SHREE H.N.SHUKLA INSTITUTE OF PHARMACEUTICAL EDUCATION AND RESEARCH

B.PHARM

(SEMESTER-VIII)

SUBJECT NAME: BIOSTATISTICS AND RESEARCH METHODOLOGY

SUBJECT CODE: BP801TT

UNIT 1

Prepared by: Ms. Renuka Dabhi

Topic

Introduction

- Statistics
- Biostatistics
- Frequency distribution

Measures of central tendency

- Mean
- Median
- Mode
- Pharmaceutical Examples

Measures of dispersion

- Dispersion
- Range

- **Standard deviation**
- **Pharmaceutical Problems**

Correlation

- **Definition**
- **Karl Pearson's coefficient of correlation**
- **Multiple correlation**
- **Pharmaceutical Examples**

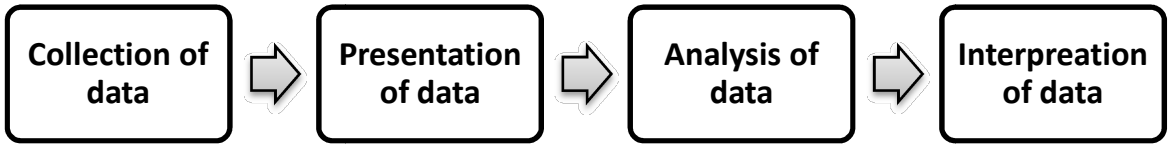
Introduction of Statistics

- Statistics is the technique to analysis the numerical data.
- It may define a universe or an entire population, based on various sampling procedure.
- It also includes the various techniques for the collecting as central tendency, tabulation, average, dispersion etc. which help in describing and summering the characteristic or feature of sampling of data in medical field.
- The word 'Statistics' has been derived from the Latin word 'Status'.
- In plural sense it means a set of numerical figures called 'Data' obtained by counting.
- Also, in the singular sense it means collection, classification, analysis, comparison and meaningful interpretation of 'Raw Data'.
- According to Croxton and Cowdon, Raw data as "It is the science which deals with the collection, analysis and interpretation of numerical data".
- Data collection in its original form is known as a raw data.
- Basically, statistics is a branch of applied mathematics and it may be defined as the collection, presentation, analysis and interpretation of numerical data.

Definition of Statistics

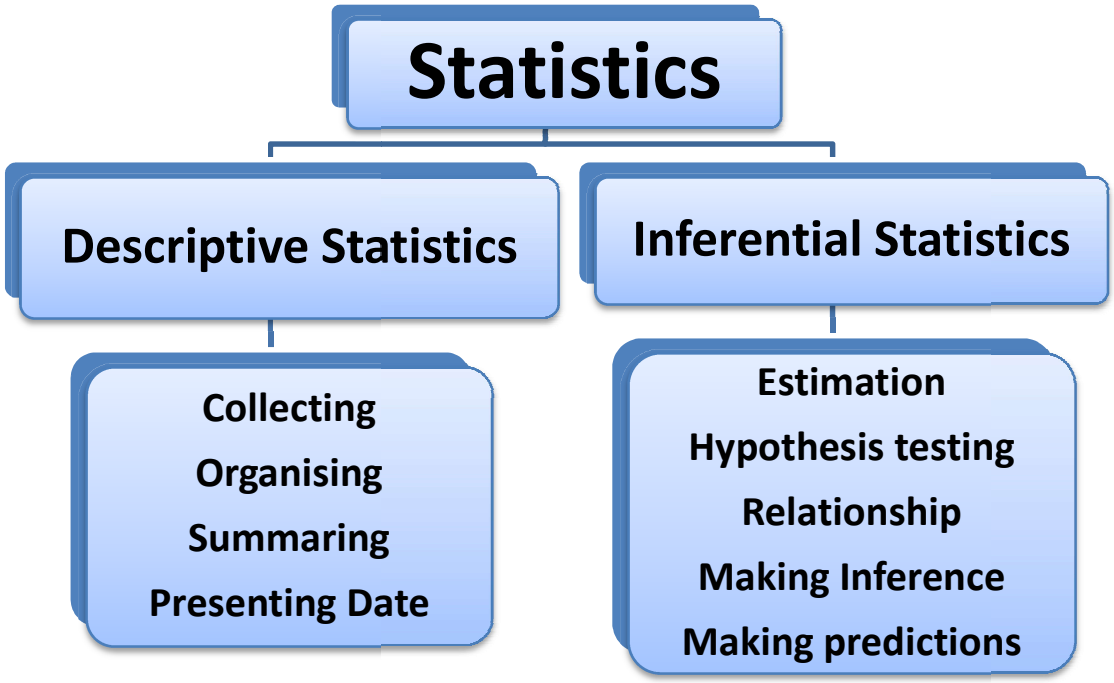
- Statistics means a measured or counted fact or piece of information stated as figure such as height of a person, birth of a body.
- It is a field of study concerned with various techniques or methods of collection, classification, summarizing, interpretation of data and drawing inference, testing hypothesis and making recommendations.
- Statistics is the science of collection, presentation, analysis and interpretation of numerical data from logical analysis.

- There are four main components of the statistics according to Croxton and Cowden.



Branches of Statistics

- Statistics is divided into two main categories, as Descriptive statistics and inferential statistics.



1) Descriptive Statistics:

- Descriptive Statistics are brief informational co-efficient that summarize a given data set, which can be either a representation of the entire population or a sample of a population.

- Descriptive Statistics are broken down into measures of central tendency and measures of variability.
- Methods used to summarize and describe the main features of a data set.
- Example; Measures of central tendency such as Mean, median and Mode which provides information about the typical value in the data set.
- Measures of dispersion like Range, Standard deviation and Variance.
- Uses; Descriptive statistics are specific methods basically used to calculate, describe and summarize collected research data in a logical, meaningful and efficient way.

2) Inferential Statistics:

- Inference means an idea or conclusion that's drawn from evidence and reasoning.
- Inferential statistics describe the many ways in which statistics derived from observations on samples from study populations can be used to deduced whether or not these populations are truly different.
- Example; Hypothesis testing, Confidence intervals, Regression analysis, Analysis of variance (ANOVA), Chi-square test.
- The goal of inferential statistics is to discover some property or general pattern about a large group by studying a smaller group of people in the hopes that the results will generalize to the larger group.

Introduction of Biostatistics

- Biostatistics is a branch of statistics applied to biological or medical sciences.
- Biostatistics covers, applications not only from health, medicines but also from field such as genetics, biology, drug discovery, epidemiology and many others. It mainly consists of various steps like generation of hypothesis, collection of

data and application of statistical analysis.

- It is branch of statistics that deals with data relating to living organism.
- Biostatistics can be able to answer many research questions in medicine and public health by using the tools of statistics. It comprises a set of principles and methods for generating and using quantitative evidence to address scientific questions.
- Biostatistics represents a key element of successful translation process that often generate an abundance of data on in-vitro tests, animal and clinical biomarkers and clinical endpoints.
- Selection of appropriate mathematical hypothesis, biological models and statistical tests are essential for adequate study design as a mandatory prerequisite for useful study outcomes.
- New statistical tools and software are often used to interpret the massive amounts of data and to detect correlations and causations.
- Biostatistics is used to study of a normal and a healthy population and for imposing limits for abnormality. In anatomy and physiology, used to study mean pulse rate, mean and variance of height and weight and their correlation in a healthy person.
- Biostatistics can be used for comparing the action of different drugs or dosage forms or to assess the relative potency of the drugs. It is used to compare the efficacy of important drug.
- Statistical tools are commonly used in community medicine and public health to find the usefulness of sera and vaccines. It is also used for comparison of vaccinated and unvaccinated death and to find whether the difference observed is statistically significant. It is useful for epidemiological studies to find the role of causative factors.

- Statistical tools are also useful in detection of reasons for reduction of birth rate may be because of higher age of marriage, family planning due to rise in living standards.
- Biostatistics is the study of data analysis and statistical reasoning applied practically to medicine and public health.
- In short, Biostatistics is a term used when the tools of statistics are applied to the data that is derived from biological sciences such as medicine. The tools and theories of statistics are very important in the field and medical sciences.

Application of Biostatistics

- To define, what is normal or healthy in a population?
- To find relative potency of a new drug with respect to a standard drug.
- To compare the efficiency of a particular drug.
- To find an association between two attributes such as disease and smoking, filariasis and social class.
- To identify sign and symptoms of a disease or syndrome.
- Biostatistics is used to check whether the difference between two populations is real or a chance occurrence for a particular attribute. It studies the correlation between two or more attributes in the same population.
- By control studies, it evaluate the efficiency of vaccines, sera etc.
- It is also used to evaluate the achievements of public health programs.
- Find out the normal values like pulse rate, blood pressure etc.
- Test usefulness of vaccines
- Leading cause of death, important causes of sickness, rise and fall of particular disease etc.

Frequency distribution

- Frequency is the number of occurrences of a repeating event per unit time/ Frequency is the number of times that any particular value comes in a data.
- For example;
  1) The frequency distribution of income in a population would show how many individuals (or households) have the income of a certain level (say 5000 a month)
  2) Consider the following data:

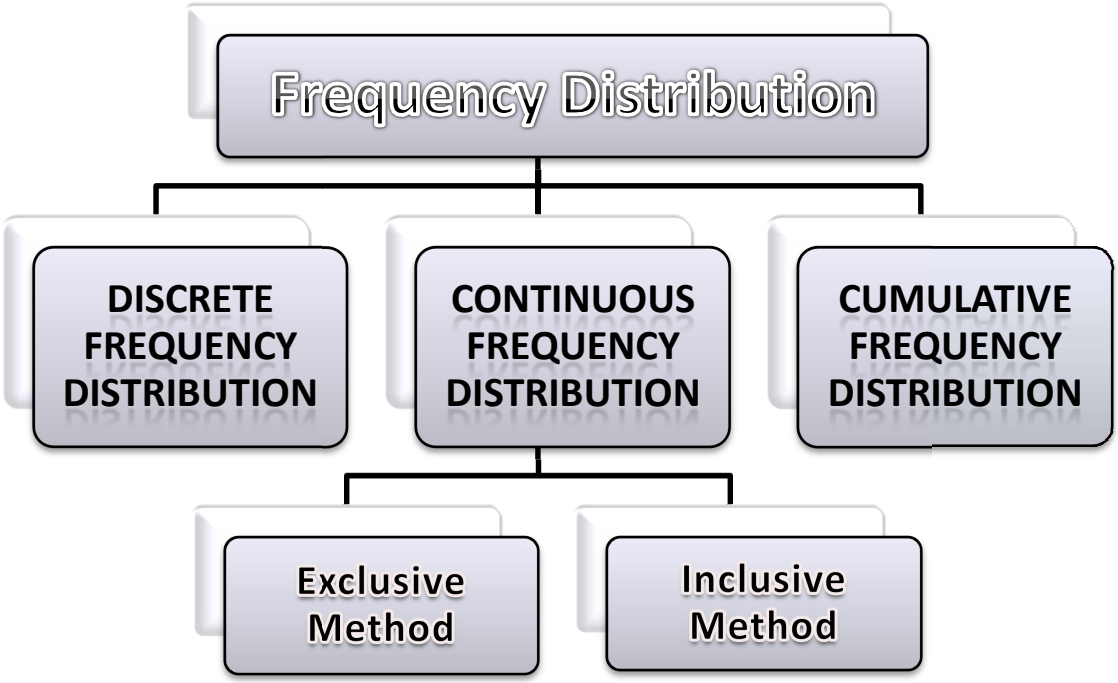
     Scores: 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 5

     The frequency of 2 is 5.

  - Frequency distribution is a series when a number of observations with similar or closely related values are put in separate bunches or groups, each group being in order of magnitude in a series.
  - It is simply a table in which the data are grouped into classes and the numbers of cases which fall in each class are recorded.
  - It shows the frequency of occurrence of different values of a single phenomenon.
  - According to Croxton and Cowden, "Frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in groups, with their corresponding frequencies side by side."
  - Frequency distribution is a representation, either in a tabular or graphical format that displays the number of observations within a given interval. The interval size depends on the data being analyzed and the goals of the analyst.
  - Frequency distribution provides a visual representation for the distribution of observations within a particular test.

Types of Frequency distribution

- Basically there are three type of frequency distribution.



1) Discrete or Ungrouped frequency distribution:

- In this form of distribution, the frequency refers to discrete value. Here the data are presented in a way that exact measurement of units is clearly indicated.
- There is definite difference between the variables of different groups of items. Each class is distinct and separate from the other class. Non-continuity from one class to another class exists.
- Data such as facts like the number of rooms in a house, the number of companies registered in a country, the number of children in a family, etc.
- The process of preparing this type of distribution is very simple. We have just to count the number of times a particular value is repeated, which is called the frequency of that class.

- In order to facilitate counting prepare a column for tally marks. In another column, place all possible values of variable from the lowest to the highest.
- Then put a bar (vertical line) opposite the particular value to which is relates. To facilitate counting, blocks of five bars are prepared and some space is left in between each block. We finally count the number of bars and get frequency.
- **For example;**

  In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

1	0	3	2	1	5	6	2
2	1	0	3	4	2	1	6
3	2	1	5	3	3	2	4
2	2	3	0	2	1	4	5
3	3	4	4	1	2	4	5

Represent the data in the form of a discrete frequency distribution.

Solution;

Number of Children	**Tally marks**	**Frequency**
0	|||	3
1	𝍸 ||	7
2	𝍸 𝍸	10
3	𝍸 |||	8
4	𝍸 |	6
5	||||	4
6	||	2
Total		**40**

2) Continuous or Grouped frequency distribution:

- Continuous series is one where measurements are only approximations and are expressed in class intervals. That means, within certain limits.
- **According to Boddington,** "the variable which can take any intermediate value between the smallest and longest value in the distribution."
- In a Continuous frequency distribution the class intervals theoretically continue from the beginning of the frequency distribution to the end without break.
- **For example;** Wage distribution of 100 employees

Weekly Wages in Rupees	Number of Employees
50-100	4
100-150	12
150-200	22
200-250	33
250-300	16
300-350	8
350-400	5
Total	**100**

Basic Component of Continuous Frequency Distribution

- The following are some basic technical terms when a continuous frequency distribution is formed or data are classified according to class intervals.

i) Class Limits:

- The class limits are the lowest and the highest values that can be included in the class.

- For example; take the class 30-40, the lowest value of the class is 30 and highest value of the class is 40.
- Two boundaries of class are known as the lower limits and the upper limit of the class. The lower limit of a class is the value below which there can be no item in the class. The upper limit of a class is the value above which there can be no item to that class.
- The way in which class limits are stated depends upon the nature of the data. In statistical calculations, lower class limit is denoted by L and upper class limit by U.

ii) Class Intervals:

- The difference between the lower limit and the upper limit of the class is known as the Class-interval.
- For example; in the class 10-20, the class interval is 10 (20-10=10).

Types of Class Intervals

- There are two method of classifying the data according to class intervals namely;

a) Exclusive Method:

- When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class, it is known as the exclusive method of classification.
- **For example;**

  Consider the following data, which are classified on the basis of this method.

Expenditure in Rupees	No. of families
0-5000	60
5000-10000	95
10000-15000	122
15000-20000	83
20000-25000	40
Total	**400**

- From the above table, it is clear that the exclusive method ensures continuity of data as much as the upper limit of one class is the lower limit of the next class.
- In the example, there are 60 families whose expenditure is between 0 and 4999.99. A family whose expenditure is 5000 would be included in the class interval 5000-10000.
- This method is widely used in practice.

b) Inclusive Method:

- In this method, the overlapping of the class intervals is avoided. Both the lower and upper limits are included in the class interval.
- This type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers in a factory etc. where the variable may take only integral values. It cannot be used with fractional values like age, height, weight etc.

- **For example;**

  Consider the following table which described the inclusive method

Class Interval	Frequency
5-9	7
10-14	12
15-19	15
20-29	21
30-34	10
35-39	5
Total	**70**

- Thus, to decide whether to use the inclusive method or the exclusive method, it is important to determine whether the variable under observation in a continuous or discrete one.
- In case of continuous variables, the exclusive method must be used. The inclusive method should be used in case of discrete variable.

iii) Range:

- The difference between largest and smallest value of the observation is called the Range and is denoted by 'R'.

  R = Largest value – Smallest value = L – S

iv) Mid value OR Mid point:

- The central point of a class interval is called the mid value OR midpoint.
- It is calculated by adding the upper and lower limits of a class and dividing the sum by 2.

$$Mid\ Value = \frac{L + U}{2}$$

- **For example;**

  If the class interval is 20-30 then the mid value is 25.

v) Frequency:

- Number of observations falling within a particular class interval is called frequency of that class.
- **For example;**

  Let us consider the frequency distribution of weights of persons working in a company.

Weight (in kgs)	Number of Persons
30-40	25
40-50	53
50-60	77
60-70	95
70-80	80
80-90	60
90-100	30
Total	**420**

- In this example, the class frequencies are 25, 53, 77, 95, 80, 60 and 30.
- The total frequency is 420. The total frequency indicates the total number of observations considered in a frequency distribution.

3) Cumulative frequency distribution:

- A cumulative distribution of frequencies shows the number of data items with values less than or equal to the upper class limit of each class.

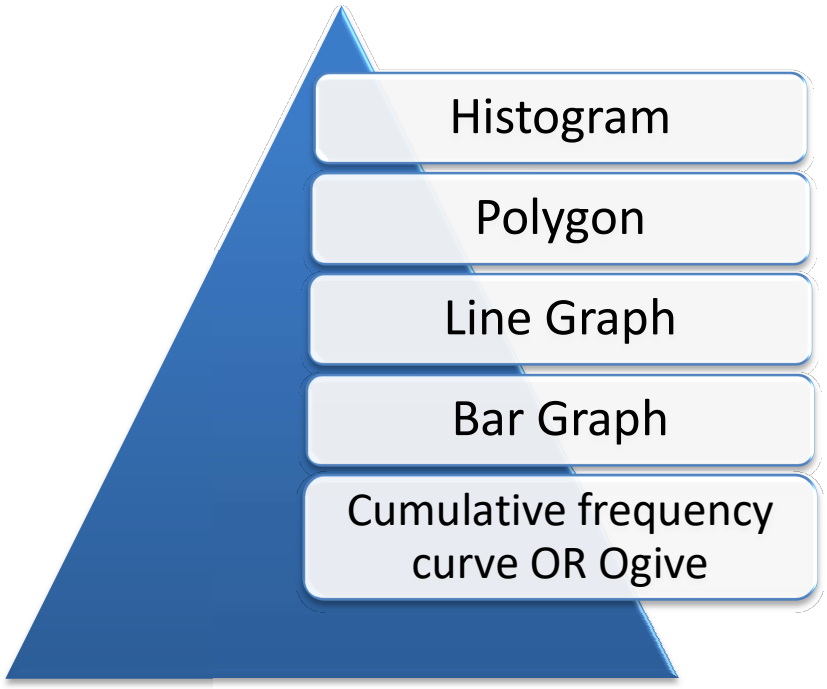
- For example; the following data represent GCS scores of table sharing the Cumulative frequency

GCS Score	Frequency (No. of Patients)	Cumulative frequency (Cumulative no. of Patients)
3	10	10
4	5	15
5	6	21
6	2	23
7	12	35
8	15	50
9	18	68
10	14	82
11	15	97
12	21	118
13	13	131
14	17	148
15	6	154

- The cumulative frequency for each category tells us how many subjects there are in that category, and in all the lesser-valued categories in the table.
- For example, 35 patients had a GCS score of 7 or less.

Graphical representation of Frequency distribution

- The important forms of frequency distribution graphs are as follows;



1) Histogram:

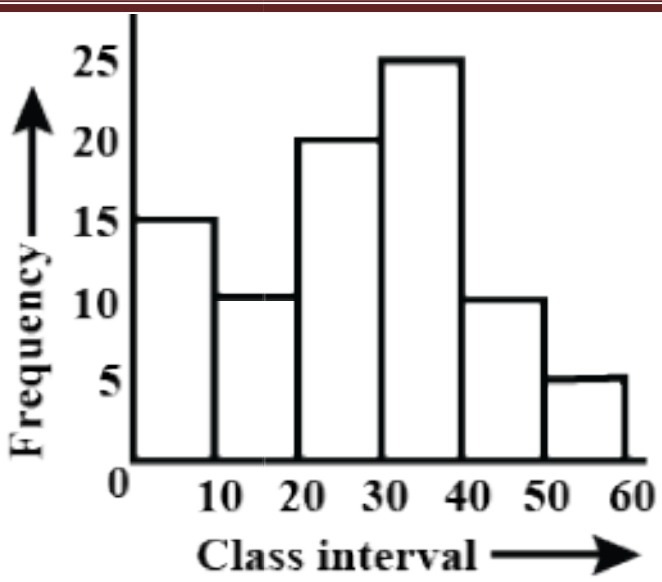
a) Histogram for equal class interval;

- A histogram is a representation of frequency distribution by a set of rectangular bars with area proportional to the class frequency.
- In graph, the classes are shown on the X-axis and the frequencies on the Y-axis.

- **For example;**

  Prepare a Histogram from the following table,

Class interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	15	10	20	25	10	5



b) Histogram for unequal class interval;

- In case of unequal class interval, to prepare a histogram use the following rules:
  - i) Divide the class interval into equal class interval
  - ii) Calculate the adjusted frequency by dividing the frequency of that class interval by 2.
- Similarly, follow the procedure for other unequal class intervals. The class interval is shown on the X-axis and the frequencies on the Y-axis.

- **For example;**

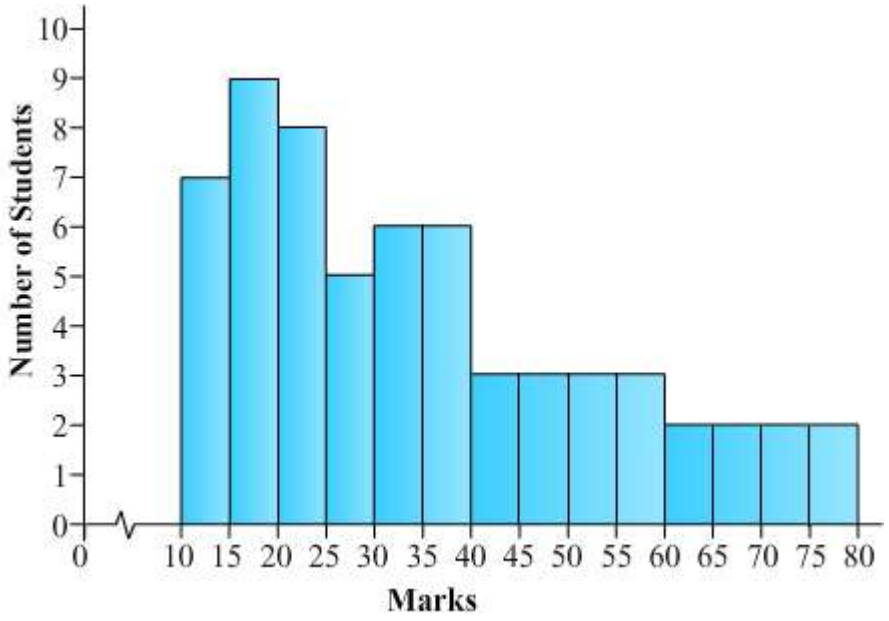
Prepare a Histogram from the following table,

Marks	10-15	15-20	20-25	25-30	30-40	40-60	60-80
No. of Students	7	9	8	5	12	12	8

Rearrange the given data and the adjusted frequencies and class intervals are given below:

Marks	10-15	15-20	20-25	25-30	30-35	35-40	40-45
No. of Students	7	9	8	5	6	6	3

45-50	50-55	55-60	60-65	65-70	70-75	75-80
3	3	3	2	2	2	2



c) Histogram for Inclusive data:

- To preparing Histogram for an Inclusive data, first it has to be converted into an Exclusive data.
- **For example;**

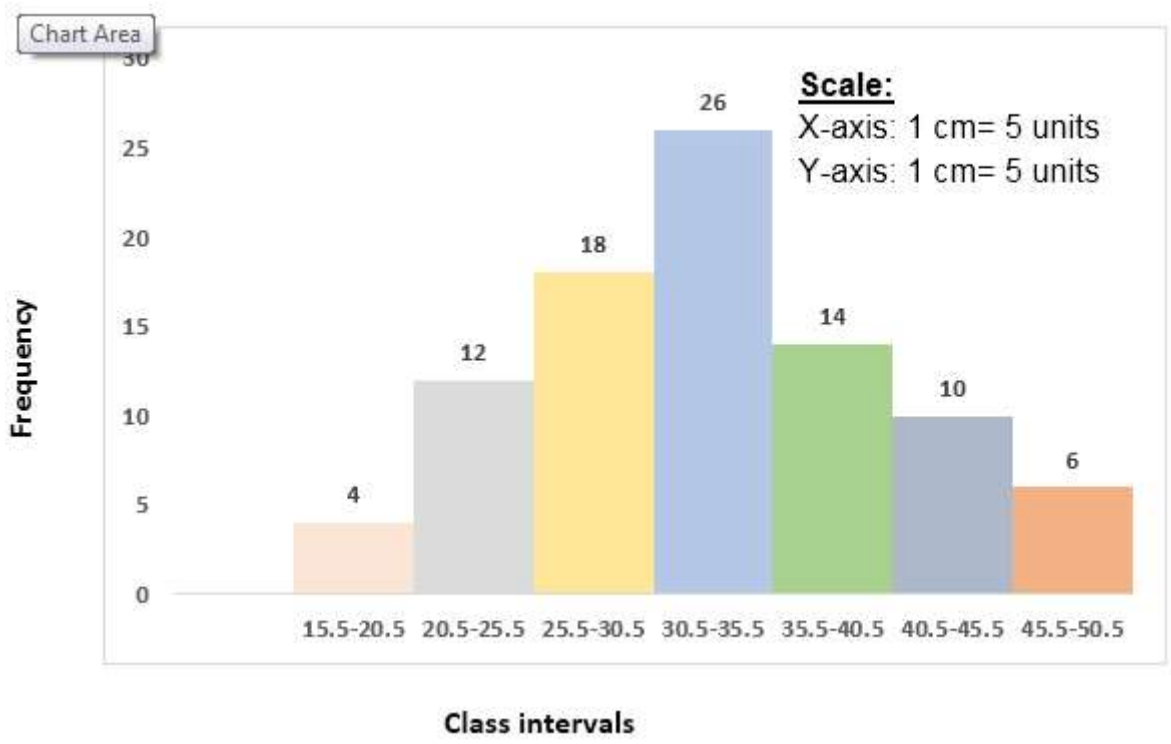
Draw a Histogram for the following data,

Class Interval	16-20	21-25	26-30	31-35	36-40	41-45	46-50
Frequency	4	12	18	26	14	10	6

- First convert given Inclusive data into Exclusive data,

Class Interval	15.5-20.5	20.5-25.5	25.5-30.5	30.5-35.5	35.5-40.5	40.5-45.5	45.5-50.5
Frequency	4	12	18	26	14	10	6

- Now, draw the graph for updated data



d) Histogram for Mid Value Series:

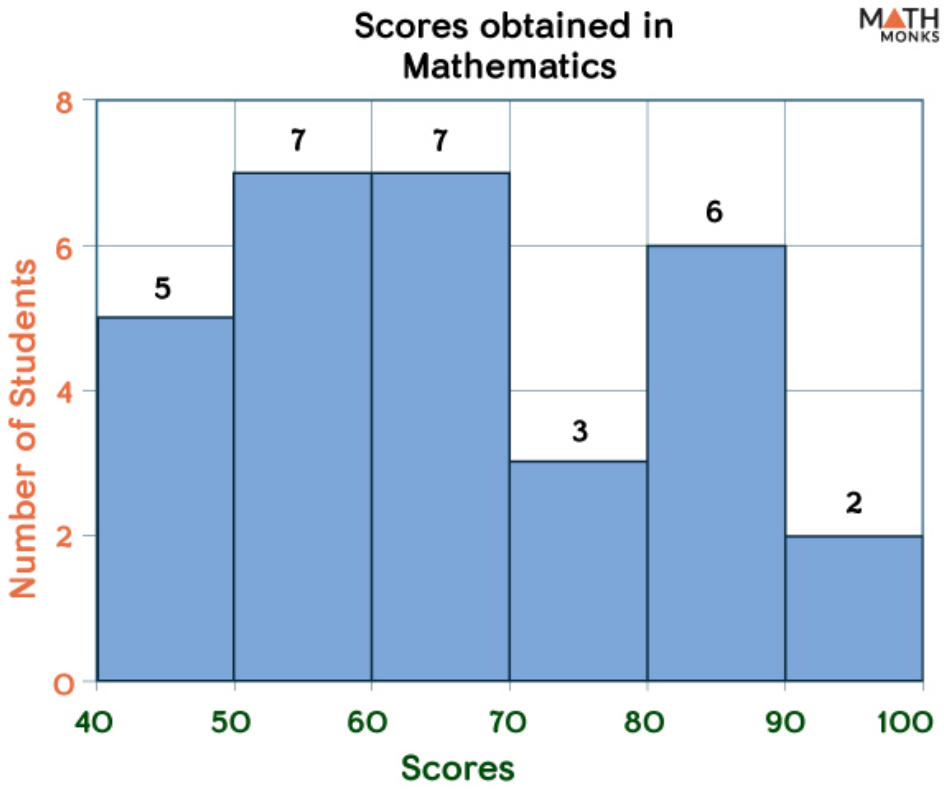
- To preparing histogram for mid value series, first it has to be converted into continuous series.
- **For example;**

  Construct a histogram with the help of the following data,

Mid Values (Score)	45	55	65	75	85	95
No. of Students	5	7	7	3	6	2

- Here, we convert given mid values series into a continuous series.

Class Interval (Score)	40-50	50-60	60-70	70-80	80-90	90-100
No. of Students	5	7	7	3	6	2



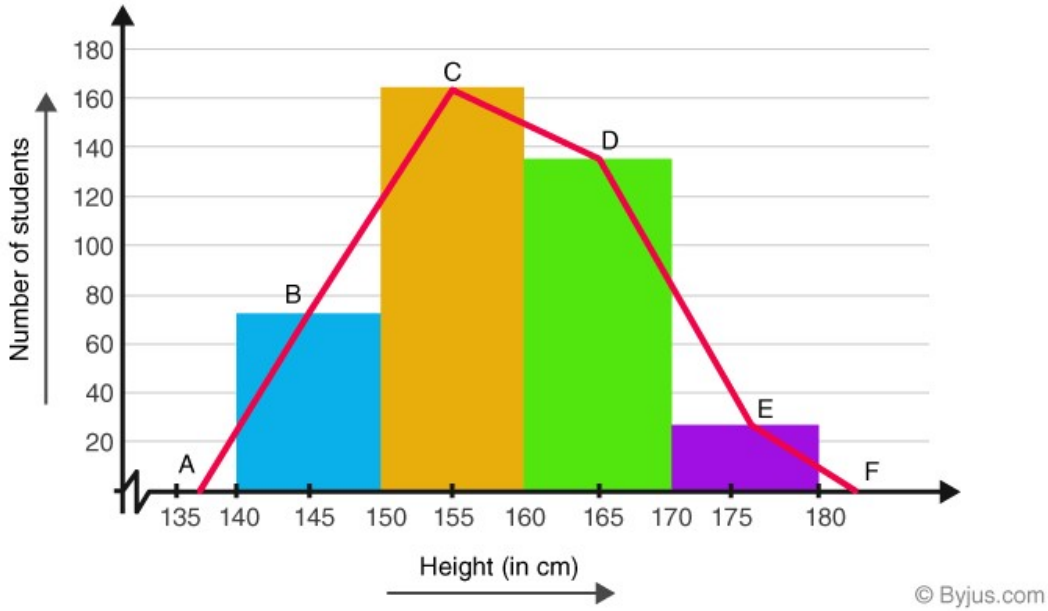
2) Polygon:

- Any graph, which has more than four sides, is called a Polygon.
- Frequency polygon is a graph in which the values of variable are taken on X-axis and the frequencies on the Y-axis.
- It is the curve obtained by joining the mid points of the tops of the rectangles in a histogram by the straight line.

- **For example;**

  Draw a frequency polygon from the following data,

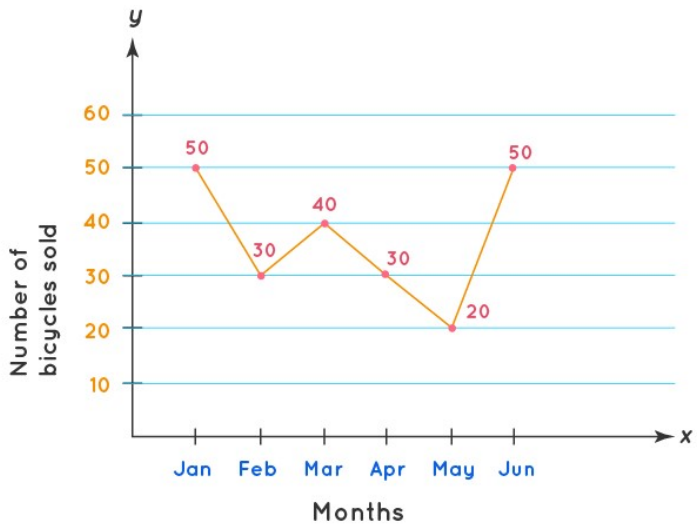
Height (in cm)	140-150	150-160	160-170	170-180
No. of Students	75	165	135	25



3) Line Graph:

- Line graph is used to depict a discrete data.
- In this graph the size is depicted on the X-axis and the frequencies on the Y-axis.
- Lines are drawn according to the given frequencies of different sizes.
- **For example;**

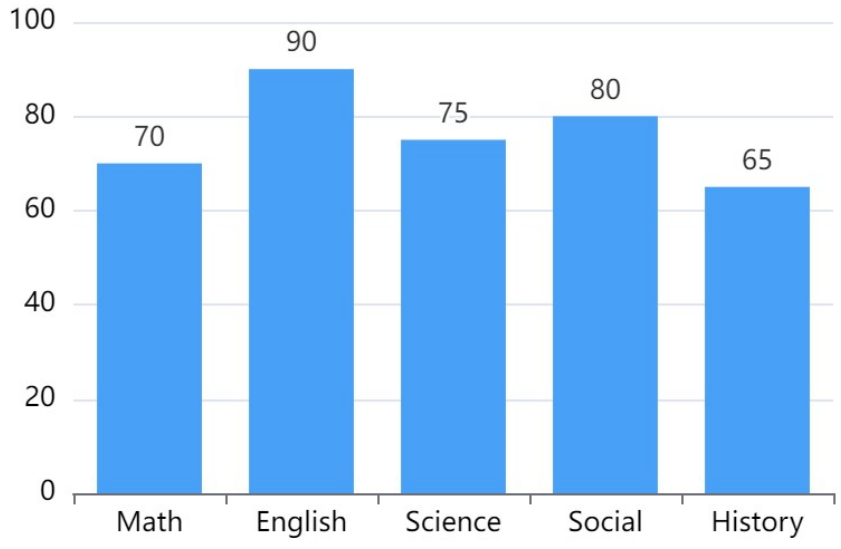
Months	Jan	Feb	Mar	Apr	May	June
No. of Bicycles sold	50	30	40	30	20	50



4) Bar Graph:

- Bar graph used to display the category of data and it compares the data using solid bars.
- **For example;**

Subject	Math	English	Science	Social	History
Marks	70	90	75	80	65



5) Cumulative frequency curve OR Ogive:

- Graph can be used to depict a cumulative frequency distribution. For drawing on ogive, an ordinary frequency distribution table is converted into cumulative frequency table.
- The cumulative frequencies are then plotted corresponding to the upper limits of the classes. The points, corresponding to cumulative frequency at each upper limit of the classes are joined by a free hand curve. The obtained graph is called an ogive.

- The ogive further classified as "less than ogive" and "more than ogive".
  a) Less than ogive: For drawing the less than ogive, the frequencies are added cumulatively in an increasing order.
  b) More than ogive: For drawing more than ogive, the cumulative frequencies of different classes are estimated in a diminishing order. The upper limit of a class interval and its respective cumulative frequency is considered for the construction of more than ogive.

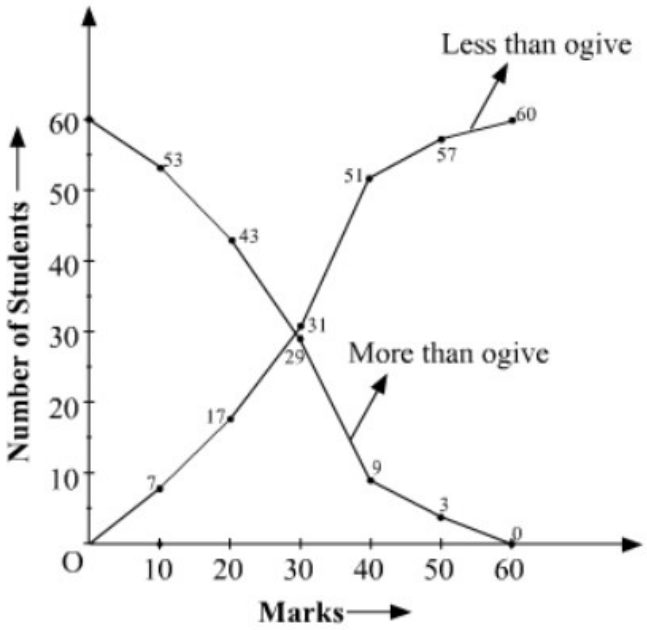
- **For example;**

  Construct the less than ogive and more than ogive from the following table,

Marks	0-10	10-20	20-30	30-40	40-50	50-60
Students	7	10	14	20	6	3

- To draw ogive, firstly we construct cumulative frequency table;

Less than Ogive		More than Ogive	
Marks	C.F.	Marks	C.F.
Less than 10	7	More than 0	60
Less than 20	7+10=17	More than 10	60-7=53
Less than 30	17+14=31	More than 20	53-10=43
Less than 40	31+20=51	More than 30	43-14=29
Less than 50	51+6=57	More than 40	29-20=9
Less than 60	57+3=60	More than 50	9-6=3
		More than 60	3-3=0

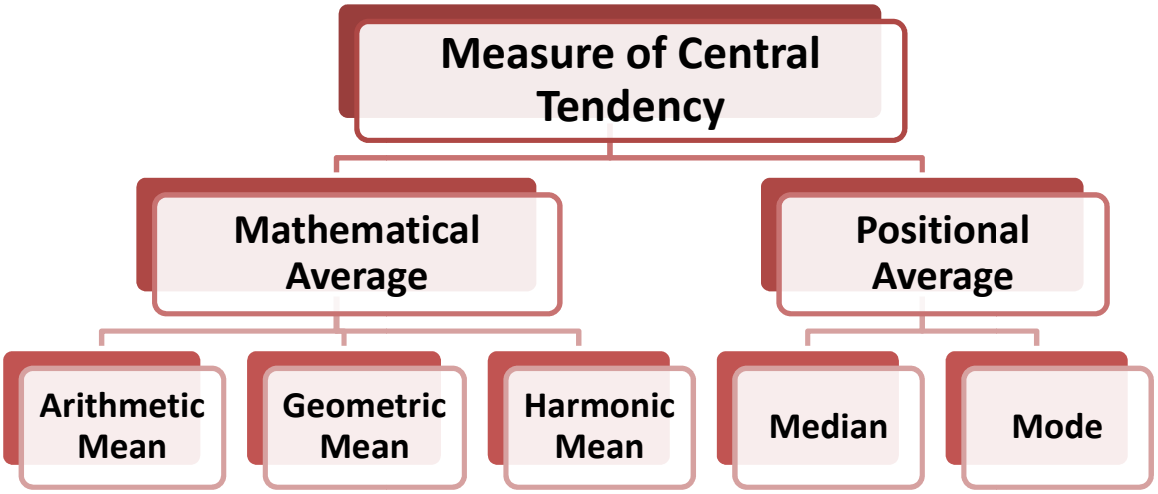


Measures of central tendency

- Measure of central tendency or an average refers to the value, which is used to represent an entire series.

- This property of concentration of the value around a central value is known as central tendency. The central value around which there is a concentration is called the measure of central tendency.
- Central tendency refers to the average value of any data.
- It is a measurement in which we calculate a single number (average) that represent the whole data.
- Measure of central tendency is also known as measure of central value OR measure of location OR average of first order.
- Average is a single figure which gives the complete picture of the phenomenon. For example, if we collect the data regarding heights of 2000 students of a school, we will not able to remember all 2000 numbers. So we calculate one single number known as average.
- Thus, it will be natural to expect that the representative should have closer to most of members of data.
- Hence, this representative is known as average and it lies more at the centre of the data due to which it is also known as central tendency.
- The main objective of calculating an average is to obtain a single number which will represent the whole data.
- Sometimes we need to compare one set of figures with another. Comparison of such bulky data is critical, so taking the average and comparison of data is made simpler and quicker.
- Taking an average is not the last stage of data analysis but it is starting point of calculations needed for further analysis.

- There are different types of averaging such as Mean, Median and Mode.



Mathematical Average

1) Arithmetic Mean:

- Arithmetic mean is the most commonly used measure of the central tendency.
- Its value is obtained by dividing the sum of the values of various items in a series by the number of total items.
- It is denoted by $\bar{x}$.

Calculation of Mean is an Individual series OR ungrouped data:

- Let $x_1, x_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots, x_n$ be the 'n' values of the variable x. Then the arithmetic mean,

$$\bar{x} = \frac{Sum\ of\ the\ values}{Total\ no\ of\ values} = \frac{x_1 + x_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots + x_n}{n} = \frac{\sum x}{n}$$

Example-1:

The birth weights of 6 babies are $2.0, 2.4, 2.6, 3.1, 3.4\ and\ 2.5kg$.

Find the mean birth weight.

Solution:

Here, $n = 6, x_1 = 2.0, x_2 = 2.4, x_3 = 2.6, x_4 = 3.1, x_5 = 3.4\ and\ x_6 = 2.5$

Then

$$\bar{x} = \frac{\sum x}{n} = \frac{2.0 + 2.4 + 2.6 + 3.1 + 3.4 + 2.5}{6} = \frac{16.0}{6} = 2.667kg$$

$$\boxed{\bar{x} = 2.667}$$

Exercise

1) The heights of 10 students are in cm: 160, 162, 175, 158, 156, 169, 173, 192, 165, and 167cm. find the mean height of students.
2) Find the mean from following data: 10, 8, 15, 12, 2, 9
3) Calculate the arithmetic mean of marks obtained in Pharmaceutical Microbiology by 10 students of B.Pharm

Students	A	B	C	D	E	F	G	H	I	J
Marks	6	10	15	6	18	17	12	14	8	14

4) The weekly wage of 5 workers is as given: 1350, 1400, 1450, 1370 and 1480 then find arithmetic mean.

Calculation of Mean is a Discrete series OR ungrouped data:

- There are two methods: Direct Method & Indirect Method (Assumed mean method)

Direct Method:

- To calculate mean follow the procedure:

- The value of each item (x) is multiplied by its frequency (f) and takes its total say $\sum fx$.
- Make sum of all frequencies and using to given formula to find mean

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Example-2:

From the following data, calculate the mean by direct method.

Class(x)	20	30	40	50	60	70
Frequency(f)	5	7	8	10	11	8

Solution:

Class(x)	**Frequency(f)**	**fx**
20	5	100
30	7	210
40	8	320
50	10	500
60	11	660
70	8	560
	$\sum f = 49$	$\sum fx = 2350$

Now,

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{2350}{49} = 47.96$$

$$\boxed{\bar{x} = 47.96}$$

Exercise

1) From the following distribution of data, calculate the mean.

Class(x)	10	20	30	40	50	60
Frequency(f)	3	2	5	10	11	8

2) Calculate the arithmetic mean from the following table that shows marks secured in pharmaceutics.

Marks(x)	40	48	52	58	64	69	74	78
No of Students(f)	5	2	7	8	5	3	2	1

3) Find average wages of 10 workers.

Daily Wage(x)	4	6	10	11	14	Total
No of Workers(f)	2	1	4	2	1	10

Indirect Method/ Assumed mean Method:

- Follow the steps of calculation of mean.
- Choose assumed mean (A) and calculate the value $dx = x - A$
- Multiply dx and its frequency(f) then obtain the total sum ($\sum f \cdot dx$)
- Using given formula to calculate mean,

$$\bar{x} = A + \frac{\sum f \cdot dx}{\sum f}$$

Example-3:

From the following frequency distribution finds out the mean weight of the 100 persons.

Weight(x)	64	65	66	67	68	69
No of persons(f)	15	13	18	5	20	11

70	71	72	73
7	6	3	2

Solution:

Weight(x)	**No of persons(f)**	$dx = x - A$	$f \cdot dx$
64	15	-4	-60
65	13	-3	-39
66	18	-2	-36
67	5	-1	-5
68=A	20	0	0
69	11	1	11
70	7	2	14
71	6	3	18
72	3	4	12
73	2	5	10
	$\sum f = 100$		$\sum f \cdot dx = -75$

Let assumed mean $A = 68$

$$\bar{x} = A + \frac{\sum f \cdot dx}{\sum f} = 68 + \frac{(-75)}{100} = 68 - 0.75 = 67.25Kg$$

$$\boxed{\bar{x} = 67.25}$$

Exercise

1) Calculate arithmetic mean from the following table that shows marks secured in pharmaceutics.

Marks(x)	40	48	52	58	64	69	74	78
No of Students(f)	5	2	7	8	5	3	2	1

2) From the following distribution of data, calculate the mean by using assumed mean method.

Class(x)	10	20	30	40	50	60
Frequency(f)	3	2	5	10	11	8

Calculation of Mean is an Continuous series OR Grouped data:

- There are two methods: Direct Method & Indirect Method (Assumed mean method and Step deviation method)

Direct Method:

- To calculate mean follow the procedure:
- The value of each item (x) is multiplied by its frequency (f) and takes its total say $\sum fx$.
- Make sum of all frequencies and using to given formula to find mean

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Example-4:

Calculate mean from the following data.

Age	18-20	21-23	24-26	27-29	30-32	33-35
Child(f)	18	20	32	7	2	1

Solution:

Class Interval/ Age	Frequency/ Child(f)	Mid Value (x)	$f \cdot x$
18-20	18	19	342
21-23	20	22	440
24-26	32	25	880
27-29	7	28	196
30-32	2	31	62
33-35	1	34	34
	$\sum f = 80$		$\sum f \cdot x = 1874$

$$Mid\ Value = \frac{Lower\ limit + Upper\ limit}{2} = \frac{18 + 20}{2} = 19$$

All mid values of given class interval find out from using above formula.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1874}{80} = 23.4$$

$$\boxed{\bar{x} = 23.4}$$

Exercise

1) From the following find out the mean profits.

Profits	100-200	200-300	300-400	400-500	500-600	600-700	700-800
No of Shares(f)	12	20	18	30	32	26	22

2) Find mean from following data,

Income	10-20	20-30	30-40	40-50	50-60	60-70
(f)	4	7	16	20	15	8

Indirect Method/ Assumed mean Method:

- Follow the steps of calculation of mean.
- Choose assumed mean (A) and calculate the value $dx = x - A$
- Multiply dx and its frequency(f) then obtain the total sum $(\sum f \cdot dx)$
- Using given formula to calculate mean,

$$\bar{x} = A + \frac{\sum f \cdot dx}{\sum f}$$

Example-5:

Calculate to arithmetic mean of the following distribution of patient and their weights.

Weights	30-40	40-50	50-60	60-70	70-80	80-90
No of Patients(f)	5	8	7	12	14	17

Solution:

Class Interval/ Weights	**Frequency/ No of Patients(f)**	**Mid Value (x)**	$dx = x - A$	$f \cdot dx$
30-40	5	35	-30	-150
40-50	8	45	-20	-160
50-60	7	55	-10	-70
60-70	12	65=A	0	0
70-80	14	75	10	140
80-90	17	85	20	340
	$\sum f = 63$			$\sum f \cdot dx = 100$

$$\bar{x} = A + \frac{\sum f \cdot dx}{\sum f} = 65 + \frac{100}{63} = 65 + 1.58 = 66.58$$

$$\boxed{\bar{x} = 66.58}$$

Exercise

1) Calculate mean from following data,

Age	20-30	30-40	40-50	50-60	60-70
No of workers (f)	8	15	12	9	6

2) Calculate mean from following data,

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No of Students (f)	5	10	25	30	20	10

Indirect Method/ Step deviation Method:

- Follow the steps of calculation of mean.
- Select an assumed mean and also calculate step deviation $du = \frac{x-A}{i}$
- Multiply step deviation with its frequency. $(f \cdot du)$
- Take the sum of $(f \cdot du)$ and use the following formula;

$$\bar{x} = A + \frac{\sum f \cdot du}{\sum f} \times i \quad , where\ i = class\ interval$$

Example-6:

Calculate to arithmetic mean from the following table,

Weights	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No of Children(f)	4	12	8	21	32	28	10	3	2

Solution:

Class Interval/ Weights	**Frequency/ No of Children(f)**	**Mid Value (x)**	$du = \frac{x - A}{i}$	$f \cdot du$
0-10	4	5	-5	-20
10-20	12	15	-4	-48
20-30	8	25	-3	-24
30-40	21	35	-2	-42
40-50	32	45	-1	-32
50-60	28	55=A	0	0
60-70	10	65	1	10
70-80	3	75	2	6
80-90	2	85	3	6
	$\sum f = 120$			$\sum f \cdot dx = -144$

$$\bar{x} = A + \frac{\sum f \cdot du}{\sum f} \times i = 55 + \frac{(-144)}{120} \times 10 = 55 - 12 = 43Kg$$

$$\boxed{\bar{x} = 43}$$

Exercise

1) Calculate the arithmetic mean from following data,

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency(f)	5	12	14	16	8

2) Calculate the arithmetic mean from following data,

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency(f)	6	28	51	11	4

3) Calculate the arithmetic mean from following data,

Class Interval	0-2	2-4	4-6	6-8	8-10	10-12
Frequency(f)	2	4	6	4	2	6

Application of Arithmetic Mean:

- Arithmetic mean is used to measure the standard deviation.
- Arithmetic mean is used in the construction of index numbers.
- It is also used in the hypothesis testing.

Advantages of Mean:

- It is simple to understand, easy to calculate and its calculations are based on observations.
- It takes into consideration all the scores in distribution and means of different distributions are useful for comparative purposes.
- It is least affected by fluctuations of sampling as compared with other measures.
- Some value is always determined, that means it is never indefinite.

- Can be used in other algebraic calculations.
- No need of sorting or arrangement (ascending or descending).
- It is stable and not affected by the variation of sampling.

Disadvantages of Mean:

- Mean is only quantitative measure, cannot be used for qualitative analysis.
- It cannot be used in case of open end classes.
- Mean is greatly affected by the extreme values.
- Sometimes mean may provide confusing impressions.
- The mean cannot be predicted by just inspecting the sample item.
- If a single value is missing, mean cannot be calculated.
- Graphical representation of mean is not possible.

2) Geometric Mean:

- Geometric mean is defined as the n^{th} root of the product of n values.
- It is denoted by GM and defined as,

  $GM = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \dots \dots \dots \dots \dots \dots \dots \cdot x_n}$ _____(1)

  Where, $x_1 \cdot x_2 \cdot x_3 \cdot \dots \dots \dots \dots \dots \dots \dots \cdot x_n$ are the various of the series and n=number of items.
- In case of large amount of items, we use logarithms to simplify calculates.
- From equation (1),

  $\log GM = \log \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \dots \dots \dots \dots \dots \dots \dots \cdot x_n}$

  $\log GM = \frac{1}{n}[\log x_1 + \log x_2 + \dots \dots \dots \dots \dots \dots + \log x_n]$

  $\log GM = \frac{1}{n}\left(\sum \log x\right)$

$$GM = Antilog\left[\frac{1}{n}\left(\sum \log x\right)\right]$$

Example-7:

Daily incomes of ten workers are given below.

Income (Rs.): 40, 45, 75, 70, 85, 500, 250, 36, 8, 15

Calculate the Geometric mean.

Solution:

Sr. No.	Income (x)	$\log(x)$
1	40	1.6021
2	45	1.6532
3	75	1.8751
4	70	1.8451
5	85	1.9294
6	500	2.6990
7	250	2.3979
8	36	1.5563
9	8	0.9031
10	15	1.1761
		$\boldsymbol{\sum(\log x) = 17.6373}$

Here, $n = 10$

We have,

$$GM = Antilog\left[\frac{1}{n}\sum \log x\right] = Antolog\left[\frac{1}{10}(17.6373)\right] = Antilog(1.76373)$$

$$= 58.040$$

$$\boxed{\boldsymbol{GM = 58.040}}$$

3) Harmonic Mean:

- The harmonic mean of n values is the reciprocal of the mean of the reciprocals of the values.
- It is denoted by HM and defined as

$$HM = Reciprocal\left[\frac{\sum reciprocals}{n}\right] = \frac{n}{\sum reciprocals}$$

$$HM = \frac{n}{\sum reciprocals}$$

Example-8:

Calculate the Harmonic mean from the following data,

X: 18, 12, 16, 21, 7, 9

Solution:

$\boldsymbol{X}$	$\boldsymbol{Reciprocals\left(\frac{1}{X}\right)}$
18	0.0556
12	0.0833
16	0.1667
21	0.4762
7	0.1429
9	0.1111
	$\boldsymbol{\sum\left(\frac{1}{X}\right) = 1.0358}$

$$HM = \frac{n}{\sum reciprocals} = \frac{6}{1.0358} = 5.7926$$

$$\boxed{\boldsymbol{HM = 5.7926}}$$

Positional Average

1) Median:

- Median of a set of values is the middle most value when the data is arranged in ascending or descending order of magnitude.
- The middle value will divide the whole data into two equal parts.
- The median is denoted by M.
- It is also called a positional average.

Calculation of Median is an Individual series OR ungrouped data:

- First given observations are arranged in ascending or descending order.
- If the number of observation (n) is odd then follow the given formula for find out the median,

$$Median(M) = Value\ of\ \left(\frac{n+1}{2}\right)^{th}\ term$$

- If the number of observation (n) is even then follow the given formula for find out the median,

$$Median(M) = \frac{Value\ of\ \left(\frac{n}{2}\right)^{th}\ term + Value\ of\ \left(\frac{n}{2}+1\right)^{th}\ term}{2}$$

Example-9:

Find out the median of the given items; 10, 15, 9, 25, 18

Solution:

First arrange the given data in ascending or descending order,

9, 10, 15, 18, 25

Here, $n = 5$

$$Median(M) = Value\ of\ \left(\frac{n+1}{2}\right)^{th}\ term = Value\ of\ \left(\frac{5+1}{2}\right)^{th}\ term$$

$$= Value\ of\ 3^{rd}\ term = 15$$

$$\boxed{\boldsymbol{M = 15}}$$

Example-10:

Find out the median of first 10 even natural numbers.

Solution:

Write down first 10 even natural numbers;

2, 4, 6, 8, 10, 12, 14, 16, 18, 20

Here, $n = 10$

$$Median(M) = \frac{Value\ of\ \left(\frac{n}{2}\right)^{th}\ term + Value\ of\ \left(\frac{n}{2}+1\right)^{th}\ term}{2}$$

$$= \frac{Value\ of\ \left(\frac{10}{2}\right)^{th}\ term + Value\ of\ \left(\frac{10}{2}+1\right)^{th}\ term}{2}$$

$$= \frac{Value\ of\ 5^{th} term + Value\ of\ 6^{th} term}{2} = \frac{10+12}{2} = 11$$

$$\boxed{\boldsymbol{M = 11}}$$

Exercise

1) Find out the median of the given items; 8, 11, 15, 18, 31, 30, 24
2) Find out the median of the given items; 46.4, 29.3, 48.2, 35.1, 46.4, 39.5, 41.3, 25.2
3) Find out the median of the given items; 9, 10, 3, 5, 9, 7, 10, 3
4) Find out the median of the given items; 16, 17, 10, 13, 20, 18, 13, 14, 18, 19

Calculation of Median is a Discrete series OR ungrouped data:

- First given observations are arranged in ascending or descending order.
- Calculate the total frequency ($\sum f = n$)
- Calculate the cumulative frequency.
- Locate the median by using the given formula;

$$Median(M) = \left(\frac{n+1}{2}\right)^{th} Value, \qquad where, n = \sum f$$

- The value, for which the cumulative frequency includes, $\left(\frac{n+1}{2}\right)^{th} Value$ is selected as a median.

Example-11:

To find the median of the following distribution;

X	1	5	6	3	2	4
f	8	20	25	16	12	19

Solution:

First of all given data convert into ascending order and also find out its cumulative frequency,

X	***f***	***cf***
1	8	8
2	12	8+12=20
3	16	20+16=36
4	19	36+19=55
5	20	55+20=75
6	25	75+25=100
	$\boldsymbol{n = 100}$	

$$Median(M) = \left(\frac{n+1}{2}\right)^{th} Value = \left(\frac{100+1}{2}\right)^{th} Value = 50.5^{th}\, Value$$

This value lies in cumulative frequency (55) for the value 4.

$\boxed{\boldsymbol{M = 4}}$

Example-12:

To find the median of the following distribution;

X	25	30	26	27	28	31	29
$\boldsymbol{f}$	15	17	16	18	20	19	17

Solution:

First of all given data convert into ascending order and also find out its cumulative frequency,

X	$\boldsymbol{f}$	$\boldsymbol{cf}$
25	15	15
26	16	31
27	18	49
28	20	69
29	17	86
30	17	103
31	19	122
	$n = 122$	

$$Median(M) = \left(\frac{n+1}{2}\right)^{th} Value = \left(\frac{122+1}{2}\right)^{th} Value = 61.5^{th}\, Value$$

This value lies in cumulative frequency (69) for the value 28.

$\boxed{\boldsymbol{M = 28}}$

Exercise

1) To find the median of the following distribution;

X	145	170	180	190	200	210
f	3	16	8	20	6	2

2) To find the median of the following distribution;

X	10	12	14	16	18	20	22
f	2	5	12	20	10	7	3

Calculation of Median is a Continuous series OR Grouped data:

- Median can be calculated with the use of following formula;

$$Median(M) = l + \frac{{}^{n}/_{2} - cf}{f} \times i$$

Where, l = lower limit of the median class

n = Total number of frequencies

f = Frequency of median class

cf = Cumulative frequency of before median class

i = Width of class interval

Example-13:

Find out median of the following data;

Class Interval	15-25	25-35	35-45	45-55	55-65	65-75
Frequency(f)	5	10	18	15	0	3

Solution:

Class Interval	f	cf
15-25	5	5
25-35	10	5+10=15
35-45	**18**	**15+18=33**
45-55	15	33+15=48
55-65	0	48+0=48
65-75	3	48+3=51
	$\boldsymbol{n = 51}$	

$\frac{n}{2} = \frac{51}{2} = 25.5 = 26^{th}\ element$

And this value lies in cumulative frequency (33) and its class interval is (35-45).

So, the median class is (35-45)

Here,

$\frac{n}{2} = 26, l = 35, f = 18, cf = 15, i = 10$

$$Median(M) = l + \frac{{}^{n}/_{2} - cf}{f} \times i = 35 + \frac{26 - 15}{18} \times 10 = 35 + \frac{11}{18} \times 10$$

$$= 35 + 6.11 = 41.11$$

$\boxed{\boldsymbol{M = 41.11}}$

Exercise

1) To find the median of the following distribution;

Class Interval	60-70	70-80	80-90	90-100	100-110
Frequency(f)	8	10	12	16	14

2) To find the median of the following distribution;

Class Interval	80-100	100-120	120-140	140-160	160-180
Frequency(f)	8	12	16	8	6

Advantages of Median:

- It is unaffected by the extreme values.
- In case of individual observation and discrete series, it is easy to understand and calculate.
- It can be used in other algebraic calculations and also useful during calculation of mean deviation.
- It can also be measured by using 'Graphical representation'.
- Median is clearly definite in nature.

Disadvantages of Median:

- Sorting (ascending and descending order) is necessary for the calculation.
- It cannot be used to measure the combined mean of two or more groups.
- It is not based on all the data samples. It is a positional average.

2) Mode:

- Mode is the value of the variable for which the frequency is maximum and it is denoted by Z.

Calculation of Median is an Individual series/ Discrete series OR ungrouped data:

- If the data is not arranged then it has to be arranged in increasing order.

- Now find the value occurring the maximum number of times, this value is called the 'Mode' of the given data.

Example-14:

Calculate mode from the following data;

10, 27, 24, 18, 27, 27, 20, 18, 15, 32

Solution:

Since the item 27 occurs the maximum number of times that means 3.

$\boxed{Z = 27}$

Example-15:

Calculate mode for given discrete data;

Fine (Rs.)	100	120	140	160	180
No of students	15	18	25	20	17

Solution:

Since the value 140 is occurring highest number of time that means 25. So the value of mode is 140.

$\boxed{Z = 140}$

Exercise

1) Find the mode of given a set of data;
   46, 47, 48, 47, 40, 50, 97, 52
2) Calculate mode for given discrete data;

Wages	145	170	180	190	200	210
Employees	3	16	8	20	6	2

Calculation of Median is a Continuous series OR Grouped data:

- Mode can be calculated with the use of following formula;

$$Mode(Z) = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Where, l = lower limit of the modal class

f_1 = Frequency of modal class

f_0 = Frequency of previous class

f_2 = Frequency of next class

i = Width of class interval

Example-16:

Calculate mode for following data;

Data	8-9	9-10	10-11	11-12	12-13	13-14	14-15
Frequency	8	14	21	25	15	10	7

Solution:

In given table, we have modal class (11-12) with highest frequency 25.

$l = 11, f_1 = 25, f_0 = 21, f_2 = 15, i = 1$

$$Mode(Z) = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 11 + \frac{25 - 21}{2(25) - 21 - 15} \times 1$$

$$= 11 + \frac{4}{50 - 36} \times 1 = 11 + \frac{4}{14} = 11 + 0.28 = 11.28$$

$$\boxed{\boldsymbol{Z = 11.28}}$$

Exercise

1) Find the mode for following data,

Prize	2-6	6-10	10-14	14-18	18-22	22-26
Frequency	1	9	21	47	52	36

2) Calculate mode for following data;

Prize	50-55	55-60	60-65	65-70	70-75	75-80
Frequency	3	8	14	20	16	2

Advantages of Mode:

- Easy to understand and calculate.
- It is an actual value, which most frequently occurs in the series.
- Not affected by extreme values.
- It is simple and accurate and can be measured in an open end class interval without determining the class limits.

Disadvantages of Mode:

- It is ill defined and in few cases it is not possible to find a definite value.
- Not a good representative because it is not based on all observations.
- In further algebraic calculation it is of no use.

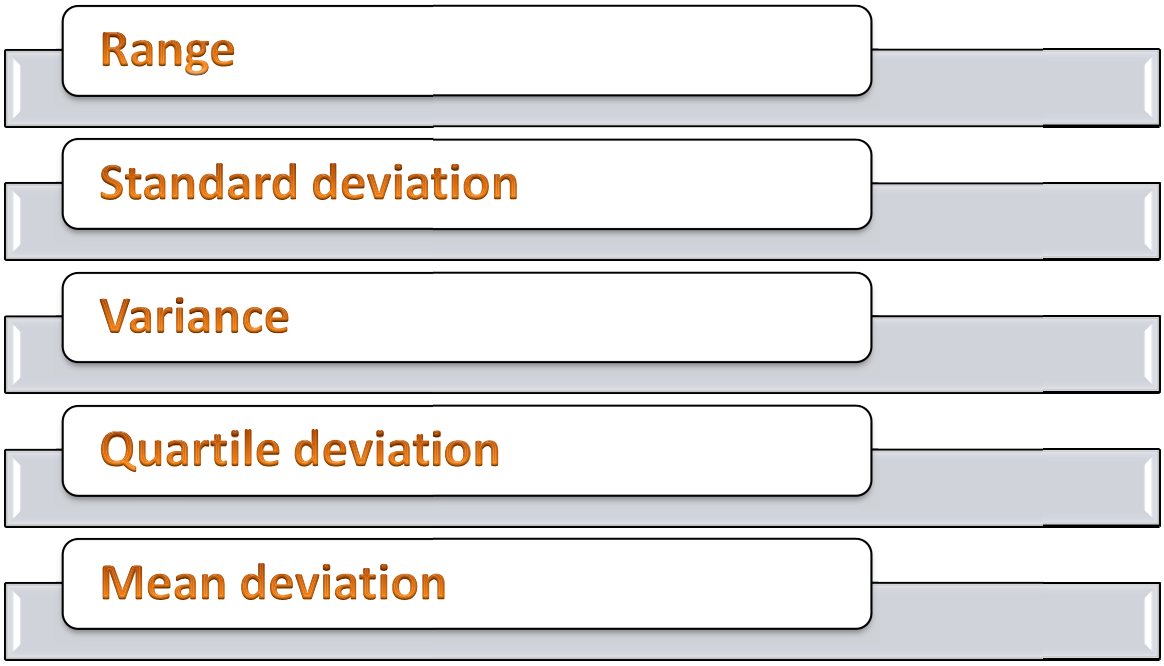
Relationship between Mean, Median and Mode:

- $\boldsymbol{Z = 3M - 2\bar{x}}$

Measures of dispersion

- Dispersion of the data is the degree to which the numerical data approached to spread about an average value.
- The variability in the data can be analyzed with the help of measure of dispersion.
- The measures of dispersion are studied to determine the reliability of an average to control the variability and to compare two or more distributions regarding their variability.
- Measure of dispersion is mainly useful in biological process of living organisms than the non-living things. E.g. RBC number, CO_2 utilization, etc.
- Dispersion is the measure of the variation of the items.

Types of Measures of dispersion



1) Range:

- Range is defined as the difference between the highest and lowest value in the sample.
- Mathematically, if H is the highest value and L is the lowest value then $Range(R) = H - L$

- Also the relative measure of range is known as the 'coefficient of range'.

$$coefficient\ of\ range = \frac{H - L}{H + L}$$

Example-17:

Calculate the range and the coefficient of range for the following data regarding Hb% of 10 patients for individual series.

8.3, 9.6, 12.3, 10.2, 11.3, 9.6, 13.2, 10.1 and 9.7

Solution:

We have the highest value = 13.2 and the lowest value = 8.3

Then,

$Range = H - L = 13.2 - 8.3 = 4.9$

$$\boxed{\boldsymbol{Range = 4.9}}$$

Also,

$$coefficient\ of\ range = \frac{H - L}{H + L} = \frac{13.2 - 8.3}{13.2 + 8.3} = \frac{4.9}{21.5} = 0.2279$$

$$\boxed{\boldsymbol{coefficient\ of\ range = 0.2279}}$$

Example-18:

Find the range and coefficient of range for discrete series from the following table,

X	5	15	25	35	45	55
f	7	14	18	10	4	1

Solution:

Here,

Highest value = 55 and the lowest value = 5

Then,

$Range = H - L = 55 - 5 = 50$

$\boxed{\boldsymbol{Range = 50}}$

Also,

$$coefficient\ of\ range = \frac{H - L}{H + L} = \frac{55 - 5}{55 + 5} = \frac{50}{60} = 0.83$$

$\boxed{\boldsymbol{coefficient\ of\ range = 0.83}}$

Example-19:

Find the range and coefficient of range for continuous series from the following table;

Data	0-10	10-20	20-30	30-40	40-50
Frequency	1	5	10	13	9

Solution:

Data	**Frequency**	**Mid value**
0-10	1	5
10-20	5	15
20-30	10	25
30-40	13	35
40-50	9	45

Here, we have $H = 45\ and\ L = 5$

$Range = H - L = 45 - 5 = 40$

$\boxed{\boldsymbol{Range = 40}}$

$$coefficient\ of\ range = \frac{H - L}{H + L} = \frac{45 - 5}{45 + 5} = \frac{40}{50} = 0.8$$

$\boxed{\boldsymbol{coefficient\ of\ range = 0.8}}$

Exercise

1) The following data shows variations in blood pressure of a patient in 7 hours. Find range and its coefficient.

Hour	1	2	3	4	5	6	7
BP	100	120	130	125	130	135	110

2) The following data shows blood sugar levels of students. Find range and its coefficient.

Blood sugar	80-90	90-100	100-110	110-120	120-130	130-140
No of students	8	12	13	17	30	90

Application of Range:

- **In quality control:** control charts are prepared for checking the qualities of a product. In this condition, range plays a very important role. In case, range increases beyond a specific point then the checking of the production machine is started.
- **In share market:** Range is useful in the study of variations in share and stocks.
- **In Weather Forecasting:** The difference between the maximum and minimum temperature is generally found by the meteorological department. This information is useful to the general public.

Advantages of Range:

- It is very easy to calculate and simple to understand.
- It is rigidly defined.
- It provides the limit within which all the data items occur.

Disadvantages of Range:

- This method is highly affected by the extreme items as it gives importance to the two extreme values.
- This method does not provide any information about the structure of the series.
- It is affected by the sampling fluctuation.
- Range is not a reliable method to measure the dispersion.

2) Standard deviation:

- Standard deviation is the square root of the arithmetic mean of the squared deviations of items taken from the arithmetic mean.
- It is used most commonly in statistical analysis.
- Standard deviation is widely used measure of variation; it is also called as root mean square deviation and denoted by Greek letter σ (sigma).
- Standard deviation is calculated on the basis of individual series, discrete series and continuous series by using actual mean and assumed mean method.
- It is given as

$$S.D.(\sigma) = \sqrt{\frac{\sum dx^2}{n}} \text{ , where } dx = x - \bar{x}$$

- In case of frequencies distribution, we have

$$S.D.(\sigma) = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}}$$

- In continuous series,

$$S.D.(\sigma) = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times c$$

Example-20:

Find out Standard deviation from the following data,

3, 7, 8, 9, 10

Solution:

Here, $n = 5$

$\boldsymbol{x}$	$\boldsymbol{dx = x - \bar{x}}$	$\boldsymbol{dx^2}$
3	-4.4	19.36
7	-0.4	0.16
8	0.6	0.36
9	1.6	2.56
10	2.6	6.76
$\boldsymbol{\sum x = 37}$		$\boldsymbol{\sum dx^2 = 29.2}$

$$Mean(\bar{x}) = \frac{\sum x}{n} = \frac{37}{5} = 7.4$$

$$S.D.(\sigma) = \sqrt{\frac{\sum dx^2}{n}} = \sqrt{\frac{29.2}{5}} = \sqrt{5.84} = 2.416$$

$$\boxed{\boldsymbol{\sigma = 2.416}}$$

Example-21:

In a survey of 150 families in a village, the following distribution of ages of children was found,

Ages of Children	0-2	2-4	4-6	6-8	8-10
No of families	40	32	25	23	30

Find the mean and standard deviation of the given distribution.

Solution:

Class Interval	Frequency (f)	Mid value (x)	$d = \frac{x - A}{c}$	d^2	$f \cdot d$	$f \cdot d^2$
0-2	40	1	-2	4	-80	160
2-4	32	3	-1	1	-32	32
4-6	25	5=A	0	0	0	0
6-8	23	7	1	1	23	23
8-10	30	9	2	4	60	120
	$\sum f = 150$				$\sum f \cdot d = -29$	$\sum f \cdot d^2 = 335$

Now,

$$Mean(\bar{x}) = A + \frac{\sum f \cdot d}{n} = 5 + \frac{(-29)}{150} = 5 - 0.193 = 4.807$$

$$S.D.(\sigma) = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times c = \sqrt{\frac{335}{150} - \left(\frac{-29}{150}\right)^2} \times 2$$

$$= \sqrt{2.23 - (0.193)^2} \times 2 = \sqrt{2.23 - 0.037} \times 2 = \sqrt{2.193} \times 2$$

$$= 1.48 \times 2 = 2.96$$

$$\boxed{\boldsymbol{\sigma = 2.96}}$$

Example-22:

Calculate the standard deviation of following data,

No of families	1	2	3	4	5
No of patients	3	5	2	4	2

Solution:

Family(x)	**No of patients(f)**	$f \cdot x$	$x - \bar{x}$	$(x - \bar{x})^2$	$f \cdot (x - \bar{x})^2$
1	3	3	-1.81	3.276	9.828
2	5	10	-0.81	0.656	3.28
3	2	6	0.19	0.036	0.072
4	4	16	1.19	1.416	5.664
5	2	10	2.19	4.796	9.592
	$n = 16$	$\sum f \cdot x = 45$			$\sum f \cdot (x - \bar{x})^2 = 28.436$

Calculation of mean,

$$Mean(\bar{x}) = \frac{\sum f \cdot x}{n} = \frac{45}{16} = 2.81$$

$$SD(\sigma) = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}} = \sqrt{\frac{28.436}{16}} = \sqrt{1.777} = 1.333$$

$$\boxed{\sigma = 1.333}$$

Exercise

1) The following data reveal the pharmaceutical units having number of ampoule filling machine in parental facility of total 25 manufacturing plants present in Goa state,

No of machine	2	3	4	5	6	7
No of Manufacturing Unit	2	4	5	6	4	4

Calculate mean and Standard deviation.

2) Calculate standard deviation from given data: 15, 16, 18, 17, 13, 14

3) Variance:

- The square of standard deviation is called the variance.
- It has a significant role in inferential statistics.
- It is denoted by σ^2 and defined as

$$Variance(\sigma^2) = \frac{\sum(x - \bar{x})^2}{n}$$

- For frequency distribution OR continuous series, variance defined as

$$Variance(\sigma^2) = \frac{\sum f \cdot (x - \bar{x})^2}{n}$$

Coefficient of Variance:

- Coefficient of variance (CV) is used to compare the variability of one character in two different groups having different magnitude of the values or two characters in the same group by expressing in percentage.
- It is calculated from standard deviation and mean of characteristic.
- The ratio of standard deviation and mean is found in percentage.

$$Coefficient\ of\ Variance = \frac{S.D.}{Mean} \times 100$$

OR

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

Example-23:

Find out Mean, Standard deviation, variance and Coefficient of variance from the following data,

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	3	4	2	3	5

Solution:

Class Interval	**Frequency** (f)	**Middle Value** (x)	$f \cdot x$	$x - \bar{x}$	$(x - \bar{x})^2$	$f \cdot (x - \bar{x})^2$
0-10	3	5	15	-21.76	473.50	1420.5
10-20	4	15	60	-11.76	138.30	553.2
20-30	2	25	50	-1.76	3.10	6.2
30-40	3	35	105	8.24	67.90	203.7
40-50	5	45	225	18.24	332.70	1663.5
	$n = 17$		$\sum f \cdot x = 455$			$\sum f \cdot (x - \bar{x})^2 = 3847.1$

$$Mean(\bar{x}) = \frac{\sum f \cdot x}{n} = \frac{455}{17} = 26.76$$

$$\boxed{\bar{x} = 26.76}$$

$$SD(\sigma) = \sqrt{\frac{\sum f(x - \bar{x})^2}{n}} = \sqrt{\frac{3847.1}{17}} = \sqrt{226.3} = 15.043$$

$$\boxed{\sigma = 15.043}$$

$$Variance(\sigma^2) = \frac{\sum f \cdot (x - \bar{x})^2}{n} = \frac{3847.1}{17} = 226.3$$

$$\boxed{\sigma^2 = 226.3}$$

$$CV = \frac{\sigma}{\bar{x}} \times 100 = \frac{15.043}{26.76} \times 100 = 0.5621 \times 100 = 56.21$$

$$\boxed{\boldsymbol{CV = 56.21}}$$

Exercise

1) Find standard deviation and Variance from the following figures,

Height	44-46	46-48	48-50	50-52	52-54
No of Children	5	25	28	22	5

2) Calculate standard deviation, Variance and Coefficient of variance from given data:

Marks	10	20	30	40	50	60	70	80	90
No of Students	2	6	9	12	15	29	11	9	7

4) Quartile deviation/ Inter-Quartile Range:

- The quartile deviation of a group of observations is the interval between the values of the upper quartile and the lower quartile for that group.
- Upper quartile of a group is the value above which 25% of the observations fall.
- Lower quartile is the value below which 25% of the observations fall.
- This measure gives us the range which covers the middle 50% of the observations in the group.
- If lower quartile is Q_1 and the upper quartile is Q_3 then,

  $Inter\ quartile\ range = Q_3 - Q_1$

  $R_Q = Difference\ between\ third\ and\ first\ quartile$

- Quartile deviation OR Semi-inter quartile range is defined as

$$Q = \frac{1}{2}[Q_3 - Q_1]$$

- Coefficient of Semi-inter quartile range is defined as

$$Coefficient\ of\ Semi-inter\ quartile\ range = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example-24:

Calculate the inter quartile range, Quartile deviation and coefficient of quartile deviation from the following table,

Height	58	59	60	61	62	63	64	65	66
No of Students (f)	21	25	28	18	20	22	24	23	18

Solution:

Height	**No of students (f)**	**cf**
58	21	21
59	25	46
60	28	74
61	18	92
62	20	112
63	22	134
64	24	158
65	23	181
66	18	199

We have,

$$Q_1 = Size\ of\ \left[\frac{n+1}{4}\right]^{th}\ item = Size\ of\ \left[\frac{199+1}{4}\right]^{th}\ item = Size\ of\ 50^{th}\ item$$

$$= 60$$

And

$$Q_3 = Size\ of\ \left[\frac{3(n+1)}{4}\right]^{th}\ item = Size\ of\ \left[\frac{3(199+1)}{4}\right]^{th}\ item$$

$$= Size\ of\ 150^{th}\ item = 64$$

$$Inter\ quartile\ range = Q_3 - Q_1 = 64 - 60 = 4$$

$$Semi - inter\ quartile\ range = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(4) = 2$$

$$Coefficient\ of\ quartile\ deviation = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{4}{64+60} = \frac{4}{124} = 0.032$$

Advantages of Quartile deviation:

- It is easy to understand and calculate.
- It is unaffected by the extreme values.
- It is quite satisfactory when only the middle half of the group is dealt with.

Disadvantages of Quartile deviation:

- It ignores 50% of the extreme values.
- It is not suitable for algebraic treatment.

5) Mean deviation:

- Mean deviation is defined as an average or mean of the deviations of the values from central tendency (mean, median or mode).
- Find out the mean deviation to follow the given steps,

  First define data as x.

  Calculate arithmetic mean as $\bar{x} = \frac{\sum x}{n}$

  Find the deviation of each observation from the mean, $dx = x - \bar{x}$

  Ignoring the negative sign of deviation and denoted by $\sum|dx|$

  Apply the formula, $MD = \frac{\sum|dx|}{n}, where\ n = total\ no\ of\ values$

- In case of frequency distribution,

$$MD = \frac{\sum f|dx|}{n}$$

$$Coefficient\ of\ Mean\ deviation = \frac{MD}{Mean}$$

Example-25:

Find the mean deviation from the following data,

10, 20, 26, 32, 23, 15

Solution:

Sr. No.	x	$dx = x - \bar{x}$	$\lvert dx \rvert$
1	10	-11	11
2	20	-1	1
3	26	5	5
4	32	11	11
5	23	2	2
6	15	-6	6
	$\sum x = 126$		$\sum \lvert dx \rvert = 36$

Here, $n = 6$

$$\bar{x} = \frac{\sum x}{n} = \frac{126}{6} = 21$$

$$Mean\ deviation = \frac{\sum |dx|}{n} = \frac{36}{6} = 6$$

$$\boxed{MD = 6}$$

Example-26:

Find the mean deviation and coefficient of mean deviation from the mean for the following data,

Data	0-10	10-20	20-30	30-40	40-50
Frequency	2	5	8	15	20

Solution:

Data	**Frequency (f)**	**Mid Value (x)**	$\|dx\| = \|x - \bar{x}\|$	$f\|dx\|$
0-10	2	5	2.5	5
10-20	5	15	12.5	62.5
20-30	8	25	22.5	180
30-40	15	35	32.5	487.5
40-50	20	45	42.5	850
	$n = 50$	$\sum x = 125$		$\sum f\|dx\| = 1585$

$$\bar{x} = \frac{\sum x}{n} = \frac{125}{50} = 2.5$$

$$MD = \frac{\sum f|dx|}{n} = \frac{1585}{50} = 31.7$$

$$\boxed{MD = 31.7}$$

$$Coefficient\ of\ MD = \frac{MD}{Mean} = \frac{31.7}{2.5} = 12.68$$

$$\boxed{Coefficient\ of\ MD = 12.68}$$

Correlation

- In correlation, we will define the relationship between two continuous variables.
- The main goal of correlation study is to understand the nature and strength of the linear association between the two quantitative parameters.
- The word correlation means the relation between two variables, in which change in the value of one variable changes the values of the other variable. Here, word relation has been used in the sense of mutual dependence.
- If two variables are so inter-related in such a manner that change in one variable brings about in the other variable, then this type of relation of variable known as correlation.
- If we change the value of one variable that will make corresponding change in the value of other variable on an average then we can say two variables are correlation.
- According to Croxton and Cowden, "The appropriate statistical tool for discovering and measuring the relationship of quantitative and expressing in the brief formula is known as correlation.
- The value of correlation coefficient will vary from -1 to +1.

Types of Correlation

- Correlation can be classified into three categories:
  1) Positive, Negative and Zero correlation
  2) Linear and Non-linear correlation
  3) Simple, Partial and Multiple correlation

Positive, Negative and Zero correlation

- If the values of two variables move in the same direction that means if the value of one variable is increase or decrease, then value of other variable also increases or decreases on an average, then the correlation said to be Positive correlation.

  For example,

  1) Height and Weight (as height increases weight also increases)
  2) Example of positive correlation,

X	1	2	3	4	5	6
Y	10	20	30	40	50	60

- If the value of one variable is increases or decreases, then value of other variable decreases or increases on an average or in a simple manner, if the value of both variable moves in opposite direction, then it is said to be Negative correlation.

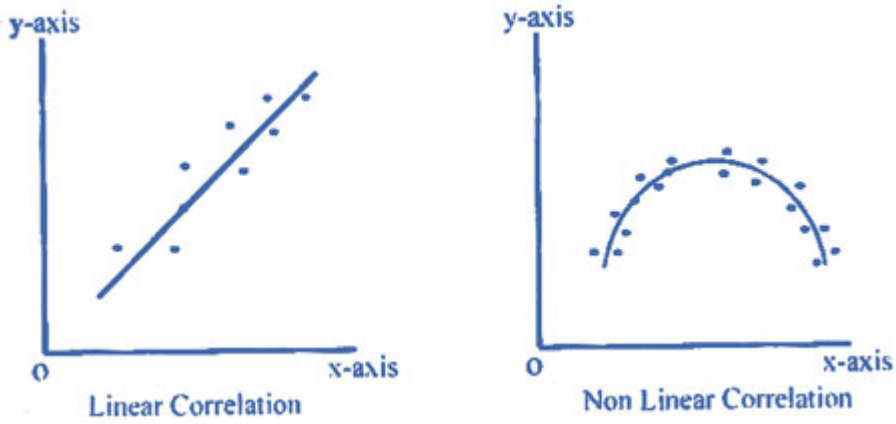
  For example,

X	1	2	3	4	5	6
Y	70	60	50	40	30	20

- If the change in the value of one variable will not affect the value of other variable then the correlation is said to be Zero correlation.

Linear and Non-linear correlation

- If the change in values of one variable makes a constant ratio with the change in value of other variable, then such type of relation known as Linear correlation.
- The correlation is said to be a Non-linear correlation, if the value in one variable does not make a constant ratio with change in the value of other variable.



Simple, Partial and Multiple correlations

- If we study the relationship between two variables X and Y then it is called Simple correlation. For example, Height and Weight
- If we study the relationship between two variables, keeping the other entire variable as constant, then it is called as Partial correlation.
- If we study the relationship between more than two variables then it is said to be multiple correlation. In multiple correlation we measure the degree of relationship between one variable on one side and combined effect of all other variable on the other side.

Note:

- Since the value of correlation lies between -1 and +1 $(-1 \leq r \leq +1)$, then
  - (i) If $r > 0$, we say that a positive correlation between variables.
  - (ii) If $r < 0$, we say that a negative correlation between variables.
  - (iii) If $r = 0$, we say that no correlation between variables.

Karl Pearson's coefficient of correlation

- Karl Pearson's coefficient of correlation is used to measure the degree of linear relationship between two variables.
- It is also called moment correlation coefficient.
- It is denoted by 'r' and defined as

$$r = \frac{\sum XY}{n \cdot \sigma_x \sigma_y} \text{ , where } X = x - \bar{x} \text{ and } Y = y - \bar{y}$$

$$n = No\ of\ pair\ of\ values\ of\ variable\ \&\ \sigma = Standard\ deviation$$

- Another form of correlation coefficient is as

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

Example-27:

Following data gives the height of father and son in inches. Find the Karl Pearson's coefficient of correlation,

Height of father (x)	65	66	67	67	69	71
Height of son (y)	67	68	64	68	70	69

Solution:

We know that,

$$\bar{x} = \frac{\sum x}{n} = \frac{405}{6} = 67.5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{406}{6} = 67.7$$

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	XY	X^2	Y^2
65	67	-2.5	-0.7	1.75	6.25	0.49
66	68	-1.5	0.3	-0.45	2.25	0.09
67	64	-0.5	-3.7	1.85	0.25	13.69
67	68	-0.5	0.3	-0.15	0.25	0.09
69	70	1.5	2.3	3.45	2.25	5.29
71	69	3.5	1.3	4.55	12.25	1.69
$\sum x = 405$	$\sum y = 406$			$\sum XY = 11$	$\sum X^2 = 23.5$	$\sum Y^2 = 21.34$

Karl Pearson's Coefficient of correlation

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} = \frac{11}{\sqrt{23.5 \times 21.34}} = \frac{11}{\sqrt{501.49}} = \frac{11}{22.39} = 0.49$$

$\boxed{r = 0.49}$

Exercise

1) Find out the Karl Pearson's Coefficient of correlation from the following data,

Age (x)	1	2	3	4	5	6
Height (y)	7	11	14	19	24	29

2) Find out the Karl Pearson's Coefficient of correlation from the following data,

x	65	40	35	75	63	80	35	20	85	65	55	33
y	30	55	68	28	76	25	80	85	20	35	45	65

3) Calculate Karl Pearson's Coefficient of correlation from the following data,

x	18	20	21	22	27	27	28	29	29	29
y	23	37	29	28	28	31	35	30	36	33

Advantages of Karl Pearson's Coefficient of correlation:

- It is important method to give a precise and quantitative result with a meaningful interpretation.
- It also gives a direction (positive or negative) as well as the degree of the correlation between the variables.

Disadvantages of Karl Pearson's Coefficient of correlation:

- This method is a time consuming.
- The limitation of value of correlation is -1 to +1.

Multiple Correlation

- The multiple correlation is a measure of linear relationship between a dependent variable and a number of independent variables.
- It is represented by R and can have any value between 0 and 1.
- For example, if one wants to measure the relationship between the yield of rice on one hand the amount of rainfall and temperature taken together, on the other hand. In this case, yield of rice is a dependent variable and rainfall and temperature are independent variables.
- Hence, multiple correlation helps us to measure the correlation between a dependent variable and a group of independent variables.
- In multiple correlation, we study the relationship between three or more variable. Suppose the dependent variable is z and x & y both are independent variables. Then the multiple correlation coefficient is defined as

$$R_{z,xy} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xz}r_{yz}r_{xy}}{1 - r_{xy}^2}}$$

- Similarly,

$$R_{y,zx} = \sqrt{\frac{r_{zy}^2 + r_{xy}^2 - 2r_{xy}r_{zy}r_{xz}}{1 - r_{xz}^2}}$$

$$R_{x,yz} = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2r_{xz}r_{xy}r_{yz}}{1 - r_{yz}^2}}$$

- A coefficient of multiple correlation lies between 0 and 1. If the value of multiple correlation is 1, then the correlation of variables is perfect, while, iff the value of multiple correlation is 0, then there is no correlation of variable.

Example-28:

The following correlation coefficients are given for a tri-variable data:

$r_{12} = 0.96, r_{13} = 0.49 \; and \; r_{23} = 0.46$

Calculate the multiple correlation coefficients treating the first variable as dependent and the second and third variables as independent.

Solution:

$$R_{1,23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{13}r_{12}r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.96)^2 + (0.49)^2 - 2(0.49)(0.96)(0.46)}{1 - (0.46)^2}}$$

$$= \sqrt{\frac{0.9216 + 0.2401 - 0.4328}{1 - 0.2116}} = \sqrt{\frac{0.7289}{0.7884}} = \sqrt{0.9245} = 0.96$$

$$\boxed{R_{1,23} = 0.96}$$

ALL THE BEST