data

It is a type of data which is collected by observing many subjects (such as individuals, firms, countries) at the same or approximately the same time point, or ~~cos~~ without regard to differences in time.

Example Here we study the changes in the value of the variable from a region to region

States	production of Rice.
Bihar	10
W.B	16
Orissa	12
UP	14
MP	12

Primary Data :-

The Statistical Data which are gathered directly from the field of Investigation for the desired purpose are called primary data.

Ex:- When a doctor, interested in the weight of his patients, records their weights using a machine, the data are primary Data.

Secondary Data

The statistical data which have already been collected by some agency and are compiled from that source by the enquirer for his use are called secondary data.

Ex:- the Census data collected from Census reports by a research scholar for his study are secondary data to the Scholar.

Measurement of scale :-

(1) Nominal Scale :-

Measurement as its weakest level exists when numbers or others symbols are used to classify an object, person or characteristic. when numbers or others symbols are used to identify the group to which various objects belong to, these numbers or symbols constitute a nominal scale.

Ex:- The numbers on automobile license plates constitute a nominal scale.

Ordinal Scale

It may happen that objects in one category of scale are not just different from the objects in other categories of that scale, but that they stand in some kind of relation to them.

Ex:- In prestige or social acceptability, all members of the upper middle class are higher than (>) all members of the lower middle class.

Ratio Scale

When a scale has all the characteristics of an interval scale

Interval Scale:-

When a scale has all the characteristics of an ordinal scale and when in addition the distances between any two numbers on the scale are of known size.

Ex:- We measure temperature on an interval scale. The unit of measurement and the zero point in measuring temperature are arbitrary- they differ for two scales: Centigrade, Fahrenheit.

The Ratio Scale.

When a scale has all the characteristics of an interval scale and in addition has a true zero point as its origin, it is called a ratio scale.

Ex:- we measure mass or weight in a ratio scale.

Presentation of Data

1. Textual Presentation of Data
2. Tabular " " "
3. Graphical " " "

Textual

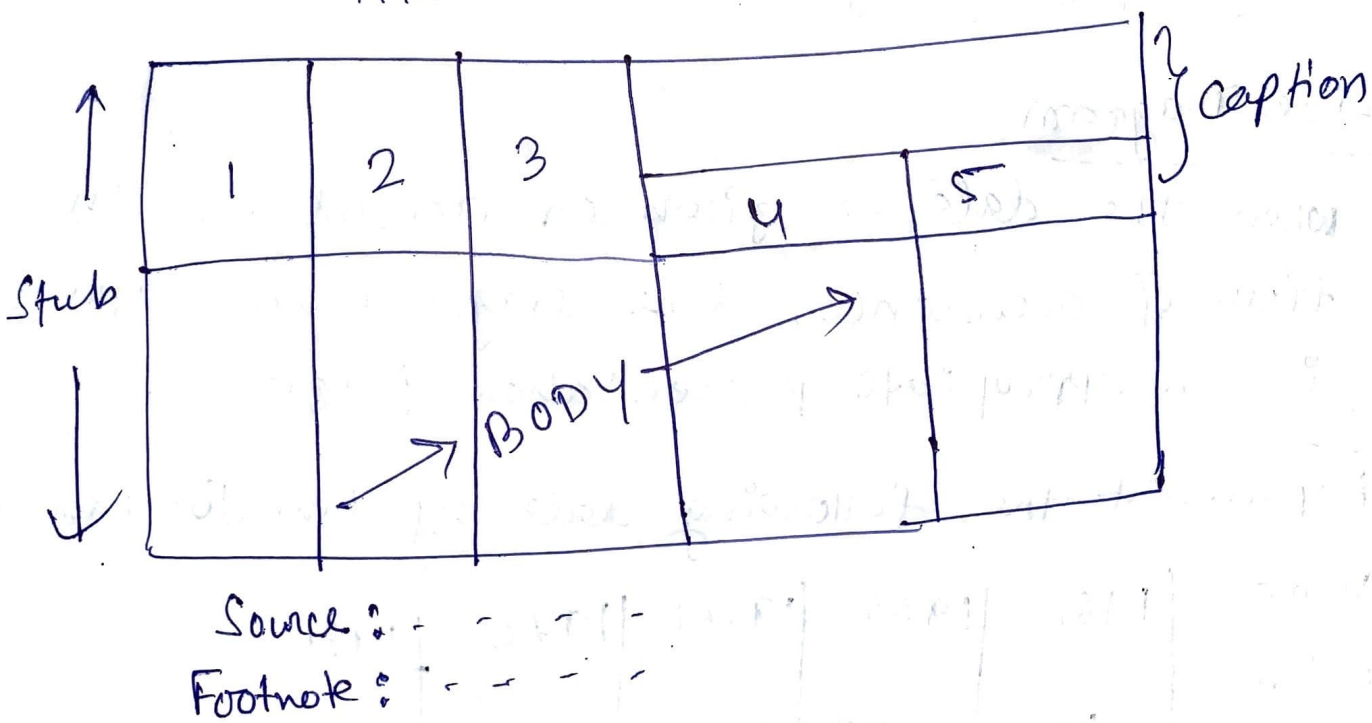
The data is mentioned in the text form which is usually written in a paragraph. The textual presentation of data is used when the data is not large and can be easily comprehended by the reader just when he reads the paragraph.

Tabulation of Data

Tabular presentation of data is an useful mode of exhibiting the data in a compact form. The systematic presentation of data in the structure of a table comprising some rows and ~~colu~~ column is called Tabulation

A table has several parts which are described below :-

1) Table Number (2) Title (3) Stub (4) Caption (5) Body (6) Footnote (7) Source.



Graphical presentation of Data

Graphs, charts, maps, pictures etc. are attractive and effective means of presentation of statistical data. Diagrams are readily capable of revealing some features of the exhibited Data. It should be noted that the selection of the appropriate diagram depends mainly on the nature of the given Data.

Some common types of charts generally used are

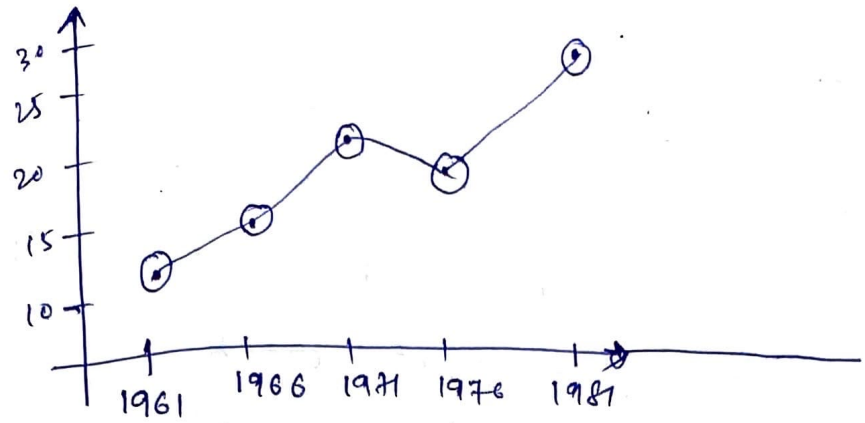
(1) Line Diagram.
(2) Bar Diagram.
(3) Pie Diagram.

Line Diagram

When the data is given in accordance with time of occurance, line diagram or graph is an appropiate presentation of Data.

Ex:
Represent the following data by line diagram.

Years	1961	1966	1971	1976	1981
Population (in lakhs)	12	15	21	20	27.

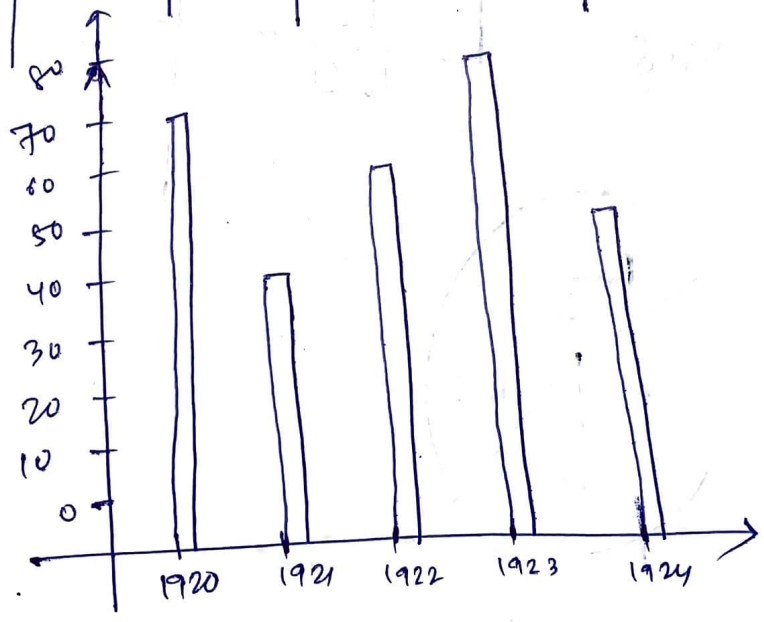


Bar diagram

This mode of representation comprises a no. of equidistant rectangles (termed as Bars), each of them being meant for some specific category of the available data. Bars of common width are drawn on the base line, the length or height of a bar representing the value of the corresponding category.

Ex:-

Year	1920	1921	1922	1923	1924
Production (tons)	70	40	60	80	50.

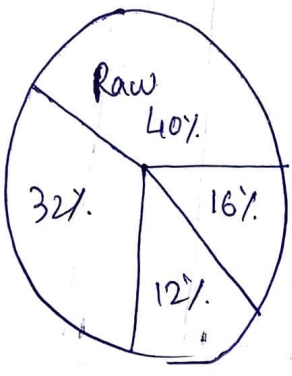


Pie diagram

A pie diagram is another appropiate diagram used for exhibiting the relative sizes of the different parts of a whole. In this case, a circle is partitioned into several sectors by drawing angles at the centre, the area of each sector indicating the corresponding percentage. Since the total angle at centre 360°, the desired angle for some particular category will be 3.6 times the relevant percentage.

Ex:-

Source	production	Percentage	Angle
Raw materia	192	40	144°
Labour	153.6	32	115.2°
Direct production	57.6	12	43.2°
Others	76.8	16	57.6°
		100	360°



Frequency Distribution

In Statistics, it is much needed to arrange the sample values in a tabular form which is generally known as Frequency Distribution.

Frequency distributions are of two types

i) Simple frequency distribution

~~ii) Grouped~~

Ex:-

No. in Maths (x_i)	Frequency (f_i)
30	2
33	1
35	3
36	2
38	1
40	1

ii) Grouped Frequency distribution.

Ex:-

Age in years	No. of Mens.
20-29	8
30-39	3
40-49	5
50-57	7.

Terms related to grouped frequency

① class interval:- when the number of sample values are large in number and are in wide range, then the whole data is divided in a number of several groups according to the size of the sample. Each of these groups, known as class interval.

Class limit

In any frequency table, if the both ends of any class interval are specified, then these end values are called class limits.

The lower end → lower class limit

The upper " → upper " "

Class boundary

Lower class boundary = Lower class limit $- \frac{1}{2}d$

Upper " " = upper class limit $+ \frac{1}{2}d$.

d = difference between upper limit of any class and the lower class limit of the next class.

Class mark

$$\text{Class Mark} = \frac{1}{2}[\text{upper class limit} + \text{lower class limit}]$$

or

$$= \frac{1}{2}[\text{upper class boundary} + \text{lower class boundary}]$$

Frequency Density

~~Frequency density~~ =

Class width

class width = Upper class boundary − lower class boundary

Frearuency Density :-

$$\text{Frearuency Density} = \frac{\text{clam Frearuency}}{\text{clam width.}}$$

Relative Frearuency

$$\text{Relative Frearuency} = \frac{\text{clam Frearuency}}{\text{Total Frearuency.}}$$

Cumulative Frearuency Distribution.

Ex:-

clam interval	Frearuency	C.F (less than type)	C.F (more than)
		0	60
10-19	6	6	60-6 = 54
20-29	10	10+6 = 16	54-10 = 44
30-39	12	16+12 = 28	44-12 = 32
40-49	15	28+15 = 43	32-15 = 17
50-59	10	43+10 = 53	17-10 = 7
60-69	4	53+4 = 57	7-4 = 3
70-79	2	57+2 = 59	3-2 = 1
80-89	1	59+1 = 60	
	60		

Graphical representation of Frearuency Distribution

There are three important diagrams to depict statistical data.

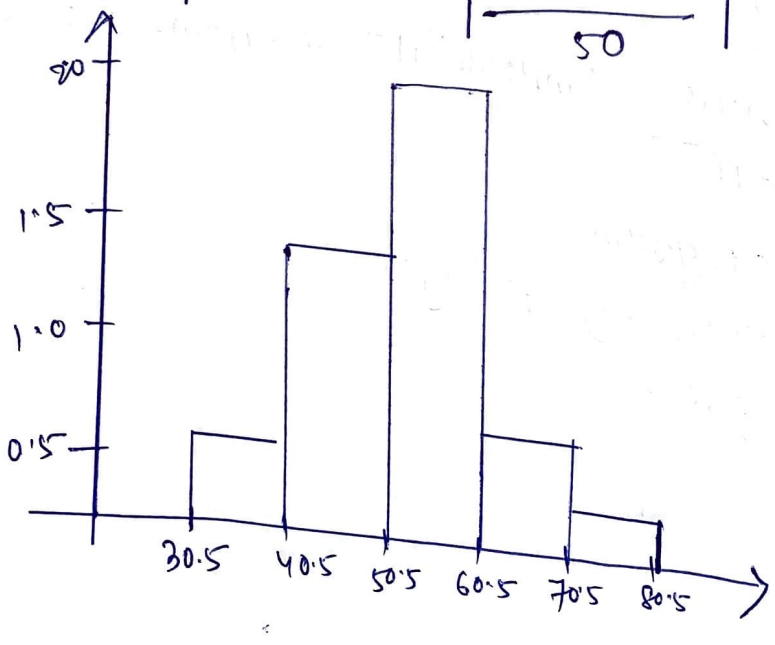
1. Histogram
2. Frearuency polygon.
3. Ogive.

Histogram

It is an appropiate diagram for representing the frequency distribution of a continuous variable in the sense that it considers the fact that the frequency of a class is dispersed over the interval. To construct histogram class-boundaries are plot on horizontal axis. Next a Rectangle is drawn over each class-interval so that its area indicates the corresponding class frequency. In other words, the height of a Rectangle becomes equal to corresponding frequency density.

Ex:-

Class limit	Class boundary	Frequency	Frequency Density
31-40	30.5-40.5	6	0.6
41-50	40.5-50.5	14	1.4
51-60	50.5-60.5	20	2
61-70	60.5-70.5	7	0.7
71-80	70.5-80.5	3	0.3
		50	



Freequency polygon.

A frequency polygon may be drawn to exhibit the frequency distribution of a continuous variable, provided the classes are of equal width. To construct frequency polygon, in the x-axis we plot class boundaries and in y-axis frequencies are plotted against the class marks of the respective classes. The plotted points are then joined succesively by line segment.

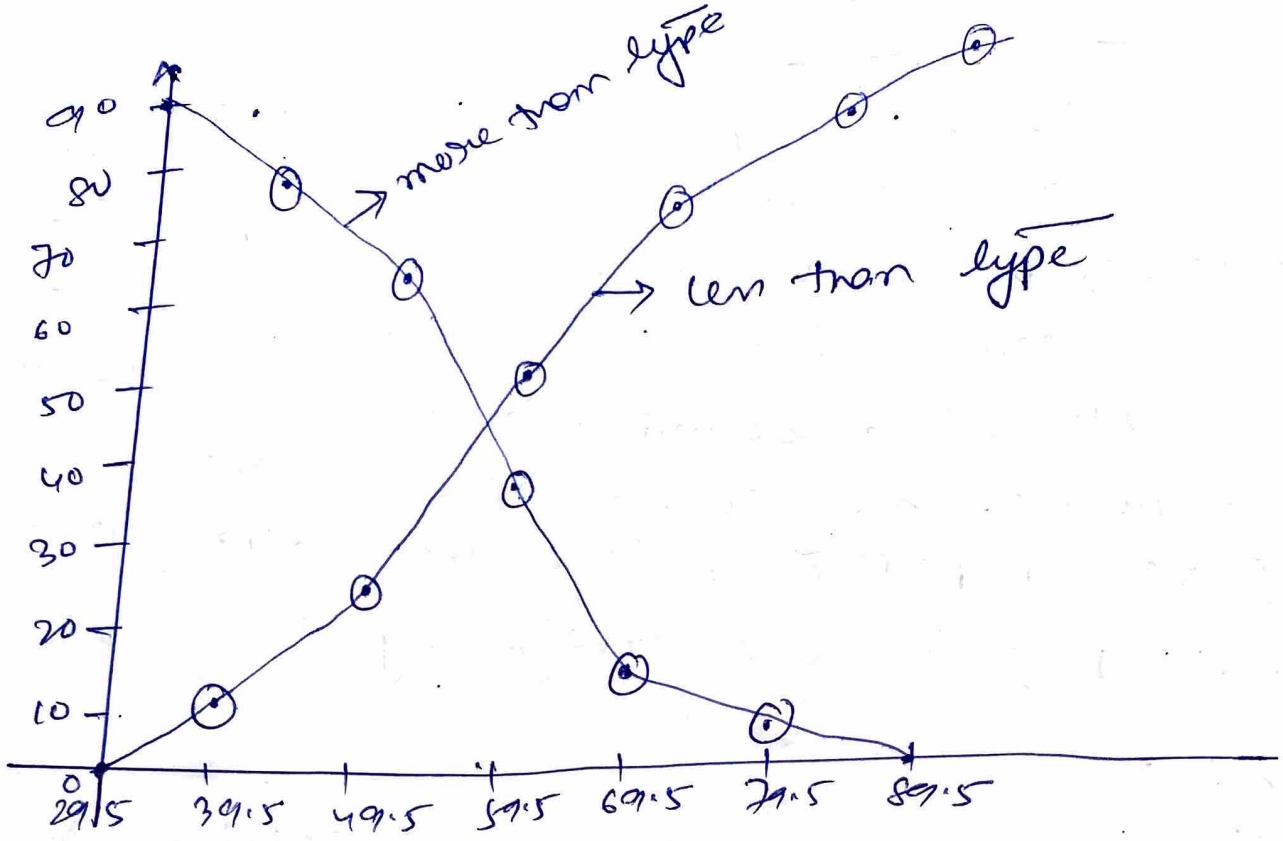
Ogive

This diagram is used for exhibiting the frequency distribution of a continuous variable in terms of cumulative frequency. (of either type).

To draw the less than type ogive, we plot the less than type C.F against the upper class boundaries of the classes and join them by straight line. Similarly, to get the more than type ogive, the more than type C.F are plotted against the lower class boundaries and points are join by straight line.

Ex:-

Marks	No. of Students	C.F(<)	C.F(>)
		12	90
29.5-39.5	12	27	78
39.5-49.5	15	54	63
49.5-59.5	27	79	36
59.5-69.5	25	87	11
69.5-79.5	8	90	3
79.5-89.5	3		


90
80
70
60
50
40
30
20
10
0
29.5
39.5
49.5
59.5
69.5
79.5
89.5
more than type
less than type