Statistics : statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis which often leads to the drawing of conclusions.

Two Major branches of statistics are

(1) Descriptive statistics.

(2) Inferential statistics.

Descriptive statistics: The part of statistics concerned with the description and summarization of data is called descriptive statistics.

Inferential statistics: The part of statistics concerned with the drawing of conclusions from data is called inferential statistics.

Descriptive Statistics: If the purpose of analysis is to examine and explore information for its own intrinsic interest only then the study is descriptive.

Statistics relies on data. In order to learn something we need data.

Data: Data are the facts and figures collected, analyzed and summarized for presentation and interpretation.

Examples: (1) To know the percentage of marks obtained by students

(2) To know how many people like a new song/ product

Collected through comments.

Data may be (i) available

(ii) need to collect

(iii) generate data.

Here we assume data is available. and our objective is to do a statistical analysis of available data. Data is of two types.

(1) Categorical data (Also called qualitative variables)

(2) Numerical data. (Also called quantitative variables)

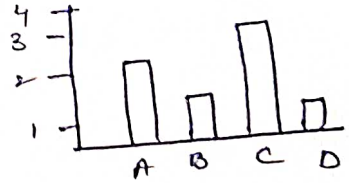
The two most common displays of a categorical variable are bar chart and Pie chart.

Both describe a categorical variable by displaying its frequency table.

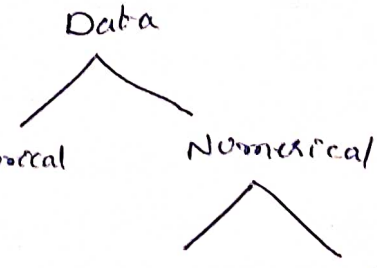
Definition:

Pie chart: A pie chart is a circle divided into pieces (wedges) proportional to the relative frequencies of the qualitative data

Bar chart: A bar chart is used to show the frequencies / relative frequencies of a categorical value. A bar chart displays the distinct values of the qualitative data on a horizontal axis and the relative frequencies of those on a vertical axis. The frequencies of each distinct value is represented by a vertical bar In a bar chart we know the count.



But In this course we are dealing with Numerical data. Numerical data is divided into two parts. Discrete data and continuous data.



Example:

Discrete: Marks obtained by students.

Continuous: weight of students

Organizing Numerical data

→ A discrete variable usually involves a count of something where as a continuous variable usually involves a measurement of something.

→ First group the observations into classes (also known as categories) and then treat the classes as the distinct values of qualitative data.

→ Once we group the quantitative data into classes we can construct frequency and relative frequency of the data.

Organizing discrete data (single value)

→ If the data set contains only a relatively small number of distinct or different values it is convenient to represent it in a frequency table we create each distinct value as a category.

→ Each class represents a distinct value (single value) along with its frequency of occurrence.

Example: Suppose the data set reports the no of people in a household. The following data is the response from 15 individuals.

2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4

The distinct values of the variable, no of people in each house hold as 1, 2, 3, 4, 5.

Frequency distribution table.

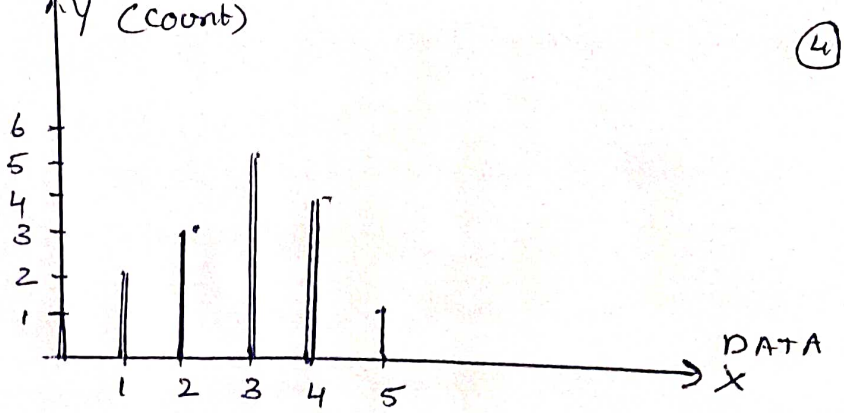
Value	Tally marks	Frequency	Relative frequency
1	II	2	0.133
2	III	3	0.2
3	IIII	5	0.333
4	IIII	4	0.267
5	I	1	0.067

Plot the data

We can see that there is an order on data because it is a numerical data.

This is called bar chart.

Observe that bars are not connected.



When observations discrete or continuous are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. Let us consider the marks in Statistics by 50 candidates.

32	47	41	51	41	30	39	18	48	53
54	32	31	46	15	37	32	56	42	48
38	26	50	40	38	42	35	22	62	51
44	21	45	31	37	41	44	18	37	47
68	41	30	52	52	60	42	38	38	34

This representation of the data does not furnish any useful information and is rather confusing on mind.

A better way may be to express in figures in an ascending or descending order of magnitude commonly termed as array. But this does not reduce the bulk of data. A much better representation is given in table.

Marks (x) Class intervals.	No of students (Tally marks)	Total frequency
15 – 19	III	3
20 – 24	II	2
25 – 29	I	1
30 – 34	𝍸 III	8
35 – 39	𝍸 IIII	9
40 – 44	𝍸 𝍸	10
45 – 49	𝍸 I	6
50 – 54	𝍸 II	7
55 – 59	I	1
60 – 64	II	2
65 – 69	I	1
		50

A bar | called tally mark is put against the number when it occurs. Having occured four times, the fifth occurrence is represented by putting (/) on the first four tallies. This technique facilitates the counting of the tally marks at the end.

Such a table showing the distribution of the frequencies in the different classes is called a frequency table and the manner in which the class frequencies are distributed over the class intervals is called the "grouped frequency distribution" of the variable

Remark: The classes of the type in which both the upper and lower limits are included are called "inclusive classes." and the classification is termed as "inclusive type classification"

Few guide lines that need to be followed. when organising the continuous data in to a no of classes to make the data understandible.

1. Number of classes The appropriate number is a subjective choice The rule of thumb is to have between 5 and 20 classes.
2. Each observation should belong to some class and no observation should belong to more than one class.
3. It is common, although not essential to choose class intervals are of equal length.

Continuous Frequency Distribution: If we deal with a continuous variable it is not possible to arrange the data in the class intervals of above type (inclusive type classification)

Let us consider the distribution of age in years. If class intervals are 15-19, 19-20 etc, then the persons with age between 19 and 20 years are not taken in to consideration. Therefore we define class intervals as

10 – 20	(Including 10	excluding 20)
20 - 30	" 20	" 30
30 – 40	" 30	" 40

This form of the frequency distribution with such classes as known as continuous frequency distribution. It should be clearly understood that in the above classes, the upper limits of each class are excluded from the respective classes. Such classes in which the upper limits of each class are excluded from the respective classes and are included in the immediate next class are known as "exclusive classes" and the classification is termed as "exclusive type classification".

Graphic representation of a frequency distribution.

1. Histogram
2. Frequency Polygon
3. Ogive Curves.

Consider the frequency distribution.

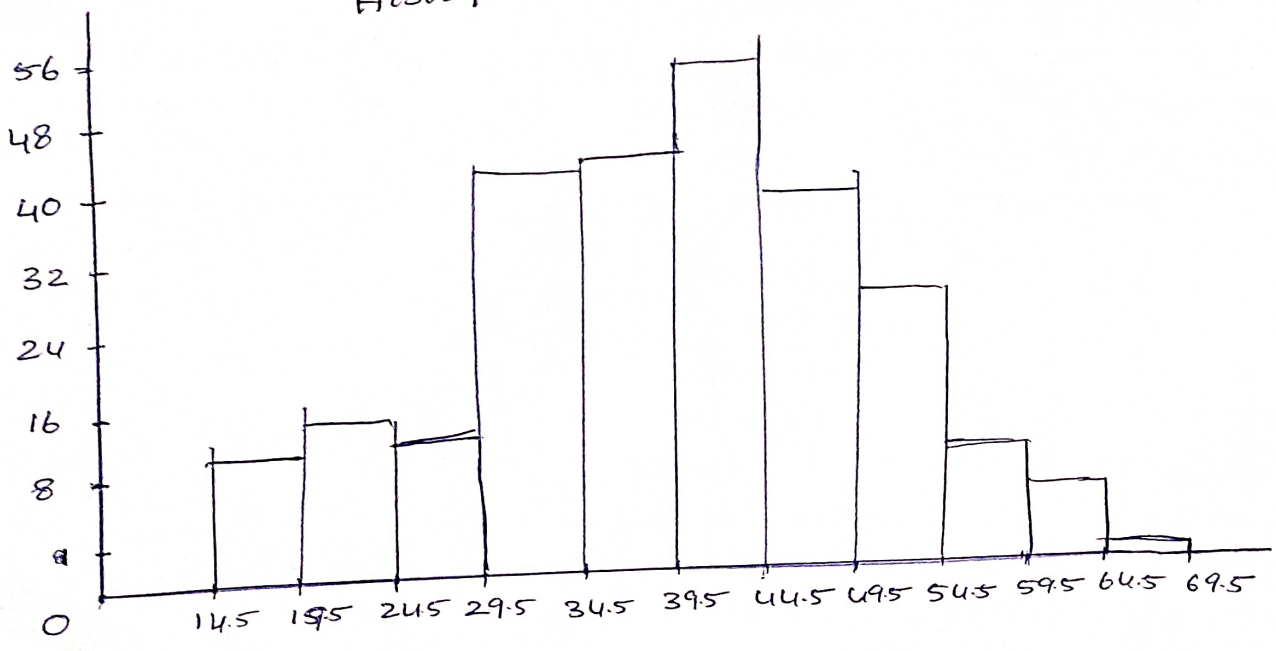
Marks	No of students
15 – 19	9
20 – 24	11
25 – 29	10
30 – 34	44
35 – 39	45
40 – 44	54
45 – 49	37
50 – 54	26
55 – 59	8
60 – 64	5
65 – 69	1

Since this grouped frequency distribution is not continuous we first convert it into a continuous distribution with exclusive type classes.

Marks	No of students
14.5 – 19.5	9
19.5 – 24.5	11
24.5 – 29.5	10
29.5 – 34.5	44
34.5 – 39.5	45
39.5 – 44.5	54
44.5 – 49.5	37
49.5 – 54.5	26
54.5 – 59.5	8
59.5 – 64.5	5
64.5 – 69.5	1

Note: The upper and lower class limits of the new exclusive type classes are known as class boundaries

Histogram for frequency distribution



Histogram is the one of the most popular graphical summary of a continuous data. How to set up a Histogram.

Step 1: Obtain a frequency distribution of the data

Step: 2 Draw a horizontal axes on which to place the classes and a vertical axes on which to display the frequencies

Step 3: For each class, construct a vertical bar whose height equals the frequency of that class.

Difference between Histogram & barchart: Because class intervals are continuous there is no gap between bars continuous display of data: Vertical height of the bar represents the count on every class interval