Chapter X

DESCRIPTIVE STATISTICS

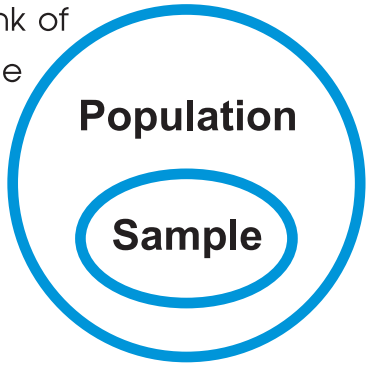
Learning Objectives:

After completion of this unit the students will be able to

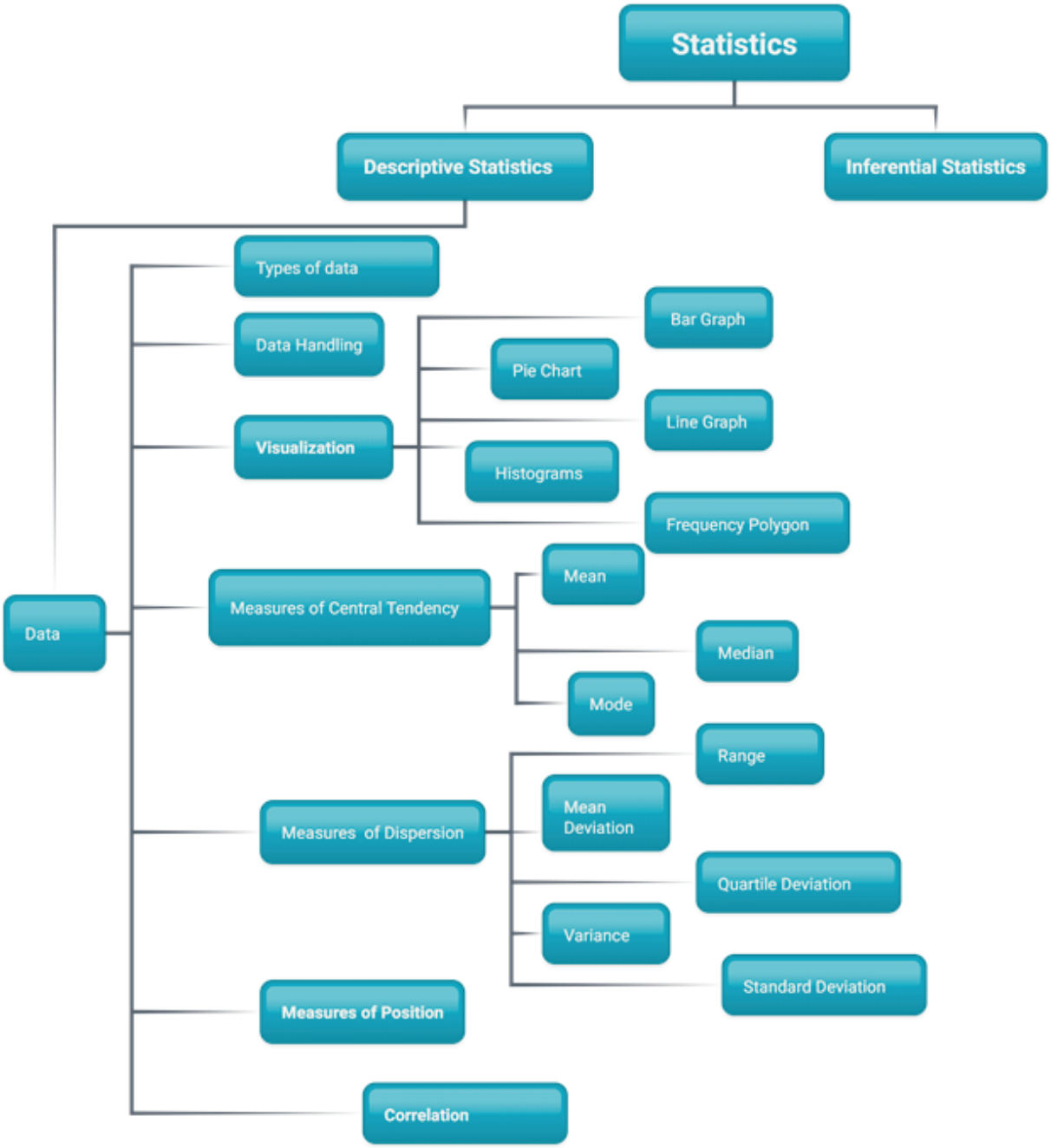
- develop an understanding of everyday data
- understand the organization, visualization and analysis of data
- create and analyze the graphical displays to summarize data
- draw meaningful conclusions from data
- make comparison among two distributions
- translate real world problems and make meaningful inferences out of it

Before you start you should know about "The population and samples."

In Statistics, population is not the same as we usually think of people in our village, town, state or country. It includes all the members of a defined group that we deal for making decisions whereas a sample is the subset of population.



Concept Map



Activity for recap

Activity 1: Suppose your class of 25 students had appeared for a test and marks obtained were;

0, 1, 1, 1, 2, 2, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 7, 7, 7, 7, 8, 8, 9, 10, 10

- How can you display and organize this given set of data?
- What conclusions can teacher draw from this data?
- What is the average of the data?

10.1 Introduction

In earlier classes, you have studied some basic concepts of statistics so you may be familiar with this through various kinds of data in your day to day life like newspaper, books, surveys, observations, etc. It plays an important role in our life for understanding the up-down in economy, population growth, climate change, weather prediction, election results and many more. We use statistics without realizing it. Every day you speak so many statistical statements unknowingly.

The facts or numerals collected for a specific purpose is called data. The word data is plural so a data set is always a group of many numbers.In fact data speaks or we can say that data can be transformed into useful information. Let us have a look on some of the newspaper headlines

(i) India to become a 5 trillion dollar economy by 2025

(ii) India's population to surpass that of China around 2024: UN

(iii) States' tax share to stay at 42% : XV Finance Panel

(iv) Inflation may drop to 2.4% in financial year 2021: RBI

We can find some sort of data in the above mentioned headlines. These data are based upon some or other information that can be verified using appropriate mathematical tools. Hence the science of collecting, studying and analyzing numerical data is known as statistics.

This is basically categorized into two branches. First, Descriptive Statistics deals with the collecting, summarizing and interpreting the data so that it can be easily understood. Second, Inferential Statistics deals with the methods for obtaining and analyzing data to make inferences by performing estimations and hypothesis tests. In this class we shall study only descriptive statistics; the other part of inferential statistics will be taken up in class XII.

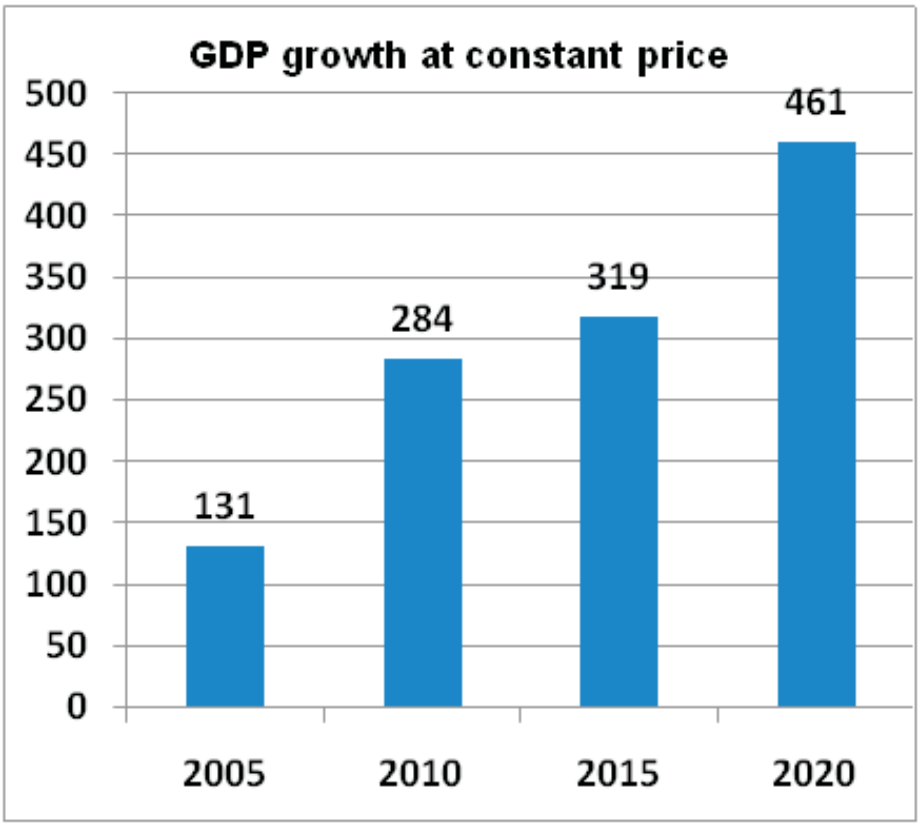
Note

- The word data is plural so a data set is always a group of many numbers. One piece of information is called a "datum"
- Statistics is concerned with the
  - (i) Collecting data
  - (ii) Organizing data
  - (iii) Summarizing data
  - (iv) Analyzing data
  - (v) Drawing conclusions / making inferences from data
- Descriptive statistics are just descriptive. They do not include generalization beyond the data.

Historical Note

Descriptive Statistics is considered to be originated during the census taken by the Babylonians and Egyptians (4500-3000 BC). They used to maintain the record for number of livestock each person owned and the crops each citizen harvested yearly.

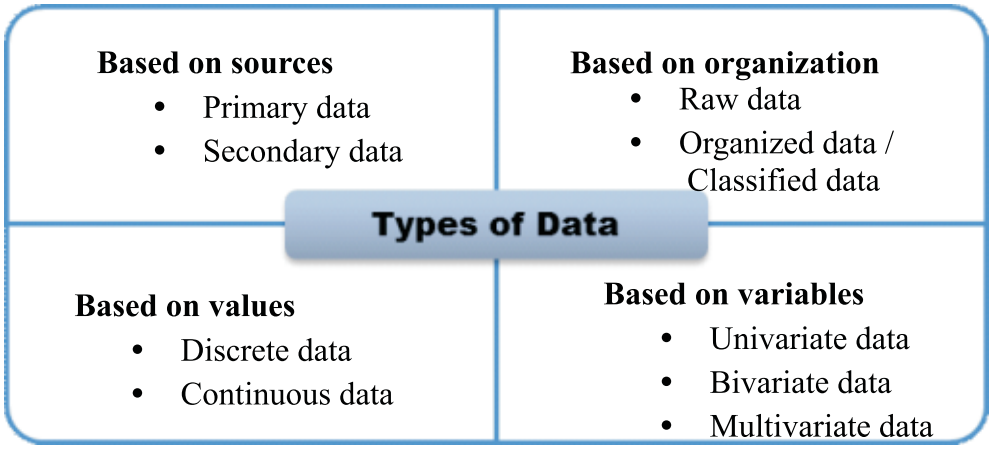
The figure given on next page gives India's Foreign Exchange Reserves (in US billion dollar) in recent years, and it make us ponder how the numbers are increasing regularly. Explanation for this rise may be anything but these descriptive statistical data offer an insight.



Source: Weekly statistical supplement of RBI

https://m.rbi.org.in/Scripts/WSSViewDetail.aspx?TYPE=Section&PARAM1=2

10.2 Types of data



Data based on sources (Primary / Secondary data)

Based upon the sources of collection, data can be either Primary or Secondary. If we collect it directly from the field or from individual (surveys, observations, case

studies), the data is called Primary Data. If it is taken from already collected data then such data is called Secondary Data.

Example 1: The teacher collected the information about the height of students in her class.

If teacher collects the data by measuring height of each student then data so collected is called the Primary Data, whereas, if teacher refer medical records of the students to note the height of students is called Secondary Data.

Data based on organization (Raw / organized data)

When we first collect the data, it is mostly unorganized and it is very difficult to make any meaningful observations on it. This first hand collected data is called Raw Data.

When we organize the raw data in order to draw any meaningful conclusion out of them, it is called organized/classified data.

Raw data: Raw data is unclassified or highly disorganized data. They require suitable organization in order to draw any meaningful conclusion out of them. For example, ODI run secured by Virat Kohli in the year 2019 given in below table is raw data.

3	104	46	45	43	60	44	116	123	7	20	18	82
77	67	72	66	26	34	1	120	114	4	0	85	23

Source: http://www.howstat.com/cricket/statistics/players/PlayerProgressBat_ODI.asp?PlayerId=3600&Year=2019

This data needs to be organized in a proper order before any systematic statistical analysis is undertaken.

Note: Raw data is also known by primary data/first hand data.

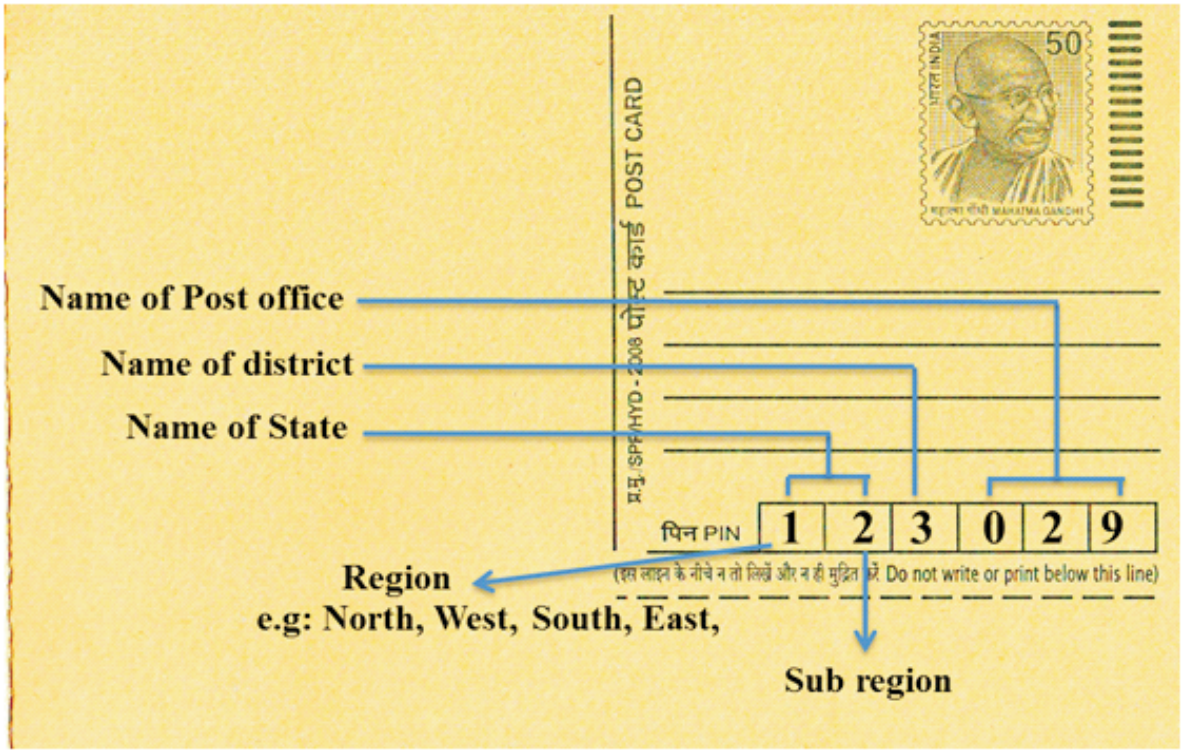
Question: Do you agree that classified data is better than raw data? Explain with an example from your daily life.

Activity 2

- Collect data of your daily pocket expenditure for a week/month and put it in a table. Find the number of observations and arrange the daily data.

Activity 3

PIN (Postal Index Number) code is the post office numbering code system used by the postal service of India. It is a 6 digits long code with each of the digits denoting a particular meaning. Here is how your PIN code is decided.



Explore the concept, how categorizing the areas with Pin codes helps postman to sort the letters and deliver at exact place.

Data based on variables (Univariate, Bivariate, Multivariate data)

(I) Univariate data involves only one variable (as Uni means one), for example, height of all the students in your class. Thus, their analysis (charts, averages) is the simplest form of analysis since information deals with only one particular quantity that changes.

Example 2: Temperature of Delhi during last seven days(in degree Celsius)

Temperature (^{0}C)	40	41	37	38	34	38	36	39	40

Here only temperature is a variable, thus this is called univariate data.

(ii) **Bivariate data** involves two different variables. The analysis is done to find out the connection between these two variables mainly the causes and relationships.

Example 3: Demand curve correctly explains the bivariate data where quantity demanded of a good has an inverse relationship with the price (keeping other factors constant like income).

Example 4: Following bivariate data depicts age and average height of a group of babies and kids.

Age (in months)	3	6	9	12	24	36	48	60
Height (in cms)	58	64	68	74	81	89	95	102

other example for bivariate data can be distance time graph, inflation, unemployment graph, caloric intake versus weight graph which compares two variables.

(iii) **Multi-variate data** involves more than two variables where analysis is done to understand the interactions between different fields in the data

Example 5: Your grade in class XI depends on the performance in various subjects (i.e multiple variables) you have opted.

Example 6: Profit of a company depends on the various statistics like cost of raw materials, labor, marketing expenditures, sales etc.

Univariate data	Bivariate data	Multi-variate data
• involves a single variable • does not deal with causes or relationships • the major purpose of univariate data analysis is to describe ❖ frequency distributions ❖ bar graph, histogram, pie chart, line graph ❖ central tendency (mean, mode, median) ❖ dispersion (range, quartiles, variance, standard deviation)	• involves two variables • deals with causes or relationships • the major purpose of bivariatedata analysis is to describe ❖ analysis of two variables ❖ correlations ❖ comparisons, ❖ relationships, causes, ❖ explanations	• involves more than two variables • deals with explanatory purpose • There exist various ways to analyze the multivariate data depending on the goals.

10.3 Characterizing the data based on variables / Data on various scales

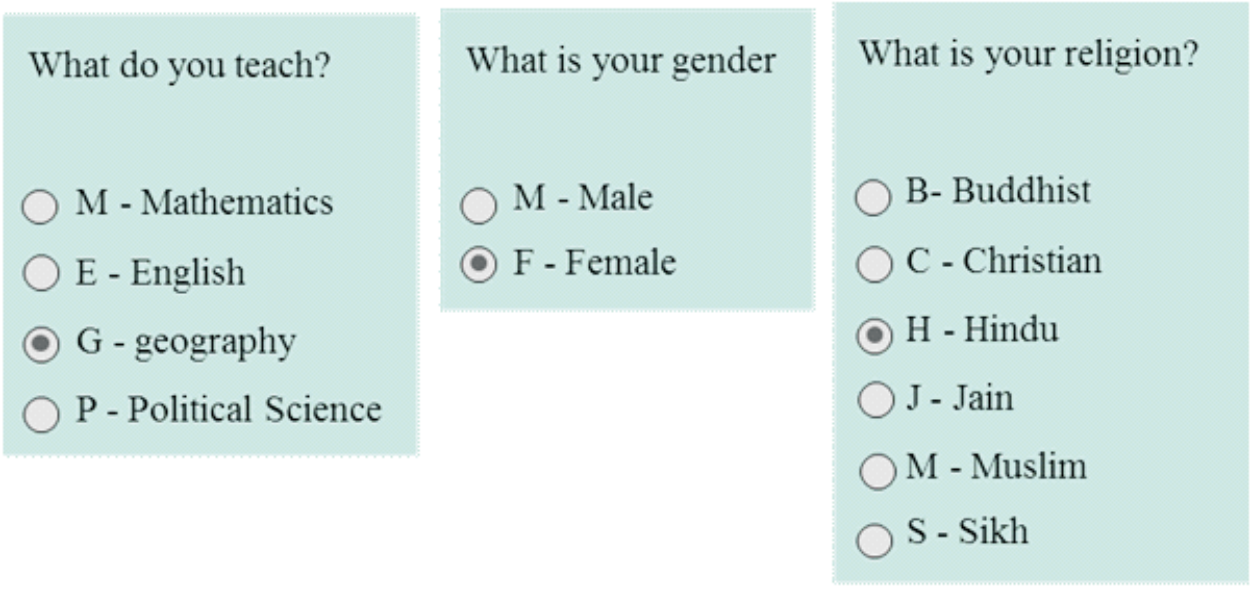
We collect data in different forms; sometimes we collect data in categories, sometimes in positional value and sometimes in absolute value. More precisely, we collect data on different scales of measurement. Classification of data based upon the method of categorizing, counting and measuring

uses four common types of measurement scales- nominal, ordinal, interval, and ratio.

10.3.1 Nominal level measurement:

Nominal level measurement differentiates data based upon their names or type of category they belong to.

Example 7: A sample of teachers in your school categorized according to subject they teache (Mathematics, English, Geography, Political Science, etc.) is an example of nominal level measurement. Putting them as per their gender (male, female), marital status (married/unmarried), religion, is also an example of this kind of measurement.



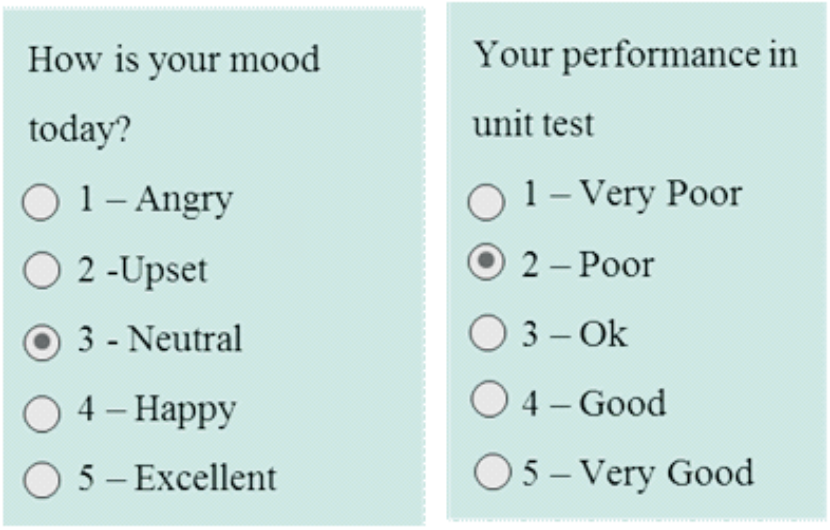
- The categories don't have to be numerical.
- It put the data into mutually exclusive (non-overlapping), exhausting categories where ranking or no order can be applied to the data.
- Each observation belongs to one of several distinct categories.
- They can be called labels (as we simply labeling the variables)

10.3.2 Ordinal level Measurement:

Ordinal level measurement allows categorizing the data and these categories can be ordered, or ranked but precise measurement of differences does not exist.

Example 8: You can put Indian cricket team players as superior, average, or poor batsman. But precise measurement can not be done as sometime they may perform differently.

Therefore ordinal level of measurement classifies data into categories that can be ranked; however, precise differences between the ranks do not exist.



- Observations can be placed into categories that can be ranked.
- The interval between each value of data in scale is not always same.

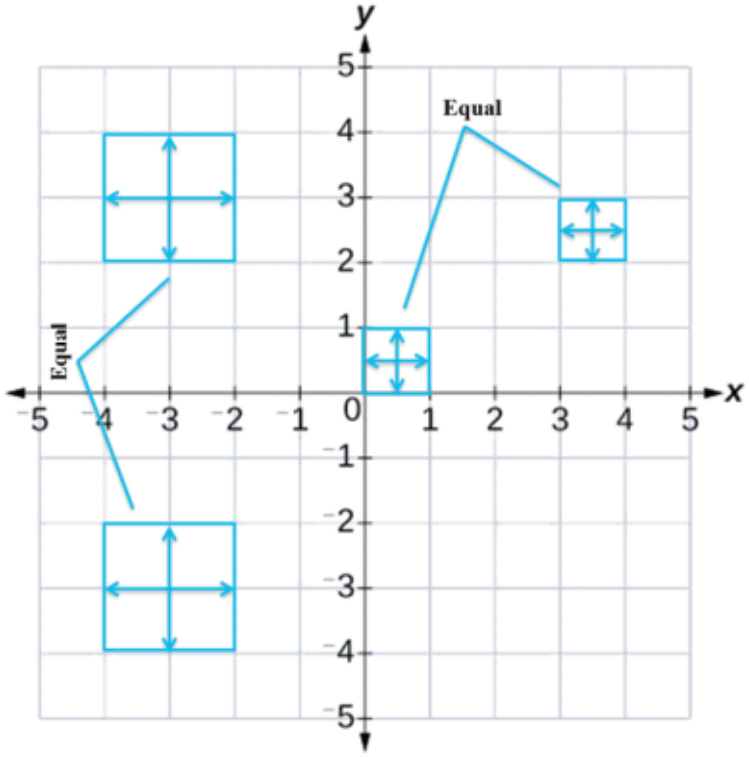
10.3.3 Interval Scale measurement

Interval level measurement specifies the equal distance between each interval, along with categorizing and putting into orders.

Example 9: 100 rupee to 200 rupee is the same interval as 700 rupee to 800 rupee.

Locations in Cartesian coordinate system is another perfect example, where

- Difference between consecutive values is always the same.
- Interval can represent below zero values as well.



10.3.4 Ratio Scale measurement

In ratio level measurement the observations have a value of zero as well along with the having equal intervals i.e Ratio level of measurement has all the features of interval measurement, in addition to existence of a meaningful true zero value.

Examples include height, weight, area, length, time duration etc.

Ratios are always more meaningful as they depicts 'how much' or 'how many' of something.For example, when we say that

i. a car takes double time than train to travel the same distance.

ii. a person has twice the height of his son.

Then ratio between them is 2 to 1 (true ratios).

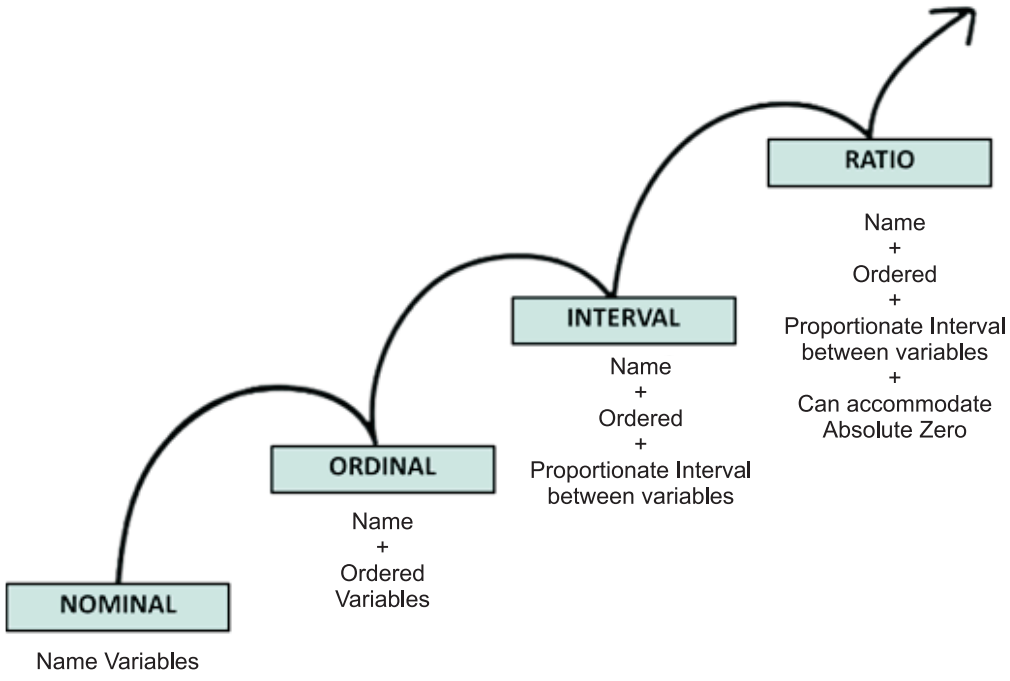
True ratios are considered to exist when the same variable is measured on two different members of the population.

Note

The main difference between interval and ratio scales comes from their ability to dip below zero. Interval can represent below zero values. For example, you can assign a point below 0 in Cartesian coordinate or

temperature below 0 degree can be measured.

But ratio variables never fall below zero value. For example, height and weight is always measured to be 0 and above.



Classification	Graphical measures	Measures of central tendency	Measures of dispersion
Nominal	Bar graphs Pie charts	Mode	Binomial or multinomial variance
Ordinal	Bar graphs Histogram	Median	Range
Interval	Histogram areas are measurable	Mean	Standard deviation
Ratio	Histogram areas are measurable	Geometric Mean, Harmonic Mean	Coefficient of variation

Table adapted from Afifi, A., S. May, and V.A. Clark. 2012. Practical multivariate analysis 5th edition. CRC Press, Taylor and Francis Group, Boca Raton, FL.

10.4 Data representation and visualization

A picture is worth a thousand words. The phrase correctly depicts the essence of presenting data in the form of diagrams and graphs (charts). There are a large number of diagrams being used for presentation of data. The selection of particular diagram depends on the nature of data, objective of presentation and the ability and experience of the person doing this task. Some popular types of diagrams are discussed below;

i. Bar Graph

ii. Pie Chart

iii. Line Graph

iv. Histograms

v. Frequency Polygons

Graph helps in showing the relation between dependent and independent variables.

Example 10 : you might have observed that amount spent on petrol is directly dependent on the distance travelled by the vehicle. Here one quantity affects the another. So here quantity of petrol consumed is an independent variable and amount of petrol bill is the dependent variable. These kind of relationship between such variables can be shown through a graph.

10.4.1 Bar Graph

Bar graph is pictorial representation of data having rectangular bars of equal width. These bars are placed on equal space on one of the axis (either x or y) i.e bars can be either vertical or horizontal. Height of the bars depends on the given data where lower end of the bar touches the base line such that the height of each bar starts from zero units.

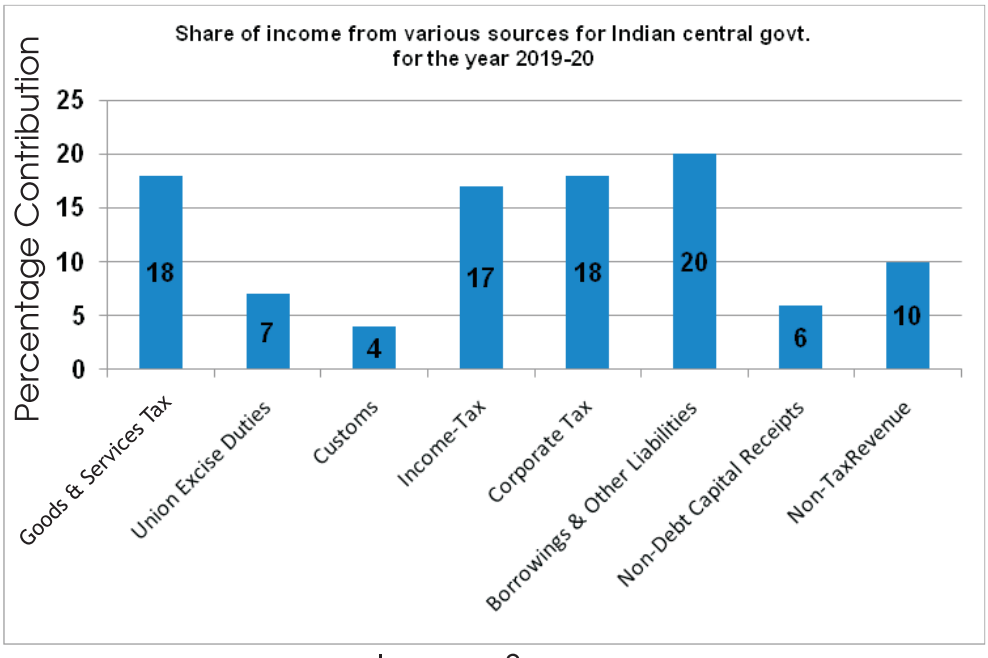
- Every bar depicts only one characteristics of the data.
- The distance between the bars should be equal.
- These bars can be either vertical or horizontal.

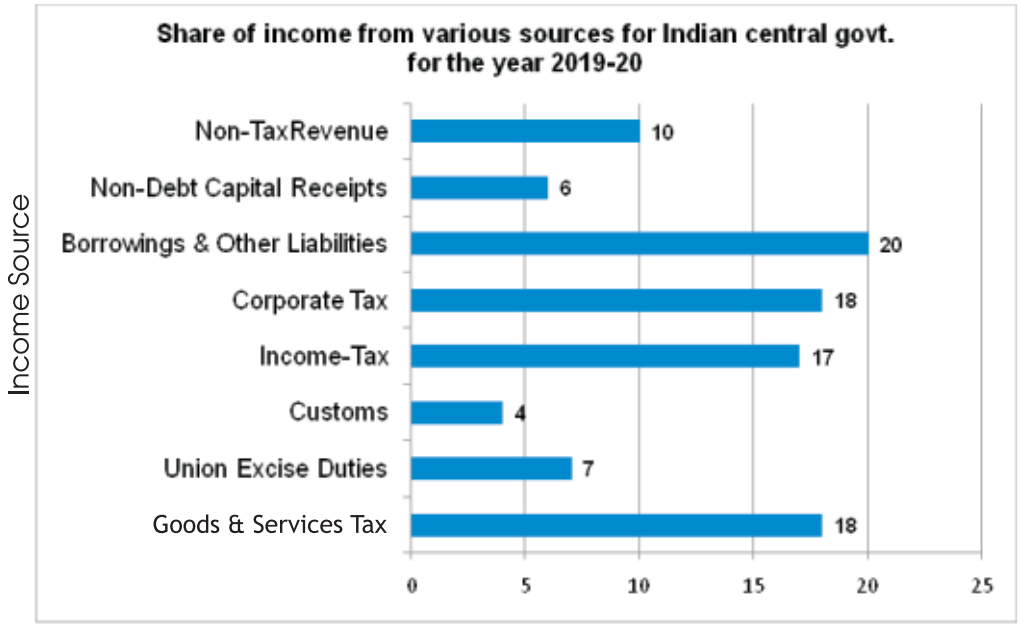
- Bars of a bar diagram can be visually compared by their relation height and accordingly data can be comprehended quickly.

Example 11: Let us take data disclosed in union budget 2020-21 for percentage shares of earning for Indian central govt. for the year 2019-20

Income source	Percentage contribution
Goods & Services Tax	18%
Union Excise Duties	7%
Customs	4%
Income-Tax	17%
Corporate Tax	18%
Borrowings & Other Liabilities	20%
Non-Debt Capital Receipts	6%
Non-Tax Revenue	10%

(Source: https://www.indiabudget.gov.in/)





By observing the above bar graphs answer the following questions:

(i) What is the maximum earning source of government?

(ii) Which two sources credited equal amount of money to government?

10.4.2 Pie Chart

Pie Chart is a circular chart in which we divide the circle into multiple sectors (pie slices). It is used when comparison of a component part is required with other component and the total.

- Pie diagrams is also known as "Angular Circle Diagram"
- To construct it, we use the fact that total of all given values corresponds to the total number of degrees in the circular arc i.e 360^0
- Each sector represents a particular component as a part of the whole
- It is used to show relative sizes

Steps of Pie Diagram:

(i) Construct a circle and divide it into number of sectors to represent the components.

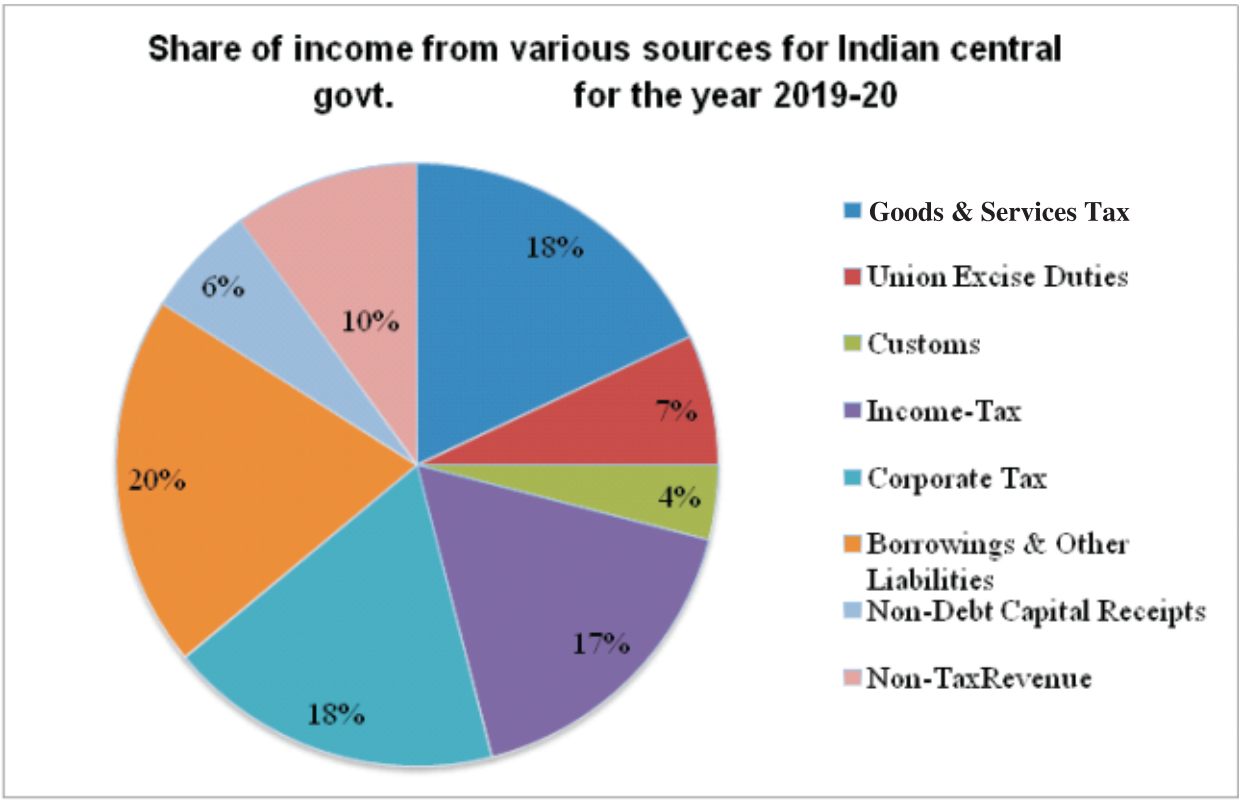
(ii) Convert the value of various components into percentage value of total.

(iii) The percentage values are converted in to corresponding degrees in the circle. Since one circle contains 360^0 and percentage value of all the items are equal to 100, therefore our percentage value is represented by $(360^0)/100 = 3.6^0$ to get size of angles.

(iv) Take base line to draw the angle represented by first component. The new line will become the base for second components angular representation. Repeat these procedures till all the components are represented.

Example 12: Data given in previous example can be represented through Pie Chart as follows:

Income source	**Percentage contribution**	**Share as a component of 360^0**
Goods & Services Tax	18%	$\frac{18}{100} \times 360^0$
Union Excise Duties	7%	$\frac{7}{100} \times 360^0$
Customs	4%	$\frac{4}{100} \times 360^0$
Income - Tax	17%	$\frac{17}{100} \times 360^0$
Corporate Tax	18%	$\frac{18}{100} \times 360^0$
Borrowings & Other Liabilities	20%	$\frac{20}{100} \times 360^0$
Non - Debt Capital Receipts	6%	$\frac{6}{100} \times 360^0$
Non - Tax Revenue	10%	$\frac{10}{100} \times 360^0$

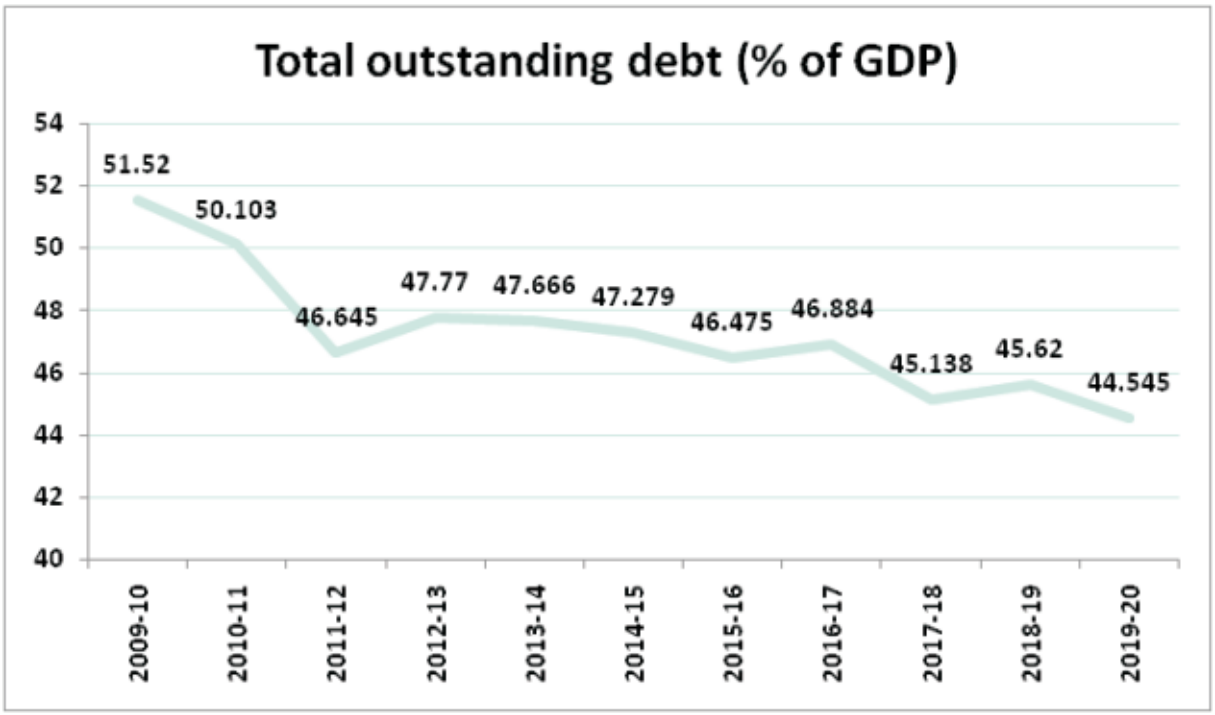


10.4.3 Line Graph

Line graph connects a series of successive data points and shows how one variable behaves over the other variable by representing each value as a dot and connecting the dots to form a line. Usually it is used to relate the time variable with other variables.

- Each dot depicts a different data point
- More than one line may be plotted in the same axis for comparison purpose.

Example 13: Following line graph presents the total outstanding debt (% of GDP) of India, which is the accumulation of borrowings over the years. A higher debt implies that the government has a higher loan repayment obligation over the years.



Source: Economic Surveys 2003-04 to 2018-19 and
https://m.rbi.org.in/Scripts/BS_ViewBulletin.aspx?Id=18703

Fact check : Total outstanding debt of the government has decreased from 55.5% of GDP in 2000-01 to 50.1% of GDP in 2020-21 (estimate). The FRBM Act sets a target of 40% of GDP for outstanding debt to be met by 2024-25.

10.4.4 Histograms

Histogram is like bar graph but does not have gaps between the bars. It is used for continuous class intervals (never drawn for a discrete variable) where class frequencies are represented by area of the corresponding bars.

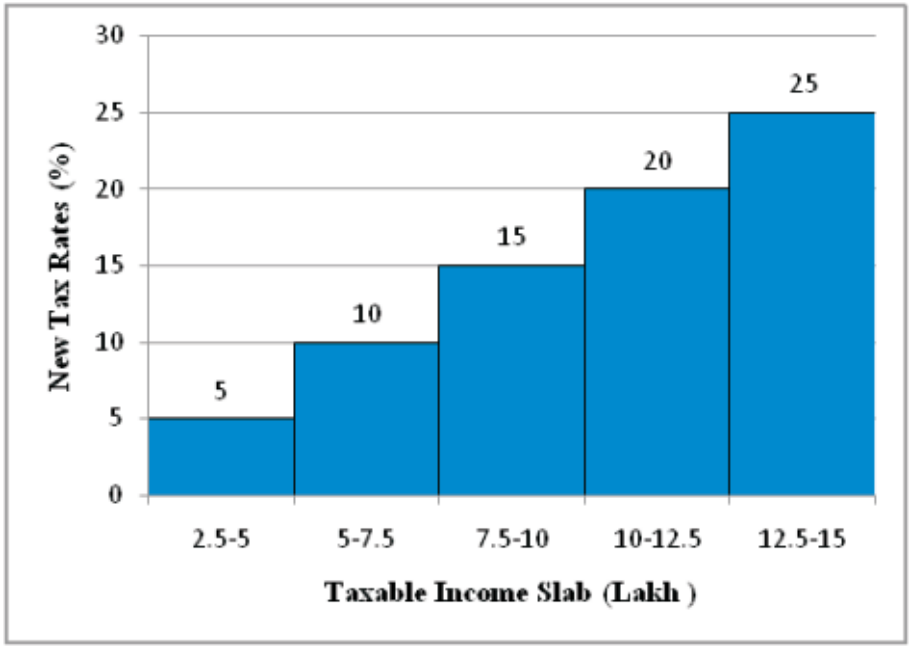
- It is a graph with consecutive bars.
- Height of every bar is equal to the corresponding frequency of the class intervals, taller the bars means more data falls in that particular range.

Example 14: Consider the frequency distribution representing the new tax rates in country as proposed in India's union budget 2020-21

Taxable Income Slab (Lakhs)	New Tax Rates (%)
2.5-5	5
5-7.5	10
7.5-10	15
10-12.5	20
12.5-15	25

Source: PRS India

https://www.prsindia.org/parliamenttrack/budgets/union-budget-2020-21-analysis



In case of **unequal class intervals** we need to first adjust the frequencies before constructing the histograms, as width of rectangular bars is going to be unequal which doesn't suit the definition of histogram.

Steps to adjust frequencies:

Step 1: Determine minimum class size. In below example, the minimum class size is 5

Step 2: The lengths of the rectangles (frequencies) are then modified to be proportionate to the minimum class size.

- Adjusted frequency of a class = (Minimum class size)/(Class size) x Frequency

- By doing so we will see that frequencies of the classes with minimum class size are not changed and frequencies of all other classes are adjusted.

Example 15 : Prepare the histogram for following data

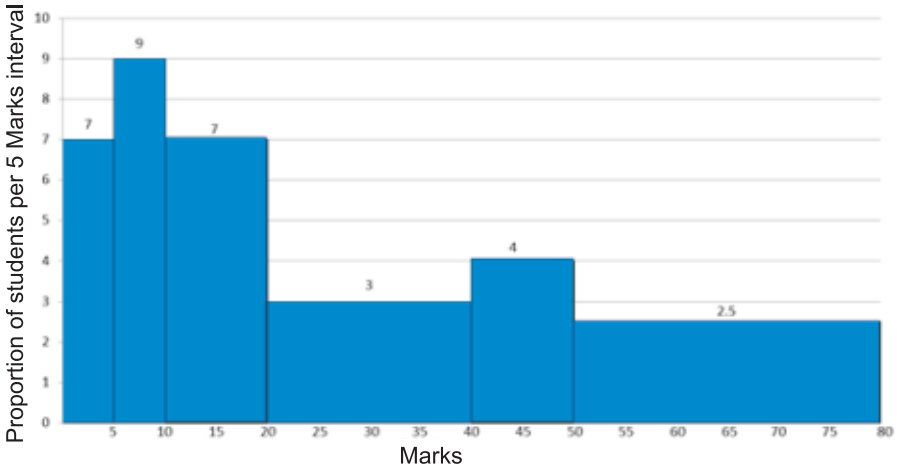
Marks	0-5	5-10	10-20	20-40	40-50	50-80
No of students	7	9	14	12	8	15

Here class intervals are unequal so we need to adjust the frequencies for preparing the histogram.

The histogram for adjusted frequencies is shown as under

Marks	**Frequency (No of students)**	**Class size**	**Adjusted frequency**
0-5	**7**	**5**	$\frac{5}{5}\times 7 = 7$
5-10	**9**	**5**	$\frac{5}{5}\times 9 = 9$
10-20	**14**	**10**	$\frac{5}{10}\times 14 = 7$
20-40	**12**	**20**	$\frac{5}{20}\times 12 = 3$
40-50	**8**	**10**	$\frac{5}{10}\times 8 = 4$
50-80	**15**	**30**	$\frac{5}{30}\times 15 = 2.5$

Now we have calculated the class sizes of 5 marks in each case, so the correct representation of histogram with unequal class intervals is constructed in below figure



Question 1:

Discuss the difference between bar graph and histogram from charts given in examples of respective concepts.

10.4.5 Frequency Polygon

A frequency polygon is similar to line graph but it is used for a continuous frequency distribution. The intervals in the continuous distribution are represented by the midpoint of each corresponding interval and mid points and corresponding frequencies are plotted as points in XY plane. All points are joined using free hand. When joined by line segments we obtain a figure and to complete the polygon we assume a class interval with frequency zero.

Frequency polygon can also be drawn independently without drawing histograms. For this, we need to directly find out the mid-points of respective class-intervals. These mid points are called as Class marks and calculated by

$$\text{Class Mark} = \frac{\text{Upper Limit} + \text{Lower Limt}}{2}$$

A. Creating frequency polygon without histogram:

Step 1: Mark the values of variable on Horizontal x-axis and frequencies on vertical y-axis.

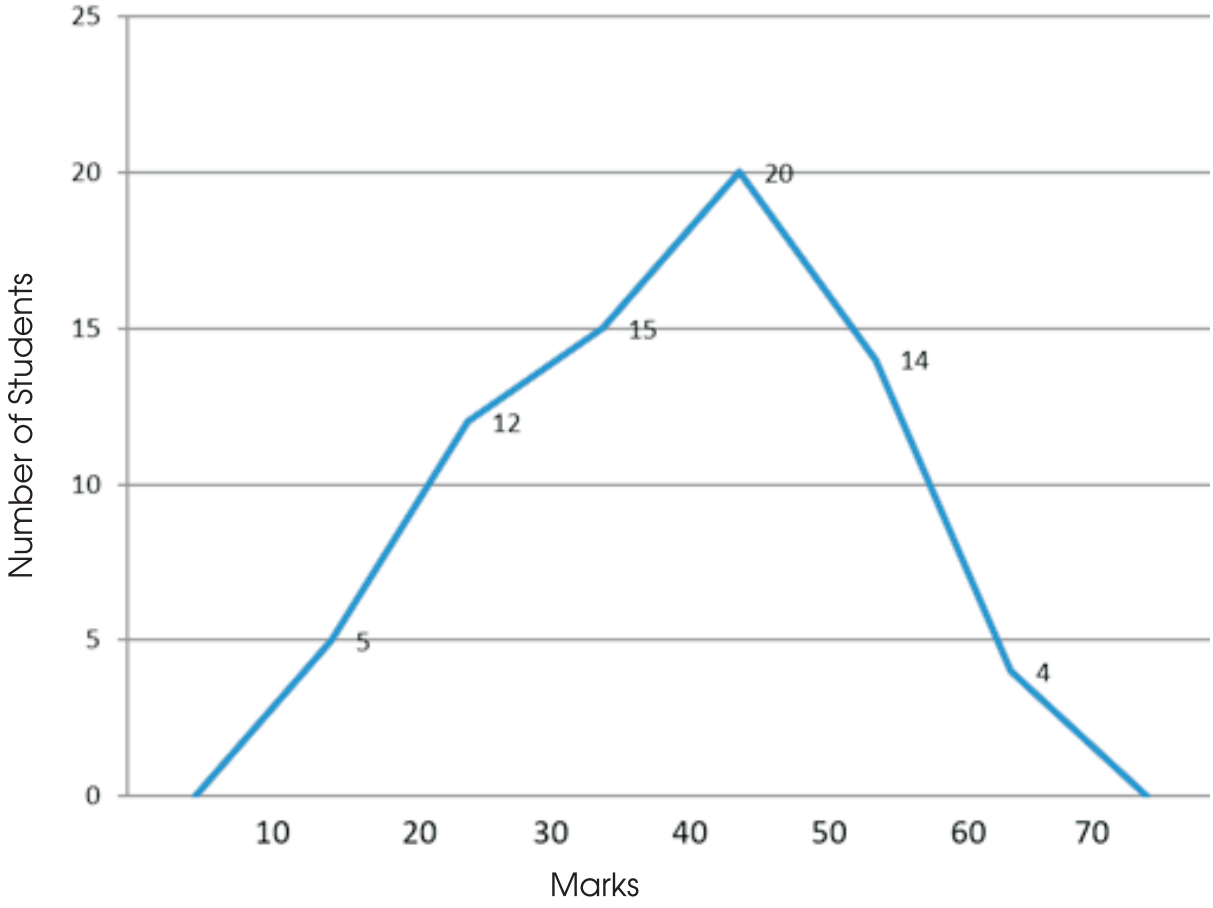
Step 2: Connect the points so drawn by a straight line.

Step 3: Connect the extreme points to the base assuming to have zero frequency.

Example 16: Consider the marks, out of 100, obtained by 70 students of a class in a test, given in table below:

Class Interval	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	5	12	15	20	14	4

Draw a frequency polygon corresponding to this frequency distribution table.



B. Creating frequency polygon with histogram:

Step 1: Draw histogram of the given frequency distribution

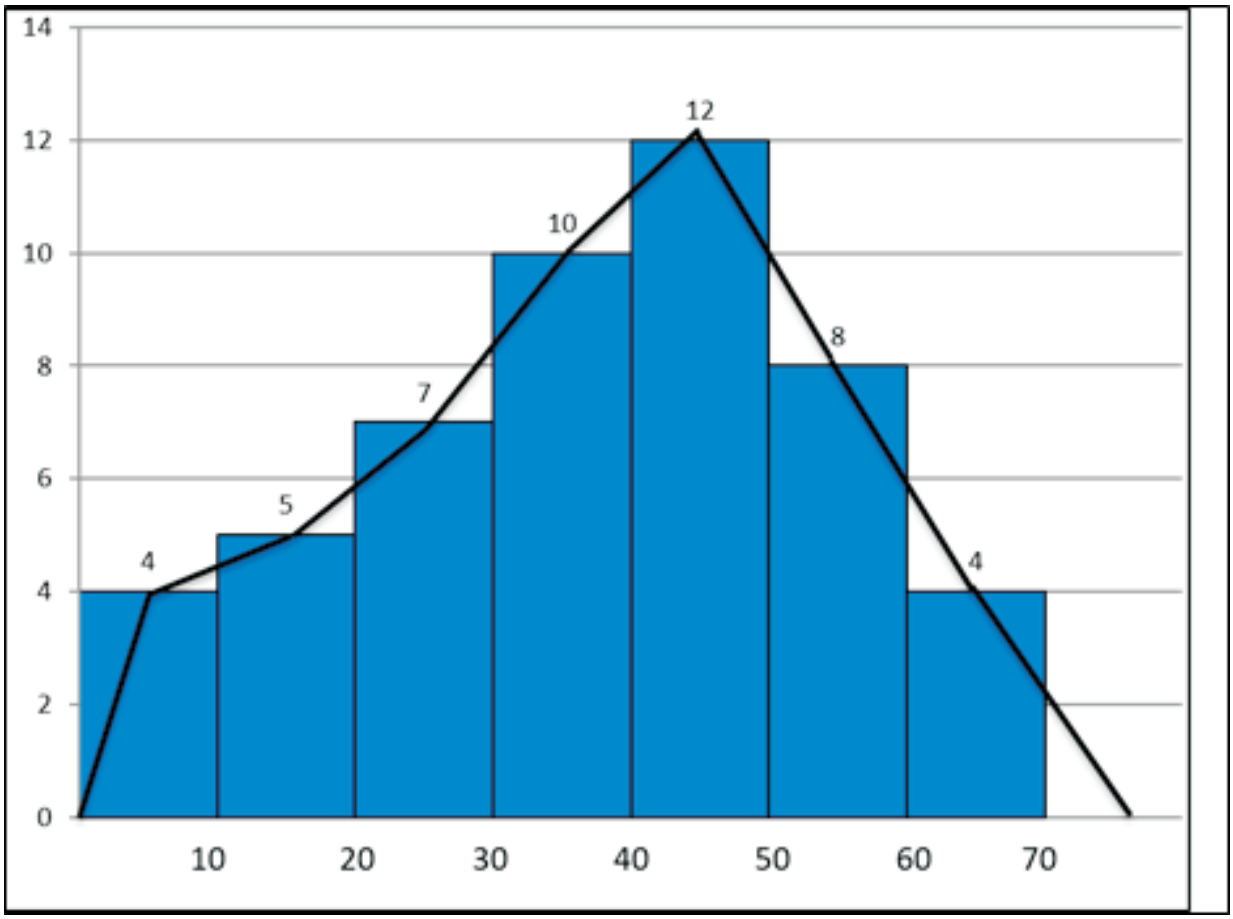
Step 2: Locate the class marks on top horizontal side of each rectangle in histogram

Step 3: Locate two imaginary interval of common size before and after the last class Interval

Step 4: Join all the midpoints, making it a polygon

Example 17: Construct frequency polygon for the following set of data

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
frequency	4	5	7	10	12	8	4



How to make choice out of various graphs

- A bar graph is used to indicate the comparison among different categories where independent variable is non-numerical.
- A pie graph is used for comparing parts of a whole. They do not show any changes over time.
- A line graph is used to display those data (independent variable) which continuously changes over the period of time, through unbroken lines.
- A histogram is used for representing the data intervals.
- Frequency Polygons are used to understand the shapes of distribution.
- The histogram and frequency polygon are two different ways to represent the same data set.

Question 2:

Collect multiple set of data and discuss which type of graph representation is most suitable for a given set of data. Is it possible to represent the data graphically in two or more ways?

Question 3:

India's union budget 2020-21 purpose to change following amendment in tax rate. Represent the information using a suitable graph. Justify your choice.

Taxable Income Slab (Rs.)	Current Tax Rates	New Tax Rates
0-2.5 Lakh	Exempt	Exempt
2.5-5 Lakh	5%	Exempt
5-7.5 Lakh	20%	10%
7.5-10 Lakh	20%	15%
10-12.5 Lakh	30%	20%
12.5-15 Lakh	30%	25%
Above 15 Lakh	30%	30%

Source: PRS India

https://www.prsindia.org/parliamenttrack/budgets/union-budget-2020-21-analysis

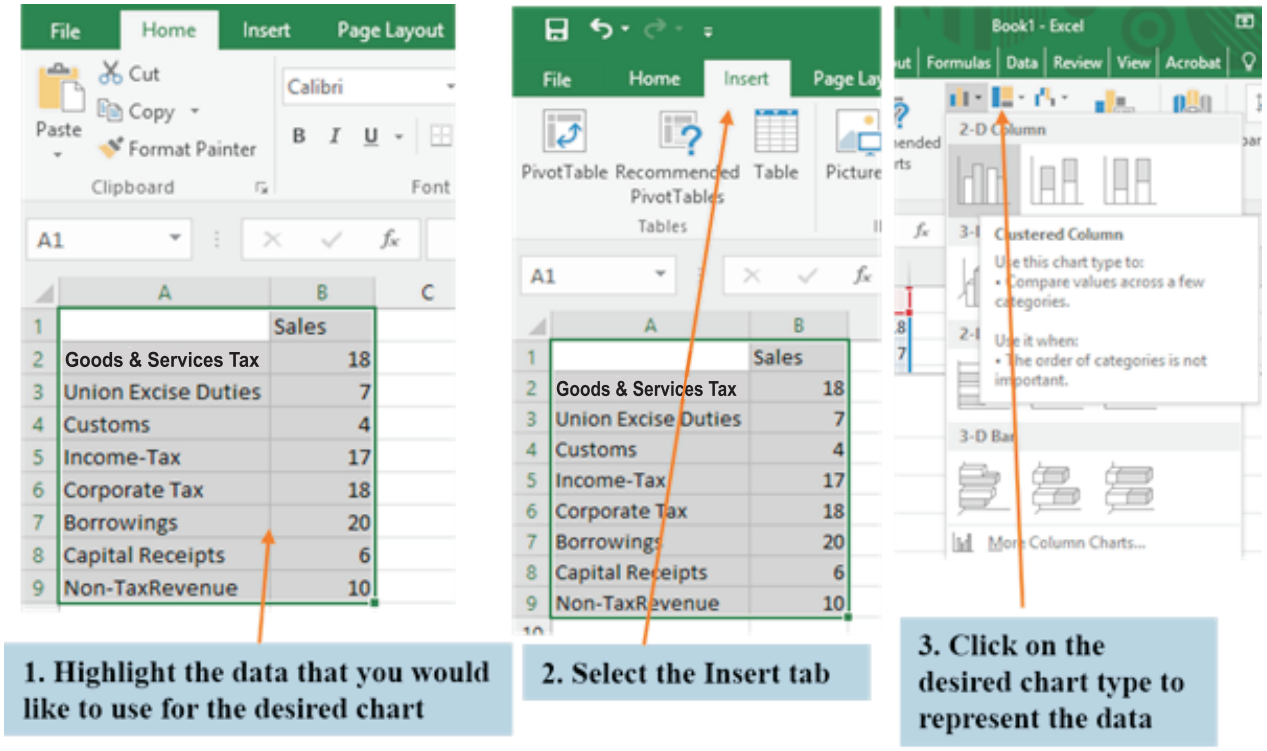
Question 4:

India's union budget 2020-21 fixes Fiscal Responsibility and Budget Management targets FRBM targets for deficits (as % of GDP)

	Actuals 2018-19	Revised 2019-20	Budgeted 2020-21	Target 2021-22	Target 2022-23
Fiscal Deficit	3.4%	3.8%	3.5%	3.3%	3.1%
Revenue Deficit	2.4%	2.4%	2.7%	2.3%	1.9%

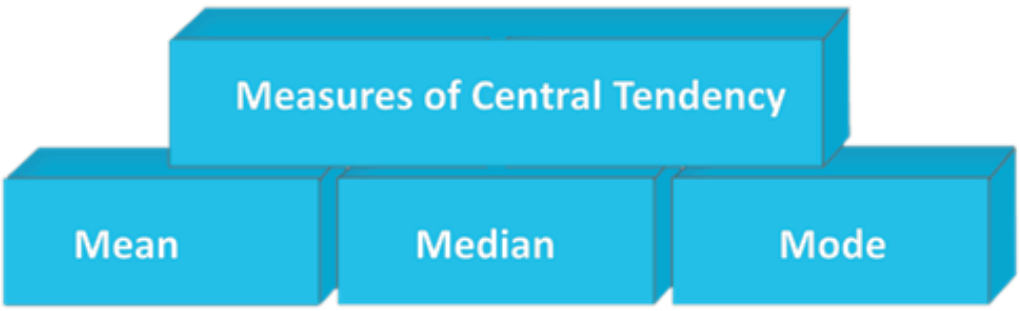
Sources: Medium Term Fiscal Policy Statement, Union Budget 2020-21; PRS.
Represent the information using a suitable graph.

How to create chart in Excel



10.5 Central Tendency

A measure of central tendency is a value that gives the central position of the given data. The most common measures of central tendency are Mean, Median and Mode. It depends on the data and intended purpose for choosing the method of calculating center point.



10.5.1. Mean

The arithmetic mean is usually called the mean or average and is the most common measure of central tendency. It includes every value of data, so it is influenced by outliers which are extreme values. It is denoted by $(\bar{x})$?.

Arithmetic mean of discrete data

$$\text{Arithmetic Mean } (\bar{x}) = \frac{\text{(Sum of all observations)}}{\text{(Number of observations)}}$$

Excel formula for Mean
= AVERAGE(data range)

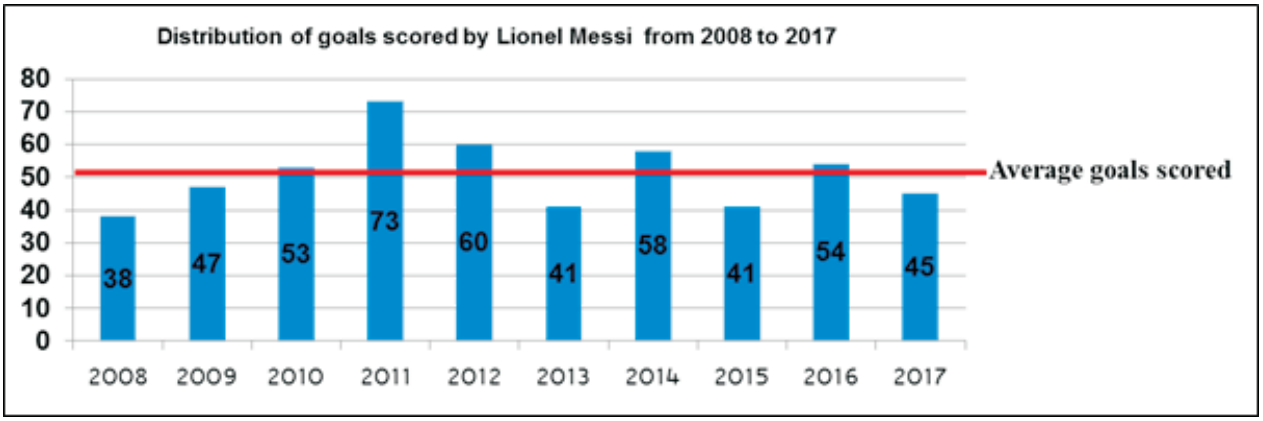
Example 18: Find out the arithmetic mean of goals scored by footballer Lionel Messi from 2008 to 2017 in all competitions;

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Goals scored	38	47	53	73	60	41	58	41	54	45

Source: https://messivsronaldo.net/

$$\text{Arithmetic Mean } (\bar{x}) = \frac{38+47+53+73+60+41+58+41+54+45}{10} = \frac{510}{10} = 51 \text{ Goals}$$

This means on an average Lionel Messi scored 51 goals in a year.



Mean of group data

Suppose $x_1 + x_2 + \ldots + x_n$ are the observations and $f_1 + f_2 + \ldots + f_n$ are their respective frequencies then sum of the values of all the observations = $f_{1x1} + f_{2x2} + \ldots..+f_{nxn}$, and the number of observations = $f_1 + f_2 + \ldots + f_n$.

The mean $\bar{x}$ of the grouped data is calculated by

i) Direct method **(ii) Assumed mean method** **(iii) Step deviation method**

The mean obtained by all the three methods is same where assumed mean method and step-deviation method are just simplified forms of the calculating mean by direct method.

(i) **Direct method:**

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \text{ or } \frac{\sum f_i x_i}{N}$$

Where $\sum f_i x_i$ gives the sum of all the observations and $\sum f_i$ gives the number of observations.

Recall the concept of frequency polygons where we talked about calculating class marks (x_i) by $\frac{\text{Upper Limit} + \text{Lower Limt}}{2}$

Example 19: Find the mean marks of the students using direct method.

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	7	4	6	3	5

Solution:

Class Interval	Frequency (f_i)	Class mark (x_i)	$f_i x_i$
0 - 10	7	5	35
10 - 20	4	15	60
20 - 30	6	25	150
30 - 40	3	35	105
40 - 50	5	45	225
	$\sum f_i = 25$		$\sum f_i x_i = 575$

$$(\bar{x}) \quad \text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{575}{25} = 23$$

(i) Assumed mean method :

This method is employed when given data (f_i and x_i) are large. We assume a mean (a) and calculate the deviations (d) from this assumed mean for every observation of data.

$$\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i}$$

where a = assumed mean

$d_i = x_i - a$ deviations from assumed mean

Example 20: Find the mean marks of the students using assumed mean method.

Class Interval	0-20	20-40	40-60	60-80	80-100
Frequency	3	2	5	6	4

Solution:

Class Interval	f_i	Class Mark (x_i)	$d_i = x_i - a$	$f_i d_i$
0-20	3	10	-40	-120
20-40	2	30	-20	-40
40-60	5	50	0	0
60-80	6	70	20	120
80-100	4	90	40	160
	$\sum f_i = 20$			$\sum f_i d_i = 120$

Let assumed mean (a) be 50, then

Mean ( ) = $\bar{x}$ $a + \frac{\sum f_i d_i}{\sum f_i} = 50 + \frac{120}{20} = 56$

(iii) **Step deviation method:** Step deviation method can be used when class-intervals for all the classes in a continuous series are of same magnitude (width) or we can say when deviations ($d_i = x_i - a$) from assumed mean is divisible by a common factor.

$$\bar{x} = a + \frac{\sum f_i u_i}{\sum f_i} \times h$$

where a = assumed mean

h = class width or common factor $u_i = \frac{x_i - a}{h}$

Example 21: Find the mean marks of the students using step deviation method.

Class interval	f_i	Class Mark (x_i)	$d_i = x_i - a$	$u_i = \frac{x_i - a}{h}$	$f_i u_i$
0-10	5	5	-20	-2	-10
10-20	8	15	-10	-1	-8
20-30	15	25	0	0	0
30-40	16	35	10	1	16
40-50	6	45	20	2	12
	$\sum f_i$ =50				$\sum f_i u_i$ =10

Let assumed mean (a) be 25.

Here class width (h) = 10

Therefore, Mean $\bar{x} = a + \frac{\sum f_i u_i}{\sum f_i} \times h = 25 + \frac{10}{50} \times 10 = 27$

10.5.2 Median

The median is the number located exactly at the center of data when arranged in an order, either ascending descending. It includes only middle most value of given data, so it is not influenced by the outliers.

Median is a positional average because its value depends upon the position of an item and not on its magnitude and changing the order of data does not changes the median.

Excel formula for Median
= MEDIAN(data range)

Calculating median of ungrouped data:

Step 1: Arrange the data in ascending/descending order

Step 2: Determine the number of observation

Step 3: Find out the middle most observation as follows

- If number of observation is odd, then Median = Middle most in the observation distribution.

  i.e $\left(\frac{n+1}{2}\right)^{th}$ observation.

  For example if n=3, then value of $\left(\frac{3+1}{2}\right)^{th}$ i.e 2nd observation will be the median

- If number of observation is even, then Median = Average of middle observation i.e median is average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations
- For example if n=4, then mean of $\left(\frac{4}{2}\right)^{th}$ and $\left(\frac{4}{2}+1\right)^{th}$ observations i.e mean of 2nd and 3rd observation will be median.

Example 22: Find median of 14, 3, 1, 7, 10.

Solution:

1	3	7	10	14

Median

Step 1: Arrange them in ascending order: 1, 3, 7, 10, 14

Step 2: It has 5 (odd) observations

Step 3: So $\left(\frac{5+1}{2}\right)^{th}$ observation is the median of this data.

i.e 7 is the median of this data.

Median of grouped data:

The median for grouped data is given by the formula:

Median = $l + \left(\frac{\frac{n}{2} - cf}{f}\right) \times h$

l = lower limit of median class,

n = number of observations,

cf = cumulative frequency of class preceding the median class,

f = frequency of median class,

h = class size (assuming class size to be equal)

To find the median class, we find the cumulative frequencies of all the classes and locate the class whose cumulative frequency is greater than or closest to. This is called the median class.

Example 23: Following set of data relates to daily wages of persons working in a factory. Compute the median daily wage.

Daily wages (Rs)	200-250	250-300	300-350	350-400	400-450	450-500
No of workers	5	6	8	9	10	12

Solution :

Daily wages (Rs) (Class Interval)	No of workers (f_i)	Cumulative frequency
200 -250	5	5
250 -300	6	11
300 -350	8	19
350 -400	9	28
400 -450	10	38
450 -500	12	50
	$\sum f_i$ =50	

The given set of data is arranged in ascending order where median class is greatest & nearest to $\frac{n^{th}}{2}$ observation (i.e 25th item) which lies in 350-400 class interval. By applying the formula of median we get

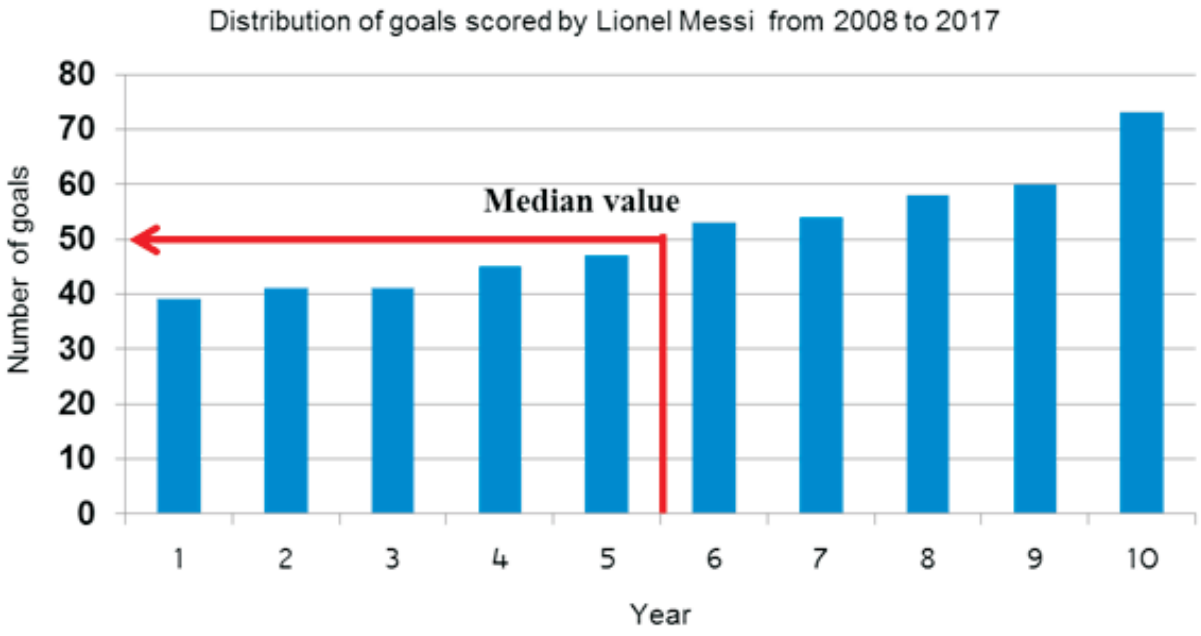
Median = $$l+\left(\frac{\frac{n}{2}-cf}{f}\right)\times h = 350+\left(\frac{\frac{50}{2}-19}{50}\right)\times 50 = 356$$

Therefore, median daily wage is Rs. 356.

This signifies that 50% of the workers receive less than or equal to Rs. 356 and 50% of the workers receive more than or equal to this value.

Note: Geometrically the median (value on x-axis) divides a histogram of data into two parts having equal areas.

Example 24: Observe the graph of data given in earlier example of mean where goals scored by footballer Lionel Messi from 2008 to 2017 in all competitions were given;



Here total number of year are even number, so positional average (median) exactly divides the graph into two parts having equal areas.

10.5.3 Mode

The mode is the value that occurs most frequently in a set of data.There can be more than one mode depending upon the given set of data. For data having one mode the distribution is said to be uni-model, and for two modes the distribution is said to be bi-model.

No mode data: When each value occurs, the same number of times in the data then there is no mode.

Multi-mode data: If two or more values occur the same number of times, then there are two or more modes and the data set is said to be multi-mode.

Excel formula for Mode
= MODE(data range)

Mode of ungrouped data:

When the data has discrete set of values, then mode may be found by simply inspection method and finding the observation having maximum frequency.

Example 25:

- The set of data 1,2,3,3,7,9,12,11,13,13,13, 15 has mode 13 (Uni-model).
- The set of data 4, 6, 8, 12, 15, 16 has no mode.
- The set if data 2, 3, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8 has two modes, 5 and 6 (bimodal).

Mode of grouped data

Mode of continuous series lies in a particular class (modal class). The following method is used in determining mode:

Mode = $l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h$

A class interval with the maximum frequency is called the modal class.

Where,

l = lower limit of the modal class,

h = class size (assuming class size to be equal)

f1 = frequency of the modal class,

f0 = frequency of the class preceding the modal class,

f2 = frequency of the class succeeding the modal class.

Example 26: Find out mode of the following series:

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	8	10	15	7	10

Solution:

Class Interval	Frequency (f_i)
0 - 10	8
10 - 20	10
20 - 30	15
30 - 40	7
40 - 50	10

Here modal class is 20-30, So l=20, $f_1 = 15$, $f_0 = 10$, $f_2 = 7$, h = 10

We know that Mode $= l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h$

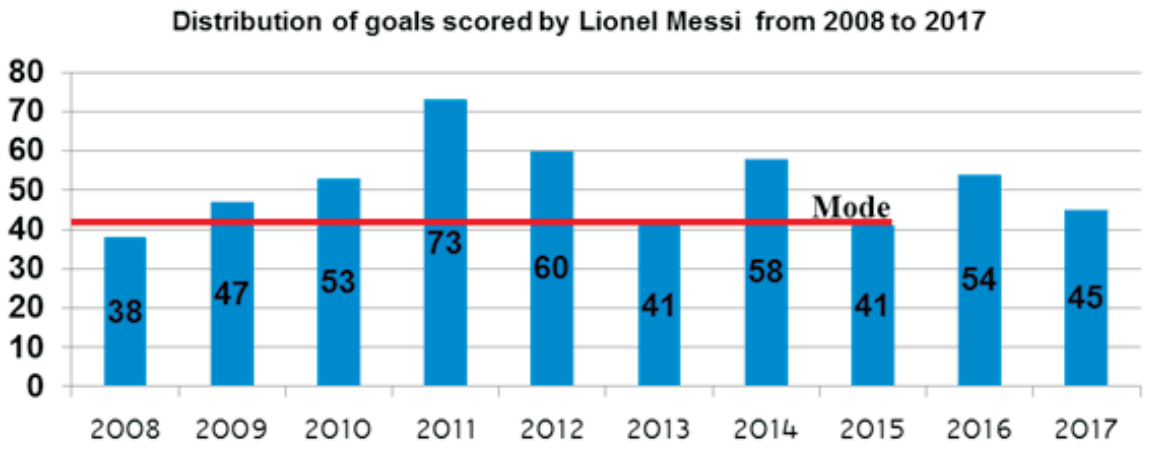
Mode = $20 + \left(\frac{15 - 10}{2 \times 15 - 10 - 7} \right) \times 10 = 23.48$

Therefore, mode of given series is 23.48

Note: Geometrically In the case of grouped data where a frequency curve has been constructed to fit the data, the mode will be the value (or values) of X corresponding to the maximum point (or points) on the curve.

Example 27: Observe the graph of data given in earlier example of mean where goals scored by footballer Lionel Messi from 2008 to 2017 in all competitions were given;

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Goals scored	38	47	53	73	60	41	58	41	54	45



Note:

For uni - model frequency data there exist a relation between mean, median, and mode:

Mean - Mode = 3 (Mean - Median)

10.6 Dispersion

We have already studied that measures of central tendency (mean, median, mode) which explores the central value of a given data set. But in order to make better interpretation from data, we should have an idea about the spread of the data around the central value. So the degree to which data tend to spread around the central value is called dispersion or variation of the data. It measures the variation of the items among themselves and the variation around the central value.

Measures of dispersion / variation in a distribution

The dispersion in a data is measured on the basis of the observations and the types of the measure of central tendency, used there. Several measures of this dispersion are available, the most common are

(i) Range (R)

(ii) Quartile Deviation (QD)

(iii) Mean Deviation (MD)

(iv) Standard Deviation (SD)

10.6.1 Range:

Range is the difference between the highest and lowest value of the given distribution. It provides a quick estimate of the variability of the two distributions. However, the range is calculated only based upon the two extreme values. So at times, it may mislead the variability of data if extremes are rare or unusual.

Range = Maximum value - Minimum value

Coefficient of range = $\frac{Max. - Min.}{Max. + Min.}$

Excel formula for range
In Excel the range is calculated using the MIN and MAX functions =MAX(data range) MIN(data range)

Example 28: Find out the range of salaries for staffs of a schools.

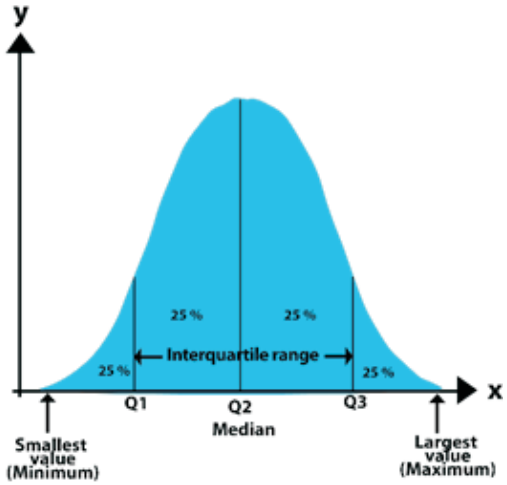
Staff	Salary (Rs.)
Principal:	100000
Senior Teacher	70000
Junior Teacher	45000
Other workers	25000

Range = Maximum value - Minimum value = 100000-25000 = Rs. 75000

Above example make it clear that range is not based on all the values of data, simply calculated by extreme values. Until maximum and minimum values remain intact, any change in other values does not affect range. So in order to have a more meaningful statistic and measure the variability, we use measures called the variance and standard deviation.

10.6.2 Quartile Deviation (QD)

Quartiles separate the original set of data into four equal parts. Each of these parts contains one-quarter (25%) of the data where two halves are less than median and other two greater than median.



- The difference between the third and first quartiles is called the interquartile range (IQR) = Q3 - Q1
- The IQR is sometimes called the middle half and effective in identifying the outliers in data.
- An outlier is any value at least 1.5 IQR above Q3 or below Q1

Quartile deviation (Q.D.) = $\frac{Q_3 - Q_1}{2}$

Quartile deviation of ungrouped data

Step 1: Arrange the values in ascending order and assign serial number to each.

Step2: Determine first quartile (Q1), third quartile (Q3) by following formula

$$Q1 = \frac{1}{4}(n+1)^{th}\ value$$

$$Q3 = \frac{3}{4}(n+1)^{th}\ value$$

Step3: Calculate the Quartile deviation using the formula

Quartile deviation (Q.D.) = $\frac{Q_3 - Q_1}{2}$

Excel formula for quartiles	
Q_1	= PERCENTILE(Data range, 0.25)
Q_2	=MEDIAN(Data range)
Q_3	= PERCENTILE(Data range, 0.75)

Example 29: Calculate the Quartile deviation of following observations:

15, 20, 22, 28, 35, 27, 44, 48, 50, 55, and 60

Solution:

Step 1: Given data: 15, 20, 22, 28, 35, 27, 44, 48, 50, 55, 60

Step 2: Q1 $= \frac{1}{4}(n+1)^{th}$ value $= \frac{1}{4}(11+1)^{th}$ = 3rd value = 22

Similarly, Q3 $= \frac{3}{4}(n+1)^{th}$ value $= \frac{3}{4}(11+1)^{th}$ = 9th value = 50

Step 3: Quartile deviation (Q.D.) $= \frac{Q_3 - Q_1}{2} = \frac{50-22}{2} = 14$

Therefore, Quartile deviation of given observations is 14.

Quartile deviation of grouped data

Step 1: Calculate cumulative frequencies corresponding to each value.

Step 2: Determine first quartile (Q1), third quartile (Q3) by following formula

$$Q1 = l + \left(\frac{\frac{n}{4} - cf}{f} \right) \times h$$

$$Q3 = l + \left(\frac{\frac{3n}{4} - cf}{f} \right) \times h$$

l = lower limit of class containing the score

n = total number of values

cf = cumulative frequency of preceding class

f = frequency of class interval

h = class size

where Q1 is the size of $\left(\frac{n}{4}\right)^{th}$ value and Q3 is the size of $\left(\frac{3n}{4}\right)^{th}$ value

Step 3: Calculate the Quartile deviation using the formula

Quartile deviation (Q.D.) $= \frac{Q_3 - Q_1}{2}$

Example 30: Calculation the Quartile Deviation from following frequency distribution:

Class Interval	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84
Frequency	1	2	6	7	9	11	8	7	5	3	1

Solution:

Class Interval	Frequencies (f)	Cumulative Frequencies (cf)
30-34	1	1
35-39	2	3
40-44	6	9
45-49	7	16
50-54	9	25
55-59	11	36
60-64	8	44
65-69	7	51
70-74	5	56
75-79	3	59
80-84	1	60
	N=60	

We know, Q1 is the size of $\left(\frac{n}{4}\right)^{th}$ value and Q3 is the size of $\left(\frac{3n}{4}\right)^{th}$ value

$\longrightarrow$ Q1 is the size of $\left(\frac{60}{4}\right)^{th}$ value = 15th value. The class containing the 15th value is 45-49.

$$Q1 = l + \left(\frac{\frac{n}{4} - cf}{f}\right) \times h = 44.5 + \left(\frac{\frac{60}{4} - 9}{7}\right) \times 5 = 48.79$$

Q3 is the size of $\left(\frac{180}{4}\right)^{th}$ value = 45th value. The class containing the 45th value is 65-69.

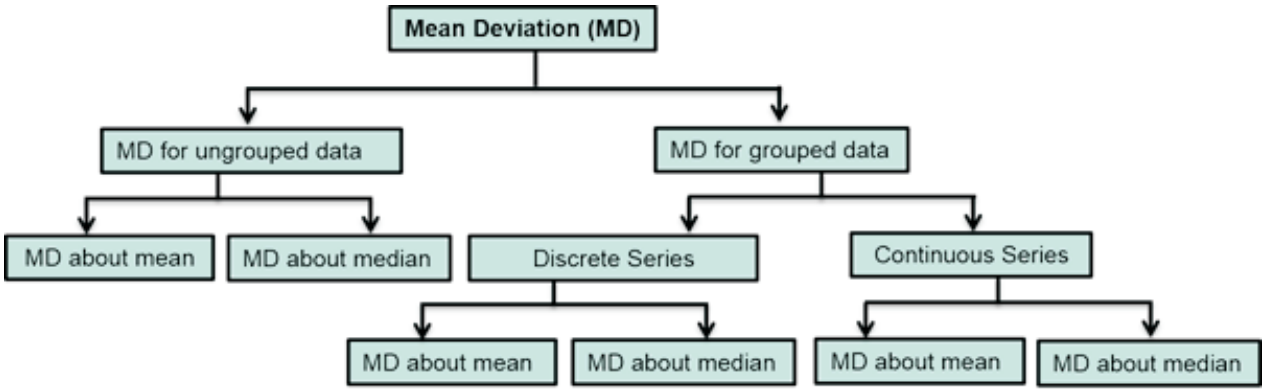
Calculating Q3: = $l + \left(\frac{\frac{3n}{4} - cf}{f}\right) \times h = 64.5 + \left(\frac{45-44}{7}\right) \times 5 = 65.21$

Quartile deviation (Q) = (Q.D.) = $\frac{Q_3 - Q_1}{2} = \frac{65.21 - 48.79}{2} = 8.21$

Therefore quartile deviation of given data is 8.21

In both individual and discrete series, Q_1 is the size of $\frac{1}{4}(n+1)^{th}$ value, but in a continuous distribution, it is the size of $\left(\frac{n}{4}\right)^{th}$ value. Similarly, for Q_3 and median (Q_2) also, n is used in place of n+1.

10.6.3 Mean Deviation (MD)



A. Mean Deviation for ungrouped data:

Let n observations be $X_1, X_2, X_3, X_4, X_5 \ldots\ldots\ldots\ldots\ldots X_n$ then

(I) Mean deviation about mean

M.D. $(\bar{x}) = \frac{1}{N}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$ where $\bar{x}$ = mean

(II) Mean deviation about median

M.D. $(median) = \frac{1}{N}\sum_{i=1}^{n}\left|x_i - M\right|$ where M = median

Example 31: Find the mean deviation about the mean and median for the data: 4,3,2,5,7,6,8.

Solution: The mean of data, $(\bar{x}) = \frac{4+3+2+5+7+6+8}{7} = 5$

The deviation of the respective observation from the mean $(\bar{x})$ ie $x_i - \bar{x}$ are

4-5, 3-5, 2-5, 5-5, 7-5, 6-5, 8-5 or -1, -2, -3, 0, 2, 1, 3

The absolute values of deviations i.e $\left|x_i - \bar{x}\right|$ are 1, 2, 3, 0, 2, 1, 3

So, mean deviation mean, M.D. $(\bar{x}) = \frac{1}{N}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|, = \frac{1+2+3+0+2+1+3}{7} = 1.71$

Now, Median $= (\frac{7+1}{2})^{th} =$ 4th observation = 5

The deviation of the respective observation from the median, i.e. x_1-median are

4-5, 3-5, 2-5, 5-5, 7-5, 6-5, 8-5 or -1, -2, -3, 0, 2, 1, 3

The absolute values of deviations i.e $|x_i - M|$ are 1, 2, 3, 0, 2, 1, 3

So mean deviation about median,

M.D. $(median) = \frac{1}{N}\sum_{i=1}^{n}|x_i - M| = \frac{1+2+3+0+2+1+3}{7} = 1.71$

B. Mean deviation for grouped data

Let the given data consist of n distinct values $x_1, x_2, x_3, x_4, x_5 \ldots\ldots x_n$

having frequencies $f_1, f_2, f_3, f_4, f_5 \ldots\ldots f_n$ respectively.

We know that data can be grouped into two ways:

(a) Discrete frequency distribution

(b) Continuous frequency distribution

(a) Discrete frequency distribution

(i) Mean deviation about mean

M.D. $(\bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i|x_i - \bar{x}|$ where $\bar{x}$ = mean

Example 32: Suppose a class of 25 students conducted a quiz and grades obtained are given in following table. Find the mean deviation about mean of data.

Grade	5	10	15	20	25
Number of students	7	4	6	3	5

Solution:

x_i	Frequency (f_i)	$f_i x_i$	$\lvert x_i - \bar{x}\rvert$	$f_i\lvert x_i - \bar{x}\rvert$
5	7	35	9	63
10	4	40	4	16
15	6	90	1	6
20	3	60	6	18
25	5	125	11	55
	$\sum_{i=1}^{n} f_i = 25$	$\sum_{i=1}^{n} f_i x_i = 350$		$\sum_{i=1}^{n} f_i\lvert x_i - \bar{x}\rvert = 158$

$$Mean = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{350}{25} = 14$$

$$\text{M.D.}(\bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i - \bar{x} \right| = \frac{158}{25} = 6.32$$

(ii) Mean deviation about median

M.D. (median) $= \frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i - M \right|$, where M = median

Example 33: Find out the mean deviation about the median for following data.

Height (cm)	147	148	150	152	155
Number of students	8	12	15	10	5

Solution:

Height x_i	Frequency (f_i)	Cumulative frequency (cf)	$\lvert x_i - M \rvert$	$f_i \lvert x_i - M \rvert$
147	8	8	3	24
148	12	20	2	24
150	15	35	0	0
152	10	45	2	20
155	5	50	5	25
	50			$\sum_{i=1}^{n} f_i \lvert x_i - M \rvert = 93$

$$Median(M) = \frac{25^{th} + 26^{th}\, observations}{2} = \frac{150 + 150}{2} = 150$$

$$M.D.(median) = \frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i - M \right| = \frac{93}{50} = 1.86\ cm$$

(b) Continuous frequency distribution

A continuous frequency distribution is a series in which the data are classified into different class-intervals without gaps alongwith their respective frequencies. Recall the process of finding the mean and median for continuous frequency distribution and apply the same for putting the values of mean and median in calculating the mean deviation.

(i) **Mean deviation about mean**

$\mathrm{M.D.}(\bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i - \bar{x} \right|$,where you find out the mean ($\bar{x}$) by methods like short cut, assumed mean or step deviation method as discussed earlier.

(ii) **Mean deviation about median**

$$\mathrm{M.D.}\,(median) = \frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i - M \right|$$

where M (Median)= $l + \left(\frac{\frac{n}{2} - cf}{f} \right) \times h$

Example 34: Find the mean deviation about mean of following data.

Grade	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of students	2	3	8	14	8	3	2

Solution:

Marks obtained	Frequency (f_i)	Mid Points (x_i)	$f_i x_i$	$\left\|x_i - \bar{x}\right\|$	$f_i \left\|x_i - \bar{x}\right\|$
10-20	2	15	30	30	60
20-30	3	25	75	20	60
30-40	8	35	280	10	80
40-50	14	45	630	0	0
50-60	8	55	440	10	80
60-70	3	65	195	20	60
70-80	2	75	150	30	60
	$\sum_{i=1}^{7} f_i = 40$		1800		400

$$\sum_{i=1}^{7} f_i x_i = 1800, \text{ and } \sum_{i=1}^{7} f_i \left| x_i - \bar{x} \right| = 400$$

Therefore, mean deviation about mean, M.D. $(\bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i - \bar{x} \right| = \frac{1}{40} \times 400 = 10$

The procedure to calculate Mean Deviation about median is the exactly same as M.D. about Mean, except that deviations are to be taken from the median as shown in next example.

Example 35: Find the mean deviation about median of following data.

Class interval	20-30	30-40	40-60	60-80	80-90
Frequency	5	10	20	9	6

Solution :

Class interval (CI)	Frequency (f_i)	Mid Points (x_i)	Cf	$\lvert x_i - M \rvert$	$f_i \lvert x_i - M \rvert$
20-30	5	25	5	25	125
30-40	10	35	15	15	150
40-60	20	50	35	0	0
60-80	9	70	44	20	180
80-90	6	85	50	35	210
	$\sum_{i=1}^{7} f_i = 50$				665

$$Median = l + \left(\frac{\frac{n}{2} - cf}{f} \right) \times h = 40 + \frac{(25-15)}{20} \times 20 = 50$$

$$\text{M.D.}(Median) = \frac{1}{N} \sum_{i=1}^{n} f_i \lvert x_i - M \rvert = \frac{665}{50} = 13.3$$

Practical example of mean deviation: Consider that Govt. is planning to build a school in a city having population N. The suitable location of school is to be decided such that average distance covered by students is minimized.

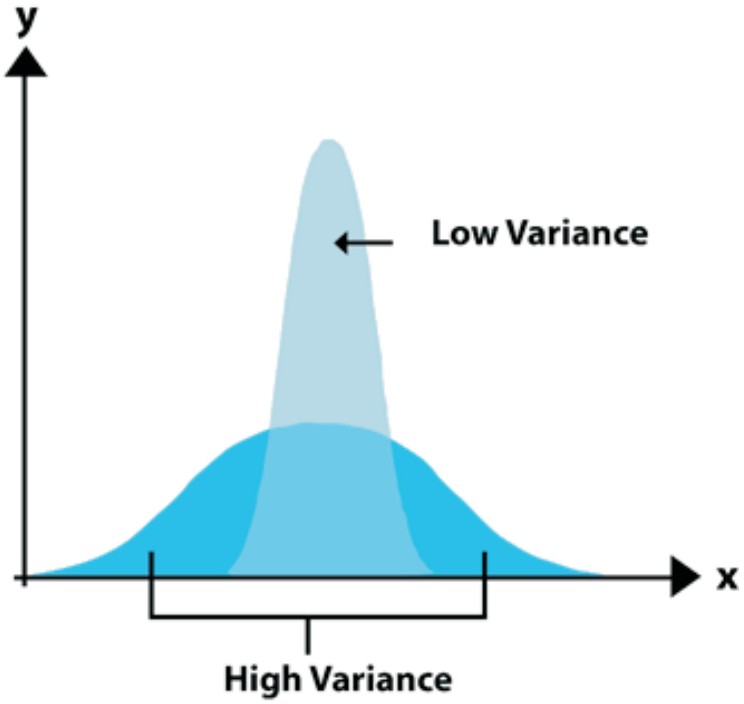
Let X_i be the distance of each student from any specific location point and 'a' be the distance of proposed site of school from same specific location. The distance covered by each student is $|X_i - a|$ and average distance is going to be MD. Now locate a point such that sum of all the $|X - a|$ is minimized. Basically we are minimizing the value of MD.

10.7 Variance

The variance is the average of the squares of the deviation of each value from mean and denoted by'σ'(read as sigma square). Therefore, the variance of n observations $x_1, x_2, x_3 \ldots\ldots x_n$ is given by;

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

So far discussed measures of dispersion (range, mean deviation, quartile deviation) describes only the measure of the theoretical averages but does not tell anything about the spread of the distribution. The variance combines all the values in a data set to produce a measure of the spread. Basically it is the arithmetic mean of the squared differences between each value and the mean value.



Example 36: Find the variance for the data set 10, 60, 50, 30, 40, 20.

Solution: The mean of data $(\bar{x}) = \frac{10+60+50+30+40+20}{6} = \frac{210}{6} = 35$

Variance $(\sigma^2) = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

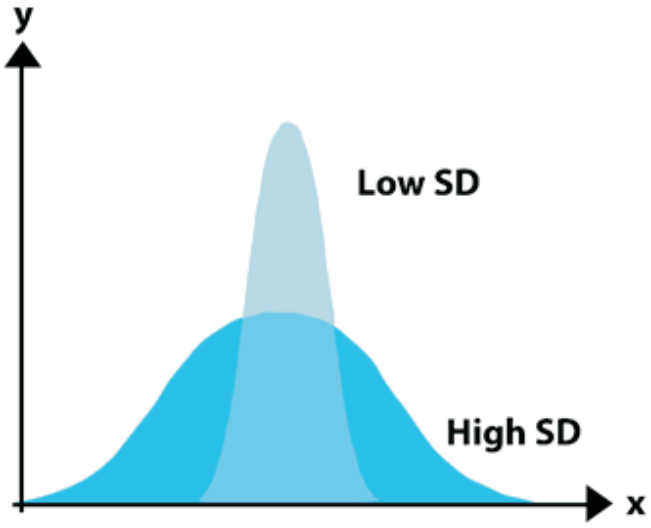
$$= \frac{(10-35)^2 + (60-35)^2 + (50-35)^2 + (30-35)^2 + (40-35)^2 + (20-35)^2}{6}$$

$$= \frac{1750}{6} = 291.7$$

10.7.1 Standard Deviation (SD)

Standard deviation (SD) measures variation (spread) in the data by finding the distances (deviations) between each data value and the mean (average).

The standard deviation (SD) gives an idea of the shape of the distribution and indicates how much variation there is from the mean and considered preferred measure of variability.



- Low SD indicates that the data points tend to be very close to the mean., thus very reliable
- High SD indicates that the data is spread out over a large range of values (large variance between data and average), thus not reliable.

The standard deviation is the square root of the variance and given by

Standard Deviation $(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

Excel formula for S.D.
= STDEV(data range)

Example 37: Find the standard deviation of data set 9, 3, 8, 8, 9, 8, 9, 18.

The mean of data $(\bar{x}) = \frac{9+3+8+8+9+8+9+18}{8} = \frac{72}{8} = 9$

Now, SD $(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

$$(\sigma) = \sqrt{\frac{(9-9)^2+(3-9)^2+(8-9)^2+(8-9)^2+(9-9)^2+(8-9)^2+(9-9)^2+(18-9)^2}{8}}$$

$$(\sigma) = \sqrt{15} = 3.87$$

Properties of Standard Deviation:

- Standard deviation is always positive.
- The outliers influence standard deviation. A single outlier can increase the standard deviation.

- For a data set having all the values same, the standard deviation is zero because each value is equal to the mean.

Effect of constant changes to the original data:

- If you add/subtract a constant value k to all the values, the arithmetic mean increases/decreases by k but the standard deviation remains the same.
- If you multiply/divide all the values by a constant value k, both the arithmetic mean and the standard deviation are multiplied/divided by k.

Standard Deviation of discrete frequency distribution

Let the given data consist of n distinct values $x_1, x_2, x_3, x_4, x_5 \ldots\ldots\ldots x_n$ having frequencies $f_1, f_2, f_3, f_4, f_5 \ldots\ldots\ldots f_n$ respectively. Then

$$(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \bar{x}\right)^2}$$ where $\bar{x}$ = mean and $N = \sum_{i=1}^{n} f_i$

Example 38: Thirty students were asked that how many days they require to revise the chapter of descriptive statistics. Their responses were: 4, 5, 6, 5, 3, 2, 8, 0, 4, 6, 7, 8, 4, 5, 7, 9, 8, 6, 7, 5, 5, 4, 2, 1 , 9, 3, 3, 4, 6, 4. Find out the standard deviation for this data.

No. of Days (x_i)	Frequency (f_i)	$(f_i x_i)$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
0	1	0	-5	25	25
1	1	1	-4	16	16
2	2	4	-3	9	18
3	3	9	-2	4	12
4	6	24	-1	1	6
5	5	25	0	0	0
6	4	24	1	1	4
7	3	21	2	4	12
8	3	24	3	9	27
9	2	18	4	16	32
	$\sum f_i = 30$	$\sum f_i x_i = 150$			$\sum f_i(x_i - \bar{x})^2 = 152$

means $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i} = \dfrac{150}{30} = 5$

Standard Deviation $(\sigma) = \sqrt{\dfrac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \bar{x}\right)^2} = \sqrt{\dfrac{152}{30}} = 2.25$

Standard Deviation of continuous frequency distribution:

The given continuous frequency distribution can be represented as a discrete frequency distribution by replacing each class by its mid-point. Then, the standard deviation is calculated by the technique adopted in the case of a discrete frequency distribution, and obtained by the formula

$(\sigma) = \sqrt{\dfrac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \bar{x}\right)^2}$ where mean and $N = \sum_{i=1}^{n} f_i$

Example 39: In a test of 50 students, marks secured by students is given in the following table. Calculate the standard deviation of scores.

Height (inch)	10-20	20-30	30-40	30-40	40-50
Number of students	5	8	15	16	6

Solution:

Class	Frequency (f_i)	Mid points (x_i)	$u_i = \dfrac{x_i - a}{h}$	$f_i u_i$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
0-10	5	5	-2	-10	484	2420
10-20	8	15	-1	-8	144	1152
20-30	15	25	0	0	4	60
30-40	16	35	1	16	64	1024
40-50	6	45	2	12	324	1944
	$\sum f_i = 50$			$\sum f_i u_i = 10$		$\sum f_i (x_i - \bar{x})^2 = 6600$

Mean $\bar{x} = a + \dfrac{\sum f_i u_i}{\sum f_i} \times h = 25 + \dfrac{10}{50} \times 10 = 27$

$(\sigma) = \sqrt{\dfrac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \bar{x}\right)^2} = \sqrt{\dfrac{6600}{50}} = \sqrt{132} = 11.48$

Short cut method of Standard Deviation S.D.:

When values of x_i in a discrete distribution or mid points of x_i in continuous distribution are large then we can apply following step deviation method

$$(\sigma) = \frac{h}{N}\sqrt{N\sum_{i=1}^{n} f_i y_i^{\,2} - \left(\sum_{i=1}^{n} f_i y_i\right)^2}$$

Where , $y_i = \frac{x_i - a}{h}$ a=Assumed mean and h = height of class

Example 40: In a survey, height of 100 students was observed and is given in the following table. Calculate the standard deviation of values.

Height (inch)	60-62	63-65	66-68	69-71	72-74
Number of students	5	18	42	27	8

Solution:

Class	Mid-point (x_i)	Frequency (f_i)	$y_i = \frac{x_i - a}{h}$	y_i^2	$(f_i y_i)$	$(f_i y_i^2)$
60-62	61	5	-2	4	-10	20
63-65	64	18	-1	1	-18	18
66-68	67	42	0	0	0	0
69-71	70	27	1	1	27	27
72-74	73	8	2	4	16	32
		100			$\sum f_i y_i = 15$	$\sum f_i y_i^2 = 97$

Let assumed mean (a) = 67

We know $(\sigma) = \frac{h}{N}\sqrt{N\sum_{i=1}^{n} f_i y_i^{\,2} - \left(\sum_{i=1}^{n} f_i y_i\right)^2}$

Where $y_i = \frac{x_i - a}{h}$ A=Assumed mean and h = height of class

So, $(\sigma) = \frac{3}{100}\sqrt{100(97) - (15)^2} = 2.92$ inch

Practical example of mean deviation: Suppose a weather forecasting agency is analyzing the high temperatures predicted for a series of days versus actual high temperature recorded on each of the day. A low value of SD would show reliable weather forecasting.

Check your Progress:

Questions 5: Find the standard deviation of the data 3, 6, 2, 1, 7, 5.

Question 6: Calculate standard deviation for the following set of scores: 40, 38, 42, 60, 72, 54.

Question 7: By multiplying each of the numbers 3, 6, 2, 1, 7, and 5 by 2 and then adding 5, we obtain the set 11, 17, 9, 7, 19, 15. What is the relationship between the standard deviations and the means for the two sets?

Use of measures of deviation

- Range is used when scores are spread broadly and need rapid, crude and at a glance measure.
- Quartile deviation is used when median has been used as a measure of central tendency.
- Standard Deviation (SD) is the most consistent and stable index of variability and used when mean has been used as measure of central tendency.

10.8 Skewness

Skewness denotes the degree of asymmetry in the data i.e tendency of a distribution to depart from symmetry. Skewness can be positive, negative or zero. The aim of measuring skewness is to know the direction of variation from an average and to compare the frequency distribution and shape of their curve.

Excel formula for Skewness
= SKEW(data range)

(i) **Positive Skewness**

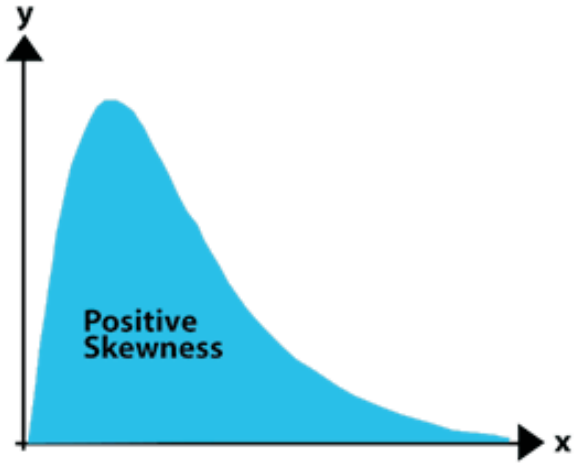
When skewness is positive, distribution is called "positively skew"

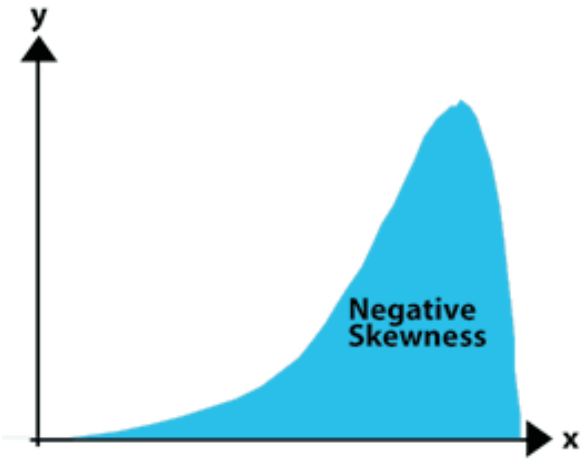
In this case, distribution has long tail on right and Mean > Median > Mode

(ii) **Negative Skewness**

When skewness is negative, distribution is called "negatively skew"

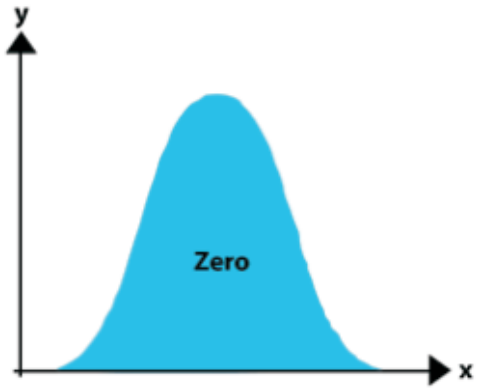
In this case, distribution has long tail on left and Mean < Median < Mode





(iii) **Zero Skewness**

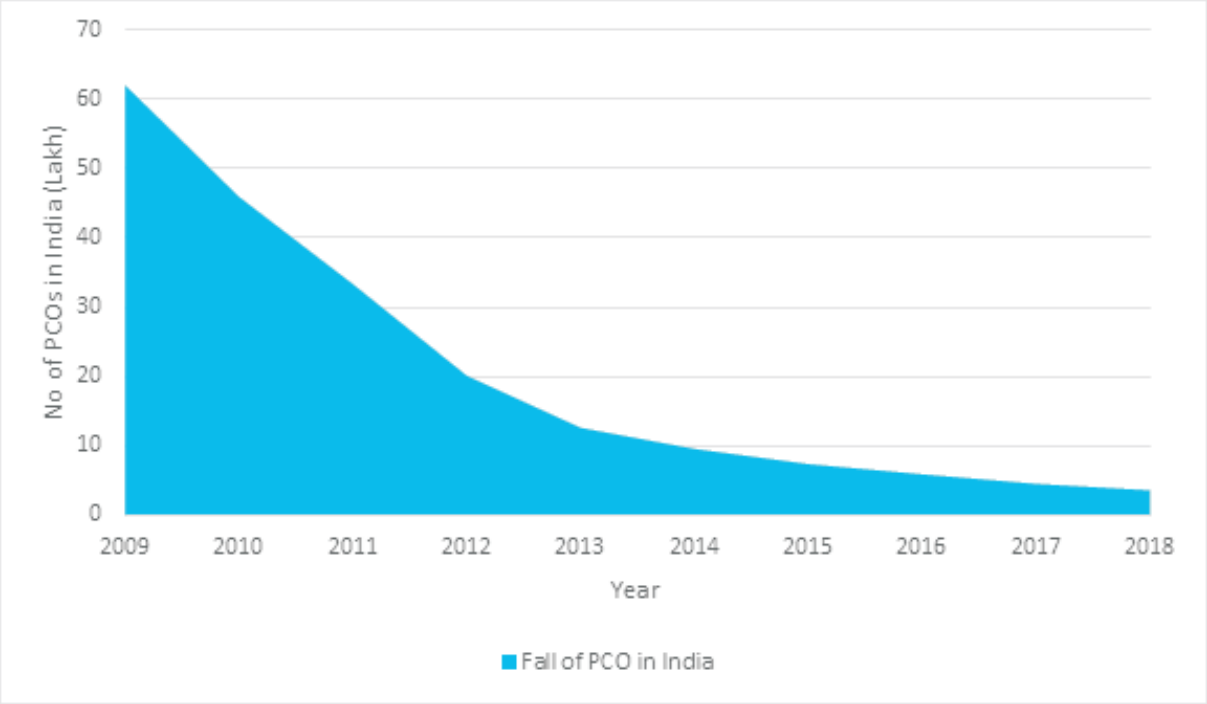
When skewness is zero, distribution is called "symmetrical"



Example 41: In the early years of the telecom sector, PCOs (Public Call Offices) became ubiquitous in India. However, spread of mobile services declined its number in subsequent years. Telecom Statistics India report (2018) gives following figures of PCOs in recent years

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
No of PCOs (in Lakh)	62	45.90	33.30	20.10	12.60	9.57	7.37	5.89	4.52	3.60

Source: Telecom Statistics India report (2018) (TRAI records)
https://dot.gov.in/sites/default/files/statistical%20Bulletin-2018.pdf



By plotting the given data, we get to know that distribution is positively skewed as it has long tail on right.

Measure of skewness: Coefficient of skewness

(i) Pearson's 1st coefficient of skewness $= \frac{Mean - Mode}{S.D.}$

(ii) Pearson's 2nd coefficient of skewness $= \frac{3(Mean - Median)}{S.D.}$

(iii) Quartile coefficient of skewness $= \frac{Q_3 + Q_1 - 2Q_2}{S.D.}$

(iv) Percentile coefficient of skewness $= \frac{P_{90} + P_{10} - 2Median}{P_{90} - P_{10}}$

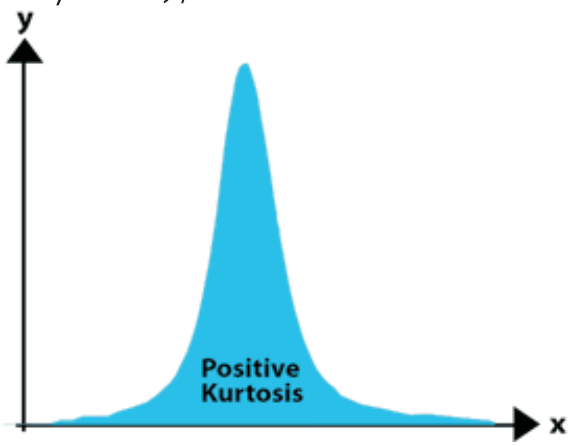
(v) Decile coefficient of skewness $= \frac{D_9 + D_1 - 2Median}{D_9 - D_1}$

10.9 Kurtosis

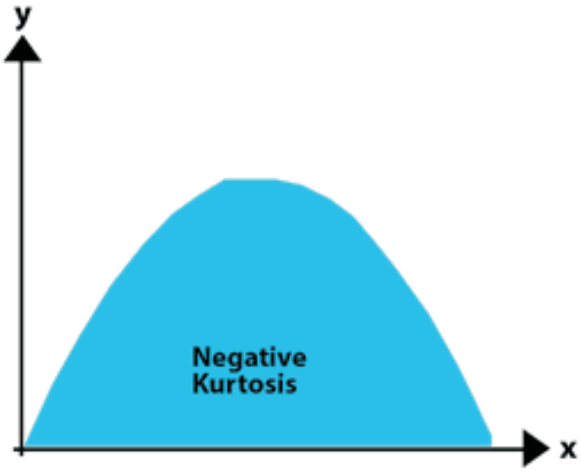
Kurtosis denotes the degree of 'peakedness' of frequency curve. It is used to specify the frequency curve as regards the sharpness of its peak. Kurtosis can be positive, negative or zero.

Excel formula for Kurtosis
= KURT(data range)

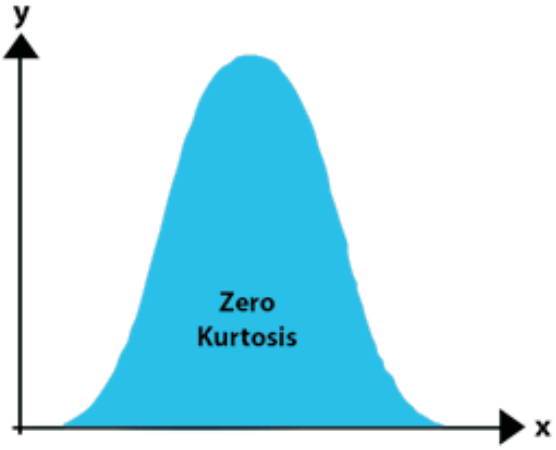
(i) **Positive Kurtosis (or leptokurtic):** A frequency distribution of data set having a **sharp peak** (heavy tailed) / outliers.



(a) **Negative Kurtosis (or platykurtic) :** A frequency distribution of data set having a blunt peak (light tailed)



(ii) **Zero Kurtosis (or mesokurtic):** A frequency distribution of data set having a moderate peak. This means the kurtosis is same as the normal distribution.



10.10 Measures of Position

Measures of position identifies the position of a value, relative to other values in a set of data. The most common measures are of position are percentiles, quartiles and deciles. Here we shall study only percentiles and quartiles.

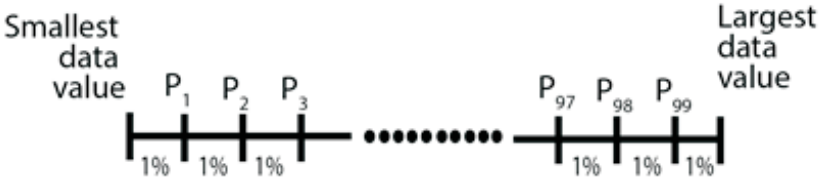
10.10.1 Percentile Rank

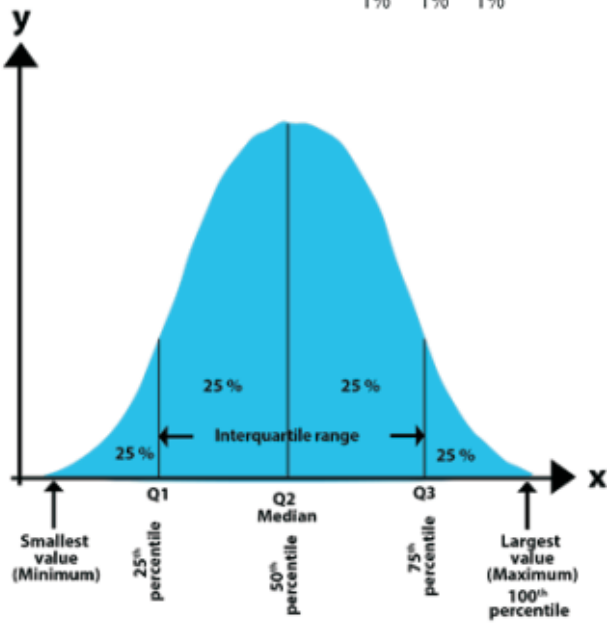
Percentile: Percentile indicate the position of an individual in a group by a number wherea certain percentage of scores fall below that particular number. The number below which P% of the values fall is called the Pth percentile

Suppose you scored 70 marks out of 80 in your Mathematics paper. This figure becomes more meaningful if you get to know that your score is 80th percentile, then it means you scored better than the 80% of students who took the same paper.

Now a days, percentile is very commonly used for declaring the results where number of candidates appeared is very large. It reflect your individual scoring in comparison to all.

Percentiles are denoted by $P_1, P_2, P_3 \ldots\ldots\ldots\ldots P_{99}$ and divide the distribution into 100 groups.





NOTE

- 25th percentile = First quartile (Q1)
- 50th percentile = Median = Second quartile (Q2)
- 75th percentile = Third quartile (Q3)
- Percentiles are not the same as percentages.

For a value X in a data set, X is the Pth percentile of the data if the percentage of the data that are less than or equal to X is P. The number P is the percentile rank of X.

Percentile rank: The percentile rank of a value is the percentage of entries smaller than that value. It is a measure of relative performance.

- 1st percentile = number such that 1% of the entries are smaller than the number, and 99% are larger
- 25th percentile = number such that 25% of the entries are smaller than the number, and 75% are larger
- The percentile rank of the highest secured marks is 100%
- The percentile rank of the median secured marks is 50%

NOTE: A percentile is a number.
A percentile rank is a percentage.

Excel formula for percentile (P_k)
= PERCENTILE(Data range, k in decimal form)

Percentile rank (PR) of ungrouped data (individual series):

Step 1: Put data in ascending order

Step 2: Find rank position (serial number) of individual score

Step 3: Percentile rank (PR) for a particular value Xi of data set is given by

$$PR = 100 - \left(\frac{100R - 50}{N}\right)$$

where R = Rank position of item within the distribution and N = Total number of items

Example 42: Suppose your class teacher conducted a 20 marks unit test. You scored 12 marks and others obtained 20, 10, 18, 15, 6, 8, 2, 3, 5, 17, 19, 20. Find out your percentile rank of your score.

Solution: Given data = 20, 10 18, 15, 6, 8, 2, 3, 5, 17, 19, 20, 12

Step 1: Put the given data in ascending order

2,3,5,6,8,10,12,15,17,17,18,19,20

We know that Percentile rank (PR) = $100-\left(\frac{100R-50}{N}\right)$

Here, score of 12 lies at 7th rank, so R = 7 & N = 13

PR = $100-\left(\frac{100\times 7-50}{13}\right)$

Hence, your score of 12 did better than 50% of the whole class.

Percentile ranks of other values can be find out by simply changing their respective rank positions.

Percentile rank of grouped data:

Step 1: Find cumulative frequency and cumulative frequency percentage of data

Step 2: Find the class interval, lower limit of class containing the score whose percentile rank is required, cf of preceding class, class size and total frequencies (i.e. N),

Step 3: Calculate x,(number of points required to be added to the exact value of the lower limit of class-interval in order to reach the score for which the PR is to be calculated)

i.e. x = Score whose percentile is required - Lower limit of class interval

Step 4: Percentile rank (PR) for a particular value Xi of data set is given by

PR = $\frac{100}{N}\left(cf+\frac{x}{h}\times f\right)$

Example 43: For the following frequency distribution, compute the percentile rank corresponding to the score of 66.

Class Interval	93-97	88-92	83-87	78-82	73-77	68-72	63-67	58-62	53-57	48-52
Frequency	4	7	5	8	3	6	7	10	5	4

Solution :

Class Interval	f	cf	Percentage of cf
93-97	4	59	100
88-92	7	55	93.22
83-87	5	48	81.36
78-82	8	43	72.88
73-77	3	35	59.32
68-72	6	32	54.24
63-67	7	26	44.07
58-62	10	19	32.30
53-57	5	9	15.25
48-52	4	4	6.78

Step 1: Find cumulative frequency and cumulative frequency percentage of data

Step 2: For score 66, Class interval is 63-67, l = 62.5 (by making it continuous),

Class size = 5 and N=59

Step 3: x = 66-62.5 = 3.5, f = 7, cf = 19

PR (for the score of 66) = $\frac{100}{59}\left(19+\frac{3.5}{5}\times 7\right)=40.50$

Thereby percentile rank of score 66 is 40.50

Check your Progress:

Question 8: Differentiate between percentile and Percentile rank.

Question 9: Calculate reasonable value for PR_{70} and PR_{80} for example no. 43 (given in Percentile rank of grouped data).

10.10.2 Quartile Rank

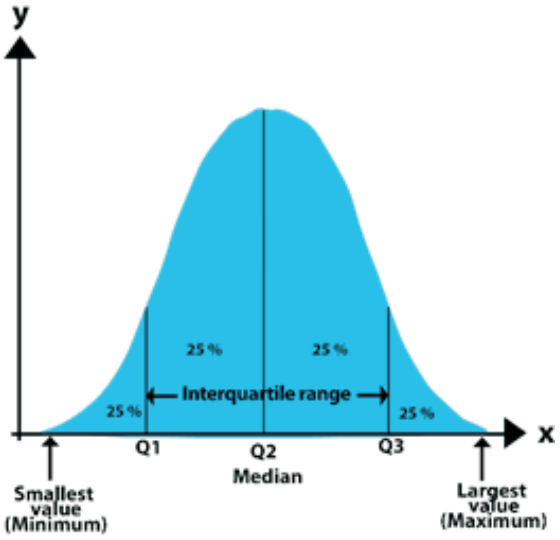
As we have already studied in section 6.6.2 that quartiles breaks the original set of data into four equal parts. Each of these parts contains one-quarter (25%) of the data where two halves are less than median and other two greater than median.

1. First quartile (Q1) contains upto 25% of data

2. Second quartile (Q2) contains 25% - 50% of data (up to the median) of data

3. Third quartile (Q3) contains 50% - 75% (above the median) of data

4. Fourth quartile (Q4) contains 75% - 100% of data



Quartile rank of ungrouped data (individual series):

Step 1: Arrange the values in ascending order and assign serial number to each.

Step2: Determine first quartile (Q1), third quartile (Q3) by following formula

Q1 = $\frac{1}{4}(n+1)^{th}$ Value

Q2 = $\frac{2}{4}(n+1)^{th} = \frac{(n+1)^{th}}{2}$ Value

Q3 = $\frac{3}{4}(n+1)^{th}$ Value

Q4 = $\frac{4}{4}(n+1)^{th}$ Value (This is not dividing the series)

Quartile rank of grouped data:

Step 1: Arrange the values in ascending order.

Step 2: Find out the quartile ranks using the formula

$$Q_1 = l + \left(\frac{\frac{n}{4} - cf}{f} \right) \times h \qquad Q_3 = l + \left(\frac{\frac{3n}{4} - cf}{f} \right) \times h$$

l = lower limit of class containing the score

n = total number of values

cf = cumulative frequency of preceding class

f = frequency of class interval

h = class size

In both individual and discrete series, Q_1 is the size of $\frac{1}{4}(n+1)^{th}$ value, but in a continuous distribution, it is the size of $\left(\frac{n}{4}\right)^{th}$ value. Similarly, for Q_3 and median (Q_2) also, n is used in place of n+1.

- For symmetrical distribution, mean and median are equal and median lies at an equal distance from the two quartiles

  i.e.Q3–Median = Median–Q1

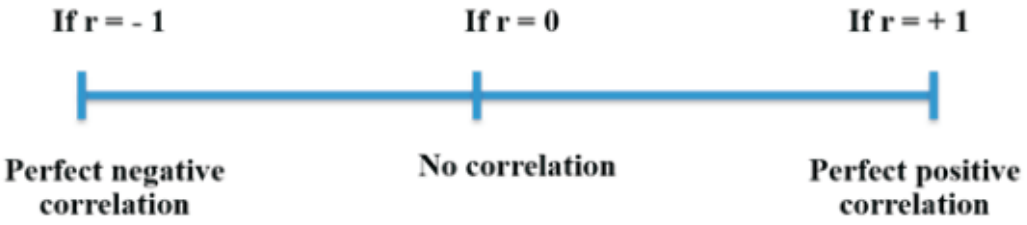
- For non-symmetrical distribution, two possibilities may arise:

  I. Q3–Median > Median–Q1 : (Positive Skewed Curve)

  II. Q3–Median < Median–Q1 : (Negative Skewed Curve)

10.11 Correlation

Correlation refers to a process for establishing relationships between two variables. It gives the extent of relation between two variables. "Scatter plot" is one of the way to get the idea about whether two variables are related or not. Theoretically, we can find out the association between two variables by calculating coefficient of correlation. It is denoted by r and value ranges from -1 to 1 i.e



If r = -1, perfect negative correlation

If r = +1, perfect positive correlation

If r > 0, positive correlation (when two variables are directly related)

If r < 0, negative correlation (when two variables are inversely related)

If r = 0, no correlation

Excel formula for Correlation
=CORREL(Variable 1 range, Variables 2 range)

Let given bivariate data (say variable x and y) is $x_1, x_2, x_3, ..., x_n$ and $y_1, y_2, y_3, ..., y_n$ then

Methods of computing correlation coefficient

1. Karl Pearson's coefficient of correlation / Product moment of coefficient of correlation :

(i) Direct method for ungrouped data (when deviations are taken from mean):

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 . \sum (y - \bar{y})^2}}$$

(ii) Direct method for grouped data:

$$r = \frac{N\sum XY - \sum X . \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} . \sqrt{N\sum Y^2 - (\sum Y)^2}}$$

(iii) When deviations are taken from assumed mean (Assumed mean method):

$$r = \frac{N\sum d_x d_y - \sum d_x . \sum d_y}{\sqrt{N\sum d_x^2 - (\sum d_x)^2} . \sqrt{N\sum d_y^2 - (\sum d_y)^2}}$$

Example 44: Result of two simultaneous tests conducted in a class have been given in following table. Calculate the Karl Pearson's coefficient of correlation for scores obtained by students in these two different tests.

Name of student	Priya	Rahul	Tanya	Priyanka	Sneha	Naveen	Neeraj	Sunil
Score in first test	1	3	4	6	8	9	11	14
Score in second test	1	2	4	4	5	7	8	9

Solution :

Name	**Score in first test (X_i)**	**Score in second test (Y_i)**	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
Priya	1	1	-6	36	-4	16	24
Rahul	3	2	-4	16	-3	9	12
Tanya	4	4	-3	9	-1	1	3
Priyanka	6	4	-1	1	-1	1	1
Sneha	8	5	1	1	0	0	0
Naveen	9	7	2	4	2	4	4
Neeraj	11	8	4	16	3	9	12
Sunil	14	9	7	49	4	16	28
	56	40		132		56	84

$$\bar{x} = \frac{56}{8} = 7 \text{ and } \bar{y} = \frac{40}{8} = 5$$

We know that Karl Pearson's coefficient of correlation $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 . \sum(y-\bar{y})^2}}$

$$r = \frac{84}{\sqrt{132 \times 56}} = 0.977$$

Hence there is a very high correlation between the given bivariate data.

Excel formula for Karl Pearson coefficient of Correlation
=PEARSON(Variable 1 range, Variables 2 range)

2. Spearman's rank correlation

Spearman rank correlation analysis can be used instead of Pearson's correlation coefficient when either given data is not normally distributed or there exist outliers. This method is not sensitive to outliers as it uses ranks for calculations, so value of observations does not affect.Spearman correlation coefficient is represented by r^s

In this method, the x and y variables are ranked and the ranks of x are compared to the ranks of y:

Steps of calculating coefficient of Spearman's correlation

Step 1: Put ranking to each value of both given variables (x, y) in descending order

Step 2: Find rank difference (d_i = rank x_i - rank y_i)of corresponding observations. If ranks of two items are same, then take their average.

Step 3: Square the rank difference

Step 4: Use formula given below to calculate the coefficient of Spearman's correlation

Case 1: If there are no tied ranks, then:

Spearman's rank correlation coefficient $r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}$

where d_i = rank x_i – rank y_i

n = number of observations

Case 2: If rank associated with two or more observations are same, then:

Spearman's rank correlation coefficient $r_s = 1 - \frac{6\left(\sum d_i^2 + \frac{\sum m\left(m^2-1\right)}{12}\right)}{n\left(n^2-1\right)}$

where d_i = rank x_i – rank y_i

n =number of observations

m= Number of times a particular rank has repeated itself

Example 45: Calculate the Spearman's rank correlation for the below given data;

IQ	99	120	98	102	123	105	85	110	117	90
EQ	3	0	30	45	16	25	17	24	26	5

Solution :

IQ (X_i)	EQ (Y_i)	Rank (X_i)	Rank (Y_i)	Rank diff (d_i)	$(d_i)^2$
99	3	4	2	2	4
120	0	9	1	8	64
98	30	3	9	6	36
102	45	5	10	5	25
123	16	10	4	6	36
105	25	6	7	1	1
85	17	1	5	4	16
110	24	7	6	1	1
117	26	8	8	0	0
90	5	2	3	1	1
					184

We know that Spearman's rank correlation coefficient $r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}$

$$r_s = 1 - \frac{6(184)}{10\left(10^2-1\right)} = -0.115$$

As the coefficient of above data is negative, we can say that there isn't much of a correlation between these two variables.

Excel formula for Spearman's rank correlation
=CORREL(rank range of Variable 1, rank range of Variable 2)

Question 10: ICC team ranking of men's Cricket for the month of May 2020 is given in following table. Calculate correlation coefficient for ODI and test rankings.

Sr. No.	Team	Rating of test matches	Rating of ODI matches
1	Australia	116	107
2	New Zealand	115	116
3	India	114	119
4	England	105	127
5	Sri Lanka	91	85
6	South Africa	90	108
7	Pakistan	86	102
8	West Indies	79	76
9	Bangladesh	55	88
10	Zimbabwe	18	39

Source: https://www.icc-cricket.com/rankings/mens/team-rankings/test

10.12 Applications of descriptive statistics using real time data

Descriptive statistics has various real-world applications and become part of business's arsenal now a day. It helps in planning strategies and making informed decisions. Some of the commonly used applications are summarized as

1. Conducting census, surveys, research and interpreting their scores widely uses various statistical tools. http://censusindia.gov.in/

2. Forecasting currency exchange rates, understanding the pattern of change in value of a particular currencies wr.t. some other currencies based upon the purchase power parity and further parameters.

   https://www.rbi.org.in/scripts/ReferenceRateArchive.aspx

3. Central Pollution Control Board (CPCB) advises the Govt. and citizens on matters concerning prevention and control of water, air pollution and improvement of the quality of air based upon the collection of water and air data and their statistical analysis. https://www.cpcb.nic.in/

4. Predicting natural calamities and weather based up the data sent by satellites and various observatories. https://mausam.imd.gov.in/

5. NavIC (Indian Regional Navigation Satellite System (IRNSS)) provides accurate position information service to users in India as well as some other extending regions based upon the data sent by constellation of 7 satellite https://www.isro.gov.in/irnss-programme

6. Big Data Analytic Companies (McKinsey, IBM, Deloitte, etc.) use it for consultation, based upon the statistical analysis of existing data.

7. For running marketing campaigns based upon the tabulation of past events such as localized sales, customer interests, arrangement of social metrics from Tweets, Facebook likes and followers.

8. Ranging from planning of national level policies / programs / schemes to pilot testing is always based upon the statistical analysis of financial market (Growth, Trends, Assumptions) and data previous data.

Online Resources

1. Statistics course on National Repository of Open Educational Resources Course:https://nroer.gov.in/home/topic_details/55b1f73181fccb7926fe54e2?nav_li=55b1f72181fccb7926fe5451,55b1f73081fccb7926fe54d5,55b1f73181fccb7926fe54e2

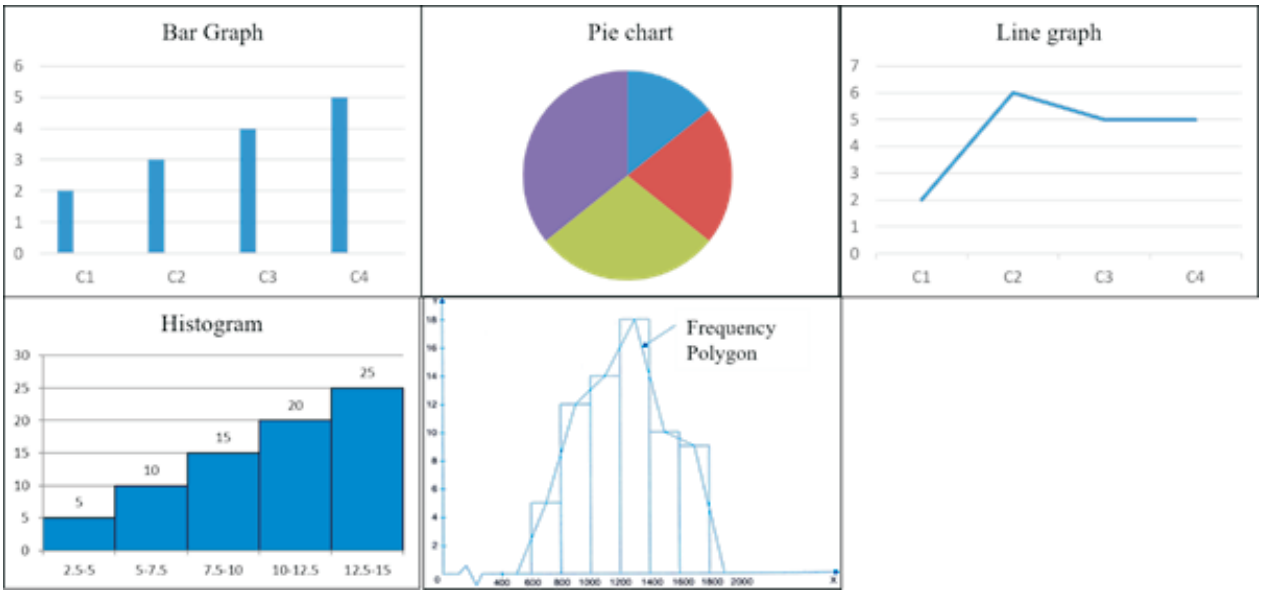
2. YouTube channel
http://www.zstatistics.com/videos/#/descriptive-statistics

3. Khan Academy Course:
https://www.khanacademy.org/math/in-in-grade-11-ncert/in-in-statistics-2

4. http://onlinestatbook.com/

Unit Summary

Representation of data:



Measures of central tendency (Mean is the most common measure of central location)		
Measure	**Definition**	**Symbol**
Mean	Sum of values, divided by total number of values	$\overline{X}$
Median	Middle point in data set that has been ordered	MD
Mode	Most frequent data value	None

Formulae			
	Individual Series	**Discrete Series**	**Continuous Series**
Measures of central tendency			
Mean Direct method Assumed mean method Step deviation method	$\frac{\text{Sum of all observations}}{\text{Number of observations}}$	$a+\frac{\sum f_i d_i}{\sum f_i}$ $a+\frac{\sum f_i u_i}{\sum f_i}\times h$	$\frac{\sum f_i x_i}{N}$ $a+\frac{\sum f_i d_i}{\sum f_i}$ $a+\frac{\sum f_i u_i}{\sum f_i}\times h$
Median	$\left(\frac{n+1}{2}\right)^{th}$ item, if odd observations Average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ items, if even observations		$l+\left(\frac{\frac{n}{2}-cf}{f}\right)\times h$ where size of $\left(\frac{n}{2}\right)^{th}$ item determines the median class.
Mode	Most frequent item	Most frequent item	$l+\left(\frac{f_1-f_0}{2f_1-f_0-f_2}\right)\times h$
Measures of dispersion			
Range			
Quartile Deviation	QD = $\frac{Q_3-Q_1}{2}$, Where Q1 = $\frac{1}{4}(n+1)^{th}$ value Q3 = $\frac{3}{4}(n+1)^{th}$ value	QD = $\frac{Q_3-Q_1}{2}$, Where Q1 = $\frac{1}{4}(n+1)^{th}$ value Q3 = $\frac{3}{4}(n+1)^{th}$ value	QD = $\frac{Q_3-Q_1}{2}$, Where Q1 is the size of $\left(\frac{n}{4}\right)^{th}$ value Q_3 is the size of $\left(\frac{3n}{4}\right)^{th}$ value

			$Q_1 = l + \left(\frac{\frac{n}{4} - cf}{f}\right) \times h$ $Q_2 = l + \left(\frac{\frac{3n}{4} - cf}{f}\right) \times h$
Mean Deviation MD about mean MD about median	$\frac{1}{N}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$ where $\bar{x}$= mean $\frac{1}{N}\sum_{i=1}^{n}\lvert x_i - M\rvert$ where M = median	$\frac{1}{N}\sum_{i=1}^{n} f_i\lvert x_i - \bar{x}\rvert$ mean $\frac{1}{N}\sum_{i=1}^{n} f_i\lvert x_i - M\rvert$ where M = median	$\frac{1}{N}\sum_{i=1}^{n} f_i\lvert x_i - \bar{x}\rvert$ where $\bar{x}$= mean $l + \left(\frac{\frac{n}{2} - cf}{f}\right) \times h$
SD	$(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$	$(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$	$(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$
	Short cut method of S.D. $(\sigma) = \frac{h}{N}\sqrt{N\sum_{i=1}^{n} f_i y_i^{\,2} - \left(\sum_{i=1}^{n} f_i y_i\right)^2}$ Note: S.D. is the most common descriptive measure of spread		
Measures of position			
Percentile Rank	$PR = 100 - \left(\frac{100R - 50}{N}\right)$		$PR = \frac{100}{N} + \left(cf + \frac{x}{h} \times f\right)$
Quartile Rank	$Q1 = \frac{1}{4}(n+1)^{th}$ $Q3 = \frac{3}{4}(n+1)^{th}$		$Q_1 = l + \left(\frac{\frac{n}{4} - cf}{f}\right) \times h$ $Q_3 = l + \left(\frac{\frac{3n}{4} - cf}{f}\right) \times h$

- **Skewness denotes the degree of asymmetry in the data**
  - **Positive skewness:** Distribution has long tail on right and **Mean > Median > Mode**
  - **Negative Skewness:** Distribution has long tail on left and **Mean < Median < Mode**
  - **Zero Skewness:** When skewness is zero

Measure of skewness: Coefficient of skewness
1. Pearson's 1st coefficient of skewness $= \frac{Mean - Mode}{S.D.}$
2. Pearson's 2nd coefficient of skewness $= \frac{3(Mean - Median)}{S.D.}$
3. Quartile coefficient of skewness $= \frac{Q_3 + Q_1 - 2Q_2}{S.D.}$
4. Percentile coefficient of skewness $= \frac{P_{90} + P_{10} - 2Median}{P_{90} - P_{10}}$
5. Decile coefficient of skewness $= \frac{D_9 + D_1 - 2Median}{D_9 - D_1}$

- **Kurtosis : Degree of 'peakedness' of frequency curve.**
  - **Positive Kurtosis:** A frequency distribution having a **sharp peak**
  - **Negative Kurtosis:** A frequency distribution having a **blunt peak**
  - **Zero Kurtosis:** A frequency distribution having a **moderate peak.**
- **Karl Pearson's coefficient of correlation:**

Ungrouped data: $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 . \sum(y-\bar{y})^2}}$

Direct method for grouped data: $r = \frac{N\sum XY - \sum X . \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} . \sqrt{N\sum Y^2 - (\sum Y)^2}}$

Assumed mean method: $r = \frac{N\sum d_x d_y - \sum d_x . \sum d_y}{\sqrt{N\sum d_x^2 - (\sum d_x)^2} . \sqrt{N\sum d_y^2 - (\sum d_y)^2}}$

- **Spearman's rank correlation coefficient**

When there are no tied ranks $r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}$

When rank of two or more observations are same $r_s = 1 - \frac{6\left(\sum d_i^2 + \frac{\sum m\left(m^2-1\right)}{12}\right)}{n\left(n^2-1\right)}$

Further Exploration

- **Additional resources**

Collect data from National Crime Record Bureau (NCRB) website https://ncrb.gov.in/ and graphically plot the various data in order to visualize the given data.

- How to calculate all the descriptive statistical results of given data in one go:

Step 1: Click the **File tab,** click Options, Click **Add-ins** category

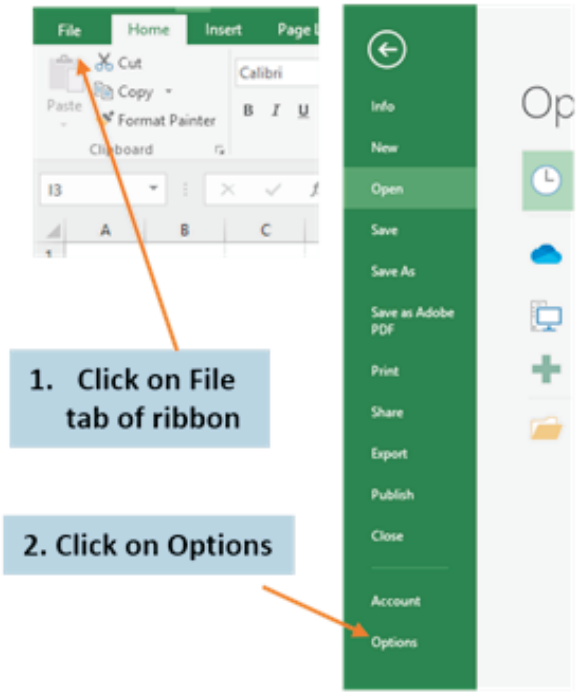
Step 2: In **Manage** box, select **Excel Add-ins** and click **Go**

Step 3: In the Add-Ins box, check the Analysis **Tool Pak & Solver Add-in** checkbox, and then click **OK**

This will bring the option of "Data Analysis" and "Solver" under the 'Data' tab of ribbon.

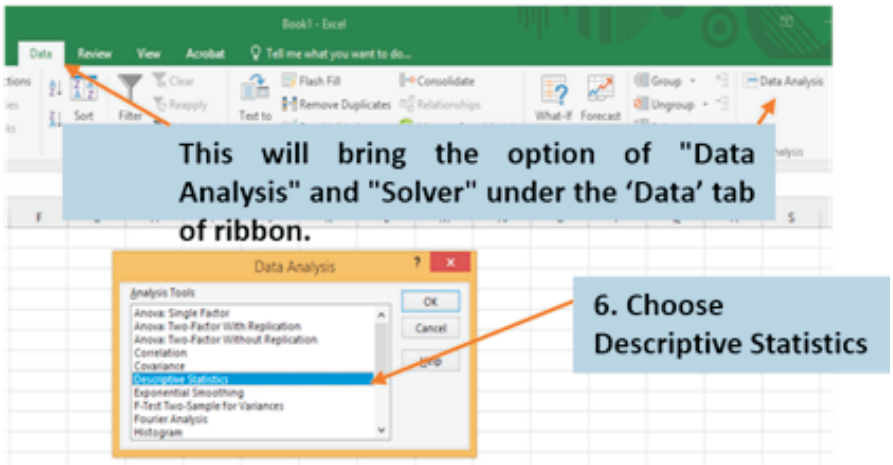
Step 4: Select the given data and output range with following steps

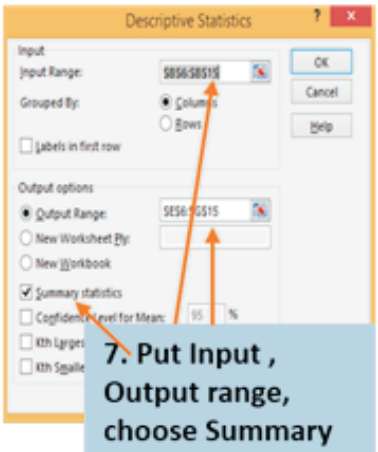
Click **Data** tab, Click **Data Analysis,** Click **Descriptive Statistics,** Select desired **Input and Output Range,** Choose **Summary Statistics** and then click **OK**


File
Home
Insert
Cut
Copy
Format Painter
Paste
Calibri
Clipboard
Info
New
Open
Save
Save As
Save as Adobe PDF
Print
Share
Export
Publish
Close
Account
Options
1. Click on File tab of ribbon
2. Click on Options


View and manage Microsoft Office Add-ins.
Add-ins
Manage:
Excel Add-ins
Go...
3. Click on Add-Ins
4. Click Go


Add-ins
Add-ins available:
Analysis ToolPak
Analysis ToolPak - VBA
Euro Currency Tools
Solver Add-in
OK
Cancel
Browse...
Automation...
5. Check Analysis ToolPak & Solver Add-in and click go
Analysis ToolPak - VBA
VBA functions for Analysis ToolPak


Data
Review
View
Acrobat
Data Analysis
This will bring the option of "Data Analysis" and "Solver" under the 'Data' tab of ribbon.
Analysis Tools
Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
OK
Cancel
6. Choose Descriptive Statistics


Descriptive Statistics
Input
Input Range:
Grouped By:
Columns
Rows
Labels in first row
Output options
Output Range:
New Worksheet Ply:
New Workbook
Summary statistics
Confidence Level for Mean:
Kth Largest:
Kth Smallest:
OK
Cancel
Help
7. Put Input , Output range, choose Summary Statistics and Click Ok