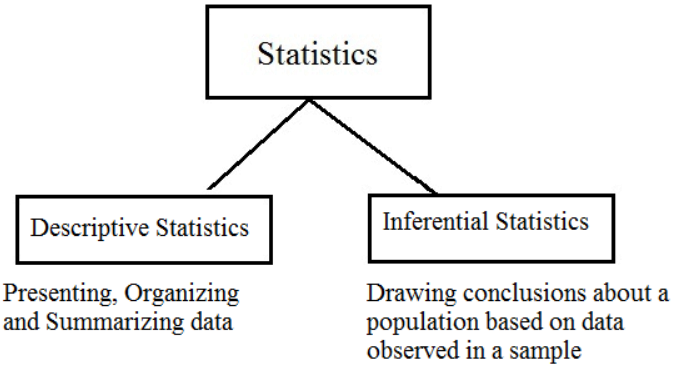
STATISTICS:

It is a field of study concerned with summarizing data interpreting data and making decisions based on data (or) a quantity calculated in a sample to estimate a value in a population is called a statistic.

The related term data science or data analysis stands for a study of processes and systems that extract knowledge or insights from data in various forms either structured or unstructured.

Data science is a continuation of some of the fields such as a statistics, data mining and predictive analytics.

STRUCTURE OF STATISTICS:



VARIABLES:

Variables are properties or characteristics of some event, object or person that can take on different values or amounts.

Constants do not vary.

Variables may be

* Independent or dependent
* Discrete or continuous
* Qualitative or quantitative.

INDEPENDENT OR DEPENDENT VARIABLES:

When conducting research, experiments often manipulate variables. For example, an experiment might compare the effectiveness of four types of antidepressants. In this case, the variable is type of antidepressant.

* When a variable is manipulated by an experiment, it is called an "independent variable".
* The experiment seeks to determine the effect of independent variable (or) relief from depression. This is called a dependent variable.

DESCRIPTIVE STATISTICS:

DATA SCIENCE:

INTRODUCTION:

Data science is the practice of using statistical techniques, regression models, machine learning and deep learning algorithms to produce advanced insights and build predictive

applications. All software applications are intended to increase productivity and efficiency by automating human activity. Traditionally, these tasks needed to be repetitive in nature and based on a deterministic set of rules. An example would be an accounting system that can take sales and expenses and automatically create a balance sheet. The intent of data science applications is to automate tasks that require human judgments and are not driven by deterministic rules. An example would be a predictive application that automatically determines if a customer feedback statement is positive or negative, or if an email is a spam email or not.

Sometimes, data science is also to provide insights that may not be otherwise available. An example would be an application that Toyota may use to predict whether an existing customer is ready to purchase a new car. This is something that an experienced sales person may also not be able to determine with high accuracy.

All applications work by creating and operating a digital model of the real life scenario that they are trying to automate. Data science applications use historical data to "learn" the chief characteristics of the real – life scenario and then create the digital model on that basis. Typically, this model is mathematical in nature, but it can take other forms.

Since data is being used to build non – deterministic models, such techniques are grouped under the umbrella term "Data science". Other terms that are applied are "Machine learning" and "Artificial Intelligence". However, Artificial Intelligence can have broader connotations and is sometimes treated as a distant category from Data science. But the popular AI techniques have many things in common with Machine Learning techniques, making these disciplines a continuum than having sharp boundaries.

DATA SCIENCE TECHNIQUES:

There are a variety of techniques that come in the scope of data science.

* Regression techniques are used to build a mathematical equation that uses input variables to predict the value of an output variable. For example, factors such location, number of rooms, age of the house can be used to predict the price of a house by using regression techniques.
* Machine learning methods create computational models with input data and use the models to determine the "class" of a variable. For example, an ML model can be used to determine if an email is "spam" or "regular". Here, "spam" and "regular" are called classes and such algorithms are called classification algorithms. Machine learning methods rely on iterative algorithms that refine the computational model and can increase in sophistication based on the computational power available. Some well – known algorithms are Decision Trees, Random Forest, XG Boost.
* There are many other techniques that can be used, depending on the available data and the use case. Clustering algorithms, Naïve Bayes, Support Vector Machines are some examples.
* Neural networks are considered AI algorithms and require large amounts of data and computational power. But they are very powerful.
* Sometimes, fitting a statistical distribution may result in a good model for the scenario.

DATA SCIENCE METHODOLOGY:

Data science applications require a particular methodology and skills.

EXPLORATORY DATA ANALYSIS (EDA):

* These applications work by identifying underlying patterns in the data and codifying that into

a computational model that represent the real life scenario (or the domain). So it is critical for a data science practitioner to understand the domain and how the data represents the domain. He needs to understand the basic properties of the data (mean, variance, range, distribution etc.) correlation between the variable that he is trying to predict and the input variables, as well as the relationship between the input variables themselves. The practitioner needs to know how to handle extreme values in the data (called outliers), how to assess the quality of the data and what to do with missing values. Often, pattern detection may require slicing or aggregation of data by various dimensions. So the data science practitioner needs to be aware of techniques to programmatically explore the data and draw insights that are meaningful for modeling .

DATA VISUALIZATION:

* Usually, the practitioner needs to deal with large volumes of data that is analyzable across several dimensions. Data visualization techniques are vital to be able to abstract data and detect patterns. The practitioner needs to be familiar with various kinds of plots and when they are to be used, and the ability to generate such plots programmatically from the data.

DATA MANIPULATION:

* Data science often needs to merge multiple datasets to create a common dataset. The student needs to be able to summarize data at various levels and in general be very well versed with data merging, splitting and computing relevant measures.

FEATURE SELECTION AND EXTRACTION:

* A data science application aims to create a digital model by choosing the key characteristics of the real–life scenario. Since real – life objects have many characteristics, the data science practitioner needs to select those characteristics that are relevant to the prediction. For example, the color of the house door is a characteristic of the house but may not be relevant to the price of the house, whereas the square footage of the garden may be a factor that influences price. Such characteristics are called features of the machine learning model. The data science practitioner should use the insights gained from the EDA and visualization exercise to determine the correct features that have predictive power.

MODEL DEVELOPMENT:

* The data science student needs to be able to determine the correct ML algorithm that can be used for a scenario. Sometimes, an empirical approach is needed to determine the best model for prediction. The student needs to be familiar with common algorithms and the merits and demerits of those algorithms.

MODEL PERFORMANCE MANAGEMENT:

* Once a model is built, its predictions need to be checked for accuracy. The practitioner needs to be aware of the various metrics of prediction and use the appropriate metric for the situation. The student also needs to be able to boost performance by using additional features or other techniques.

  Many of these techniques are available as software libraries in languages such as R programming and Python. The practitioner should be conversant with these languages and the libraries.

CONCLUSION:

Data science is a powerful discipline that can deliver great value to enterprises. It can be applied to a variety of domains and there are specialized domain specific techniques

available. But data science problems are open – ended and require experimentation and an active spirit of enquiry. Thus, a practitioner can benefit from a knowledge of these techniques but also should exhibit thorough analytical skills, a comfort with data manipulation and should also be creative in crafting the correct model for the situation.

STATISTICS:

DEFINITION:

Statistics is a tool in the hands of mankind to translate complex facts into simple and understandable statement of facts.

The word statistics is derived from the Italian word stato and it means a political state. In the singular sense statistics is as defined as a science which deals with scientific methods of collection, organization, summarization, presentation, analysis and interpretation of numerical data. Statistical methods are applied for investigation in every important science.

STATISTICAL METHODS:

1. COLLECTION OF DATA:

The first step of an investigation is the collection of data. Careful collection is needed because further analysis is based on this.

2. ORGANISATION OF DATA:

The large mass of figures that are collected from a survey needs organization.

3. PRESENTATION OF DATA:

The collected data must be edited very carefully so that irrelevant answers and wrong computations must be corrected or adjusted.

The collected data must be classified and tabulated before they can be analyzed.

4. ANALYSIS OF DATA:

After presentation of the data the next step is to analyse the presented data. Analysis included condensation, summarization, conclusion etc., through means of Measures of Central Tendencies, Dispersion, Skewness, Kurtosis, Correlation and Regression etc.

5. INTERPRETATION OF DATA:

Valid conclusions must be drawn on the basis of analysis. Correct interpretation leads to valid conclusion.

The statistical generalization provides the estimates of the characteristic behavior of population, but not of individual person.

The real purpose of statistical methods is to make sense out of facts and figures, prove the unknown and to cast light upon the situation. The statistical methods are employed as a tool for comparison between past and present results with a view to find out the reasons for changes. Statistics has become so much indispensible in all phases of human endeavor.

COLLECTION OF DATA:

The basic problem of statistical enquiry is to collect facts and figures relating to a particular phenomenon under study. The investigator is a person who conducts the statistical enquiry. He is a trained and efficient statistician. The statistician counts are measures the characteristic under study for further statistical analysis. The respondents or informants are

the persons from whom the information is collected. The statistical units are the items on which the measurement is taken. Collection of data is the process of enumeration together with the proper recording of results. The success of an enquiry depends on the proper collection of data.

PRIMARY AND SECONDARY DATA:

Statistical data may be classified as primary and secondary.

PRIMARY DATA:

If an individual or an officer collects the data to study a particular problem, the data are the raw materials of the enquiry. They are the primary data collected by the investigator himself to study any particular problem.

SECONDARY DATA:

Secondary data are those which are already collected by someone for some purpose and are available for the present study. For example, the data collected during Census operations are primary data to the department of census and the same data, if used by a research worker for some study are secondary data.

SOURCES OF SECONDARY DATA:

1. PUBLISHED SOURCES:

Such as international publications, official publications of central and state governments, semi – official publications of semi – government institutions like municipal corporations, panchayats, etc., publications of research institutions, publications of commercial and financial institutions, reports of various committees, Journals and news papers.

2. UNPUBLISHED SOURCES:

They are records maintained by various government and private offices, the research carried out by individual research scholars in the universities or research institutes.

PRECAUTIONS IN THE USE OF SECONDARY DATA:

Before using the secondary data, we must take into consideration.

a) THE SUITABILITY OF DATA:

The investigator must satisfy himself that the data available is suitable for the purpose of enquiry.

b) ADEQUENCY OF DATA:

If the data are suitable for the purpose of investigation, then we must consider whether the data is useful or adequate for the present analysis.

c) RELIABILITY OF DATA:

The reliability of data can be tested by finding whether the collecting agents used proper methods or not. If the methods are proper, the data can be relied on.

Without knowing the meanings and limitations, we cannot accept the secondary data.

POPULATION VS SAMPLE:

In statistical enquiry, all the items, which fall within the preview of enquiry, are known as **Universe of Population**. That is population is a complete set of all possible observations of

the type which is to be investigated. This is a statistical usage and the term population does not necessarily refer to the people.

1. FINITE AND INFINITE POPULATION:

Population can be either finite population or infinite population. When the number of observations can be counted and definite, it is known as "finite population". For example, when we are studying the economic background of students of a college, all the students of the college will constitute population and this number will be finite. When the number of observations cannot be counted and is infinite, it is known as infinite population. For example, the number of stars in the sky is infinite population.

2. HYPOTHETICAL AND EXISTANT POPULATION:

Universe can be classified as extent and hypothetical. A Universe containing persons or objects is known as existent or real population. The examples are the students of a college, population of a city, the employees of a factory.

Hypothetical Universe which is also known as theoretical population, is one which does not consist of concrete objects. This population exists only in imagination. For example, if we toss a coin infinite times, the result is a hypothetical population. Information on population can be collected in two ways. Census method and sample method.

A) CENSUS METHOD:

The object of census or complete enumeration is to collect information for each and every unit of the population.

In census method, every element of the population is included in the investigation. When we make a complete enumeration of all items in the population, it is known as census method of collection of data. For example, if we study the average expenditure of the students of University, which has 20,000 students, we must study the expenditure of all the 20,000 students. In the census method complete enumeration is done.

This method requires a large number of enumeration and is a costly method. Then only government alone can use this method for conducting population census, production census etc.

B) SAMPLE METHOD:

In the case of sample enquiry, only a part of the whole group of population will be studied. We can study the characteristics of a population from sampling. A study of the sample will give a correct idea of the Universe or population.

MERITS:

1. It saves time, because fewer items are collected and processed.
2. When the results are urgently required, this method is very helpful.
3. It reduces the cost of the enumeration.
4. More reliable results can be obtained, since there are fewer chances of sampling errors.
5. Expert and trained people can be employed for scientific processing and analysis.

METHODS OF SAMPLING:

1. RANDOM SAMPLING METHOD (PROBABILITY SAMPLING):

A random sample is one where each item in the universe has an equal chance of known opportunity.

2. NON – RANDOM SAMPLING:

This can be done in three methods.

A) JUDGEMENT OR PURPOSIVE SAMPLING:

The choice of the sample items depends on the judgement of the investigation.

B) QUOTA SAMPLING:

To collect data, the universe is divided into quota according to some characteristics.

C) CONVENIENCE SAMPLING OR CHECK SAMPLING:

The sampling is obtained by selecting convenient population units.

1. It is suitable when the population is not clearly defined.
2. Sample is not clear.
3. Complete source list is not available.

A sample obtained from automobile registration, telephone directories etc. is a convenient sample. They are unsatisfactory. They are biased. But they are used for pilot studies.

TYPES OF VARIABLES:

All the experiments examine some kind of variable(s). A variable is not only something that we measure, but also something that we can manipulate and something we can control for. First, we illustrate the role of dependent and independent variables. Finally, we explain how variables can be characterized as either categorical or continuous.

DEPENDENT AND INDEPENDENT VARIABLES:

An independent variable, sometimes called an experimental or predictor variable, is a variable that is being manipulated in an experiment in order to observe the effect on a dependent variable, sometimes called an outcome variable.

Imagine that a tutor asks 100 students to complete a maths test. The tutor wants to know why some students perform better than others. Whilst the tutor does not know the answer to this, she thinks that it might be because of two reasons:

1) Some students spend more time revising for their test and

2) some students are naturally more intelligent than others. As such, the tutor decides to investigate the effect of revision time and intelligence on the test performance of the 100 students. The dependent and independent variables for the study are:

DEPENDENT VARIABLE: Test Mark (measured from 0 to 100)

INDEPENDENT VARIABLE: Revision time (measured in hours) Intelligence (measured using IQ score)

The dependent variable is simply that, a variable that is dependent on an independent variable(s). For example, in our case the test mark that a student achieves is dependent on revision time and intelligence. Whilst revision time and intelligence (the independent variables) may (or may not) cause a change in the test mark (the dependent variable), the reverse is implausible; in other words, whilst the number of hours a student spends revising and the higher a student's IQ score may (or may not) change the test mark that a student achieves, a change in a student's test mark has no bearing on whether a student revises more or is more intelligent

CATEGORICAL AND CONTINUOUS VARIABLES:

Categorical variables are also known as discrete or qualitative.

* Nominal variables are variables that have two or more categories, but which do not have an intrinsic order. For example, a real estate agent could classify their types of property into distinct categories such as houses, condos, co-ops or bungalows. So "type of property" is a nominal variable with 4 categories called houses, condos, co-ops and bungalows. Of note, the different categories of a nominal variable can also be referred to as groups or levels of the nominal variable. Another example of a nominal variable would be classifying where people live in the USA by state. In this case there will be many more levels of the nominal variable (50 in fact).

* Dichotomous variables are nominal variables which have only two categories or levels. For example, if we were looking at gender, we would most probably categorize somebody as either "male" or "female". This is an example of a dichotomous variable (and also a nominal variable). Another example might be if we asked a person if they owned a mobile phone. Here we may categorise mobile phone ownership as either "Yes" or "No". In the real estate agent example, if type of property had been classified as either residential or commercial then "type of property" would be a dichotomous variable.

* Ordinal variables are variables that have two or more categories just like nominal variables only the categories can also be ordered or ranked. So if you asked someone if they liked the policies of the Democratic Party and they could answer either "Not very much". "They are OK" or "Yes, a lot" then you have an ordinal variable. Why? Because you have 3 categories, namely "Not very much", "They are OK" and "Yes, a lot" and you can rank them from the most positive (Yes, a lot), to the middle response (They are OK), to the least positive (Not very much). However, whilst we can rank the levels, we cannot place a "value" to them; we cannot say that "They are OK" is twice as positive as "Not very much".

  Continuous variables are also known as quantitative variables. Continuous variables can be further categorized as either interval or ratio variables.

* Interval variables are variables for which their central characteristic is that they can be measured along a continuum and they have a numerical value (for example, temperature measured in degrees Celsius or Fahrenheit). So the difference between 20^{o}C and 30^{o}C is the same as 30^{o}C to 40^{o}C. However, temperature measured in degrees Celsius or Fahrenheit is NOT a ratio variable.

* Ratio variables are interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. So, temperature measured in degrees Celsius or Fahrenheit is not a ratio variable because 0^{o}C does not mean there is no temperature. However, temperature measured in Kelvin is a ratio variable as 0 Kelvin (often called absolute zero) indicates that there is no temperature whatsoever. Other examples of ratio variables include height, mass, distance and many more. The name "ratio" reflects the fact that you can use the ratio of measurements. So, for example, a distance of ten meters is twice the distance of 5 meters.

DATA VISUALISATION:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Data visualization is the act of taking information (data) and placing it into a visual context, such as a map or graph. Data visualizations make big and small data easier for the human brain to understand, and visualization also makes it easier to detect patterns, trends and

outliers in groups of data.

5 types of Big data visualization categories:

1. Bar Chart
2. Line Chart
3. Scatter Plot
4. Sparkline
5. Pie Chart.

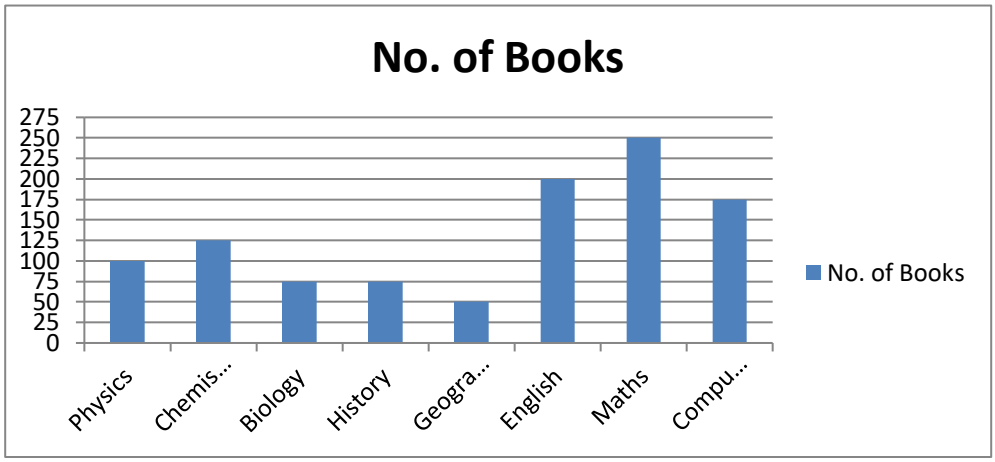
1. BAR GRAPH / CHART:

A bar graph is a pictorial representation of the numerical data by a number of bars of uniform width with different heights, erected horizontally or vertically with equal spacing between them.

Ex: The following table shows the number of books of different subjects in a library.

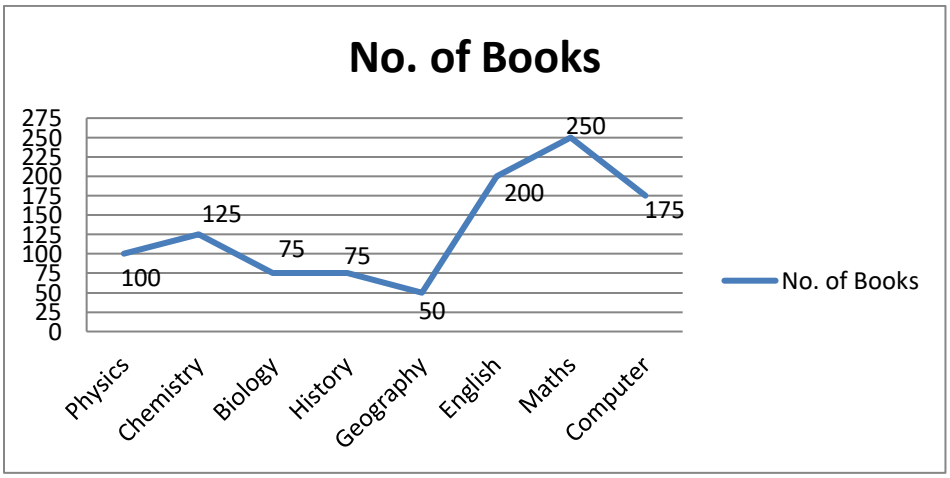
Subject	**Phy.**	**Chem.**	**Bio.**	**Hist.**	**Geography**	**Eng.**	**Maths**	**Comp.**
No. of Books	100	125	75	75	50	200	250	175

Sol: Take the subjects along the X – axis and number of books along the Y – axis. Construct the bars of same width, with same distance between them. Take the scale as 25 books = 5 small divisions or ½ cm.



2. LINE GRAPH / CHART:

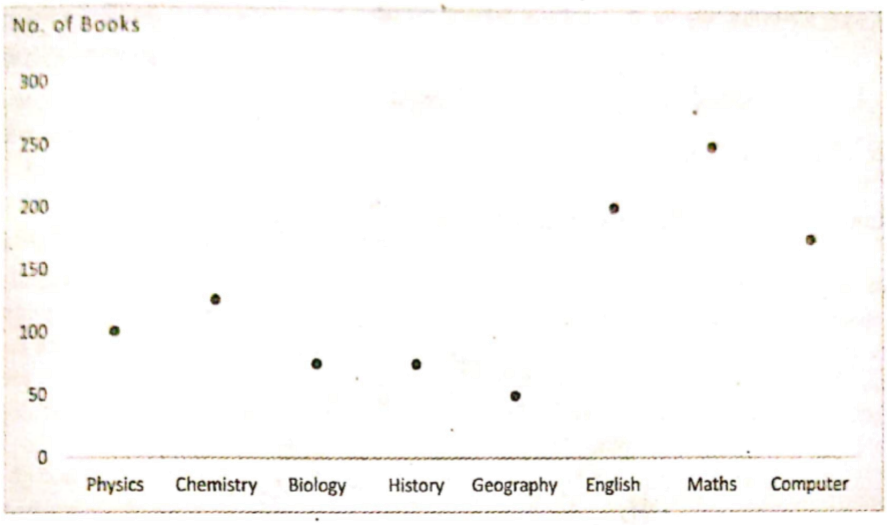
A line chart is a type of chart which displays information as a series of data points called markers connected by straight line segment.



3. SCATTER GRAPH / CHART:

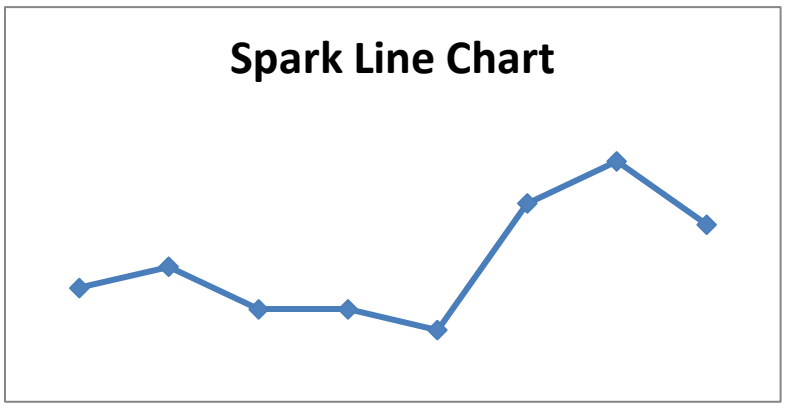
A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for two variables for a set of data. With scatter plots we often talk about how the variables relate to each other. This is called correlation.

A scatter plot is also called scatter graph, scatter chart or scatter diagram.



4. SPARKLINE:

A spark line is a very small line chart drawn without axes or coordinates. It present the general shape of the variation in some measurement, such as temperature or stock market price in a simple and highly condensed way.



5. PIE CHART / GRAPH:

A pie chart is a way of showing how something is shared or divided. This pie chart shows how 36 pupils usually come to school.

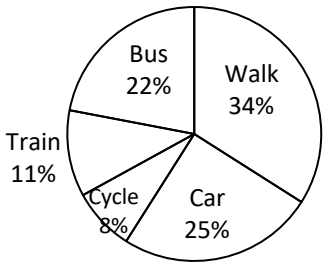
The number of pupils is 36 and the whole group of 36 pupils is represented by the complete angle 360^o. The angles at the centre are in proportion to each category. Thus, the angle of 120^o at the centre corresponds to the walk group.

Since angle of 360^o at the centre corresponds to the whole group of 36 pupils.

$\therefore 1^o$ corresponds to $\frac{36}{360}$ of the whole group.

$\therefore 120^o$ corresponds to $\frac{36}{360} \times 120 \; i.e., 12 \; pupils$. So pupil walk to school.

Similarly, number of pupils coming by car $= \frac{36}{360} \times 90 = 9$



MEASURES OF CENTRAL TENDENCY:

The objective of statistical analysis is to arrive at one single value which represents the whole series. This value is called central value or average.

While studying a population, we may get large number of observations, it is not possible to group any idea about the characteristics when we look at all the observation. So it is better to get one number for one group. That number must be represented by all the observation, such number is called the central value or the average.

DEFINITION:

According to Y-Lun-Chou "An average is a typical value in the since that it is sometimes employed to represent all the individual values in a series".

DEFINITION:

According to croxton and cowden, an average value is a single value within the range of the data that is used to represent all the value in the series. Since an average in somewhere within the range of the data, it is something called a measure of critical value.

MEASURES OF CENTRAL TENDENCY:

The following are the measures of central tendency.

1) Arithmetic Mean

2) Geometric Mean

3) Harmonic Mean

4) Median and

5) Mode

In this the first three are mathematical averages and the other two are positional averages and among this AM, Median and Mode are simple averages and GM and HM are the special averages.

CHARACTERISTICS FOR A GOOD (OR) AN IDEAL AVERAGE:

* It should be rigidly defined.

* It should be easy to calculate and easy to understand.

* It should be based on all the observation in the data.
* It is not effected by variations of sampling.
* It should be capable of being further mathematical treatment.
* It is not affected by extreme values.

	Mean		**Individual series** **X: $X_1, X_2, X_3, \ldots, X_n$**	**Discrete series** **X:** **f:**	**Continuous series** **C.I.** **f:**
1	**Arithmetic Mean**	**Direct Method**	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$\overline{X} = \frac{\sum_{i=1}^{n} f_n X_n}{N}$, $where\ N = \sum f$	$\overline{X} = \frac{\sum_{i=1}^{n} f_n X_n}{N}$, $X\ is\ mid\ of\ C.I.$ $N = \sum f$
		Shortcut Method	$\overline{X} = A + \frac{\sum d}{n}$ $d = X - A$ $A = Assumed\ mean,$ $d = deviation$	$\overline{X} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N}$ $d = X - A$ $A = Assumed\ mean,$ $d = deviation$	$\overline{X} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} \times c$ $d = \frac{X - A}{c}$ $X\ is\ mid\ value\ of\ C.I.$ $A = Assumed\ mean,$ $c = length\ of\ C.I$
2	**Weighted Arithmetic Mean**			$W.M = \frac{\sum_{i=1}^{n} X_n W_n}{N}$ $here\ N = \sum W$	$W.A.M = \frac{\sum_{i=1}^{n} X_n W_n}{N}$ $X\ is\ mid\ of\ C.I.$ $N = \sum W$
3	**Geometric Mean** **(Antilog of = 10 to the power of Ex:Antilog of 1.2 = $10^{1.2}$)**		$G.M. = Antilog\ of \left(\frac{\sum_{i=1}^{n} \log X_i}{n}\right)$	$G.M. = Antilog\ of \left(\frac{\sum_{i=1}^{n} f_n \log X_n}{N}\right)$ $here\ N = \sum f$	$G.M. = Antilog\ of \left(\frac{\sum_{i=1}^{n} f_n \log X_n}{N}\right)$ $X\ is\ mid\ of\ C,I.$ $N = \sum f$
4	**Harmonic Mean**		$H.M = \frac{n}{\sum \left(\frac{1}{x_1}\right)}$	$H.M = \frac{N}{\sum \left(\frac{f}{x}\right)}$	$H.M = \frac{N}{\sum \left(\frac{f}{x}\right)}$ $X = mid\ value\ of\ C.I.$

ARITHMETIC MEAN (OR) MEAN – A.M. :

Arithmetic Mean of the variable is defined as the sum of observation divided by the number of observations.

* **For Individual Series:**

If X_1, X_2, X_3 …….. X_n are 'n' observations then the mean is denoted by $\overline{X}$.

$\overline{X} = X_1 + X_2 + X_3 + \cdots \ldots . + X_n = \frac{\sum_{i=1}^{n} x_i}{n}$

Ex: The mean of 8, 12, 16, 20, 22, 18

Mean $\overline{X} = \frac{\sum X}{n} = \frac{96}{6} = 12.6667$

* **For Discrete Series:**

If X_1, X_2, X_3,……X_n are n variables with frequencies are f_1, f_2, f_3,……f_n respectively then the A.M. is

$\overline{X} = \frac{f_1X_1+f_2X_2+f_3X_3+\cdots..+f_nX_n}{f_1+f_2+f_3+\cdots....+f_n} = \frac{\sum_{i=1}^{n} f_iX_i}{N}$. This is direct method.

Short cut formula:

$\overline{X} = A + \frac{\sum fd}{N}, where\ d = X - A; A = assumed\ mean; d = deviation$

Ex: Calculate Arithmetic Mean from the following data:

Wages (Rs):	10	15	20	25	30	35	40
No. of Persons	5	2	3	10	3	2	5

Sol: **Direct Method:**

Wages (Rs):	10	15	20	25	30	35	40	
No. of Persons	5	2	3	10	3	2	5	N=30
Fx	50	30	60	250	90	70	200	$\sum Fx = 750$

$\overline{X} = \frac{\sum fx}{N} = \frac{750}{30} = 25$

Short cut Method:

X	**f**	**A = 20; d = X – A**	**f(d)**
10	5	-10	-50
15	2	-5	-10
20	3	0	0
25	10	5	50
30	3	10	30
35	2	15	30
40	5	20	100
	30		**150**

$\overline{X} = A + \frac{\sum fd}{N} = 20 + \frac{150}{30} = 20 + 5 = 25$

For Continuous Series:

For the grouped data, the Arithmetic Mean

$\overline{X} = \frac{\sum fx}{N}\ for\ direct\ method$

$\overline{X} = A + \frac{\sum fd}{N} \times C\ for\ step\ deviation\ method$

$d = \frac{X-A}{C}, A\ is\ Assumed\ Mean$

Direct Method:

Calculate mean for the following data:

C.I	F	Mid Value (x)	Fx
0 – 10	3	5	15
10 – 20	5	15	75
20 – 30	7	25	175
30 – 40	10	35	350
40 – 50	13	45	585
50 – 60	16	55	880
60 – 70	12	65	780
70 – 80	8	75	60
80 – 90	6	85	510
	80		**3970**

$$\overline{X} = \frac{\sum fx}{N} = \frac{3970}{80} = 49.625$$

For Step Deviation Method:

C.I	F	Mid Value	$d = \frac{x - A}{c}$	Fd
0 – 10	3	5	-4	-12
10 – 20	5	15	-3	-15
20 – 30	7	25	-2	-14
30 – 40	10	35	-1	-10
40 – 50	13	45	0	0
50 – 60	16	55	1	16
60 – 70	12	65	2	24
70 – 80	8	75	3	24
80 – 90	6	85	4	24
	80			**37**

$$\overline{X} = A + \frac{\sum fd}{N} \times C$$

$$d = \frac{X-A}{C} = 45 + \frac{37}{80} \times 10 = 45 + 46.25 = 49.625$$

$$A = Assumed\ Mean; C = Class\ Interval$$

Find the Arithmetic Mean from the following:

Marks below:	80	70	60	50	40	30	20	10
No. of Students	240	190	125	95	75	60	40	25

Marks below	Class	M.V.	C.F.	F	$d = \frac{x - A}{c}$	Fd
80	70 – 80	75	240	50	+3	150
70	60 – 70	65	19	65	+2	13
60	50 – 60	55	0125	30	+1	30
50	40 – 50	45	95	20	0	0
40	30 – 40	35	75	15	-15	-15
30	20 – 30	25	60	20	-2	-40
20	10 – 20	15	40	15	-3	-45
10	0 – 10	5	25	25	-4	-100
				240		**+110**

$$\overline{X} = A + \frac{\sum fd}{N} \times C = 45 + \frac{110}{240} \times 10 = 45 + 4.583 = 49.583$$

Combined Mean:

If $n_1 \; and \; n_2$ are the sizes and $\overline{X_1}, \overline{X_2}$ are the respective means of two groups then the mean $\overline{X}$ of the combined group of size $n_1 + n_2$ is given by

$$\overline{X_c} = \frac{n_1\overline{X_1} + n_2\overline{X_2}}{n_1 + n_2}$$

MATHEMATICAL PROPERTIES OF ARITHMETIC MEAN:

Arithmetic Mean possesses some very interesting and important mathematical properties given below:

PROPERTY 1:

The algebraic sum of the deviations of the given set of 'n' observations from the Arithmetic Mean is zero.

$\sum(x - \overline{x}) = 0$

$\sum(x - n\overline{x}) = 0 \quad \therefore \sum x = n\overline{x}$

$n\overline{x} - n\overline{x} = 0$

PROPERTY 2:

The sum of the squares of deviations of the given set of observations is minimum when taken from the Arithmetic Mean.

Mathematically, for a given frequency distribution, the sum $S = \sum f(x - A)^2$
Which represents the sum of the squares of deviations from any arbitrary value A is minimum when $A = \overline{X}$.
This means that, if for any set of data, we compute
$S_1 = Sum \; of \; squares \; of \; deviation \; from \; mean = \sum f(x - \overline{x})^2$
$and \; S = Sum \; of \; squares \; of \; deviation \; from \; any \; point \; A = \sum f(x - A)^2, \; A \neq \overline{X}.$

Then S_1 is always less than S. $S_1 < S$

MERITS OF ARITHMETIC MEAN:

1. It is rigidly defined.
2. It is easy to calculate and easy to understand
3. It is based on all the observations.
4. It is suitable for further mathematical treatments.
5. It is all the averages, Arithmetic Mean are affected least by fluctuations of sampling.

DEMERITS:

1. It is very much affected by extreme observations.
2. It cannot be obtained by inspection nor located through a frequency graph.

WEIGHTED ARITHMETIC MEAN:

For calculating the Arithmetic Mean, we suppose that all the observations in the distribution have equal importance. But in practical life this may not be so. In case some times are more important than other, an A.M. computed is not represented of the distribution. In this situation proper weightages has to be given to the various items.

For example, to have an idea of the change in cost of living of a certain group of persons, this A.M. of the prices of the commodities consumed by them will not do because commodities are not equally important. In practical life (daily life) Rice, Wheat and Oils etc are more important than Coffee, Tea, Salt etc. If X_1, X_2, X_3……. X_n be the values of a variable X with respective weights are W_1, W_2, W_3, …….W_n assumed to them, then the weighted A.M. is given by

$$\overline{X_w} = \frac{W_1X_1+W_2X_2+\cdots..+W_nX_n}{W_1+W_2+\cdots..+W_n} = \frac{\sum W_iX_i}{\sum W_i}$$

GEOMETRIC MEAN:

The Geometric Mean of series containing 'n' observations is the n^{th} root of the product of the values X_1, X_2, X_3 …… X_n are n observations then the

$$G.M. = \sqrt[n]{X_1X_2X_3 \dots \dots .. X_n} = (X_1X_2X_3 \dots \dots .. X_n)^{1/n}$$

The product calculations are very difficult for simple calculations we use Logarithms on both sides.

$$\text{Log } G.M. = \frac{1}{n}\log(X_1X_2X_3 \dots \dots .. X_n) = \frac{1}{n}(\log X_1 + \log X_2 + \log X_3 + \cdots . + \log X_n)$$

$$= \log G.M. = \frac{\sum_{i=1}^{n} \log x_i}{n}$$

$\therefore$ G.M $= Antilog\ of\ \left[\frac{\sum \log x1}{n}\right]$ for individual series

For Grouped data:

$$GM = Antilog\left[\frac{\sum f \log x}{N}\right]$$

For Individual Series:

Ex: Compute the GM of the following 2, 4, 8, 12, 16 and 24.

Sol:

X	2	4	8	12	16	24	**Total**
Log x	0.3010	0.6021	0.9031	1.0792	1.2041	1.3802	**5.4697**

$$GM = Anti\log\left[\frac{\sum \log X}{n}\right] = A\ L\left[\frac{5.4697}{6}\right] = A\ L(0.9116) = 8.158$$

Ex: Calculate the Geometric Mean of the following series is given X: 0.4, 0.5, 5.0, 10.0, 45.0, 75, 125, 130, 150, 500.

X	0.4	0.5	5.0	10.0	45.0	75.0	125.0	130.0	150.0	500.0	**n = 10**
Log x	1.6021	1.6990	0.6990	1.0000	1.6532	1.8751	2.0969	2.1139	2.1761	2.6990	**Total = 13.6143**

$$GM = Anti\log\left[\frac{\sum \log X}{n}\right] = A\,L\left[\frac{13.6143}{10}\right] = A\,L(1.36143) = 22.98$$

For Discrete Series:

Ex: Calculate G.M of the following distribution.

Variable X	8	9	10	11	12	13	14
Frequency	11	8	6	9	7	3	1

Sol:

X	**f**	**Log X**	**f log X**
8	11	0.9031	9.9341
9	8	0.9542	7.6336
10	6	1.0000	6.0000
11	9	1.0414	9.3726
12	7	1.0792	7.5544
13	3	1.1139	3.3417
14	1	1.1461	1.1461
	45		**44.9325**

$$GM = Anti\log\left[\frac{\sum f \log X}{N}\right] = A\,L\left[\frac{44.9825}{45}\right] = A\,L(0.996) = 9.991$$

For Continuous Series :

Ex: Find G.M for the following distributions.

Marks:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of Students	5	7	15	25	8

Sol:

Marks	**No. of Students**	**(X) Mid Values**	**Log X**	**f log X**
0 – 10	5	5	0.6990	3.4950
10 – 20	7	15	1.1761	8.2327
20 – 30	15	25	1.3979	20.9685
30 – 40	25	35	1.5441	36.6025
40 – 50	8	45	1.6532	13.2256
	60			**84.5243**

$$GM = Anti\log\left[\frac{\sum f \log X}{N}\right] = A\,L\left[\frac{84.5243}{60}\right] = A\,L(1.40874) = 25.63$$

MERITS:

1. Geometric Mean is rigidly defined.
2. It is based on all the observations.
3. It is suitable for further mathematical treatments.
4. It is affected to lesser extent by extreme items.
5. It is not affected much by fluctuations of sampling.

DEMERITS:

1. G.M is not easy to understand and calculate for non – mathematical person.
2. If any item is zero, G.M becomes zero and if any one of the observation is negative, G.M. becomes imaginary regardless of the size of the other item.

HARMONIC MEAN:

DEFINITION:

Harmonic Mean is the reciprocal of the Arithmetic Mean of the reciprocals of the given observations.

If X_1, X_2, X_3……. X_n are 'n' observations, then their Harmonic Mean is given by

$H.M = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\cdots+\frac{1}{x_n}\right)} = \frac{1}{\frac{1}{n}\Sigma\left(\frac{1}{x_1}\right)} = \frac{n}{\Sigma\left(\frac{1}{x_1}\right)}$ This is for individual series (ungrouped data)

For Frequency distribution:

$H.M = \frac{1}{N}\left[\frac{f_1}{X_1}+\frac{f_2}{X_2}+\cdots\ldots+\frac{f_n}{X_n}\right] = \frac{f_1}{\frac{1}{n}\Sigma\left(\frac{f_1}{x_1}\right)} = \frac{N}{\Sigma\left(\frac{f}{x}\right)}$ $where\ N = \sum f$

For Individual Series:

Ex: Find the Harmonic Mean from the following data 2574, 465, 75, 5, 0.8, 0.08, 0.005, 0.0009

X	1/X
2574.00	0.00039
465.00	0.00215
75.00	0.01333
5.00	0.20000
0.80	1.25000
0.08	12.25000
0.005	200.00000
0.0009	1111.11111
	1324.8269

For Discrete Series:

Ex: The following table gives 31 persons in sample enquiry. Calculate H.M.

Weight:	130	135	145	146	148	149	150	157
No. of Persons:	3	4	6	6	3	5	2	1

Sol:

$Weight\ (X)$	$No.of\ Persons\ (f)$	$\frac{f}{X}$
130	3	0.02307
135	4	0.02964
140	6	0.04284
145	6	0.04140
146	3	0.02055
148	5	0.03380
149	2	0.01342
150	1	0.00667
157	1	0.00637
	31	**0.21776**

$$H.M = \frac{N}{\Sigma\left(\frac{f}{x}\right)} = \frac{31}{0.21776} = 142.36.$$

For Continuous Series:

Ex: Calculate Harmonic Mean from the following data:

Class	**Frequency**	**M.V**	**f/x**
0 – 10	12	5	2.4000
10 – 20	18	15	1.2000
20 – 30	27	25	1.0800
30 – 40	20	35	0.57174
40 – 50	17	45	0.3778
50 – 60	6	55	0.1091
	100		**5.7383**

$$H.M = \frac{N}{\Sigma\left(\frac{f}{x}\right)} = \frac{100}{5.7383} = 17.4268$$

MERITS:

1. Harmonic Mean is rigidly defined.
2. It is based on all observation.
3. It is suitable for further mathematical treatments.
4. It is not affected much by fluctuations of sampling.
5. It is useful for special types of rates and rations where time factor

DEMERITS:

1. It is not easy to calculate and understand.
2. If one item is zero, Harmonic Mean cannot be calculated.
3. It is not a representative figure of the distribution.

TRY YOURSELF:

1. Find H.M for the following variables.

   15, 250, 15.7, 157, 1.57, 105, 1.06, 25.7, 0.25 [Ans: 1.7337]

2. Calculate the H.M for the following

X:	5	10	15	20	25	30	25	40
F:	5	7	9	10	8	6	5	2

3. Calculate H.M. from the following table:

Class:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency:	8	12	14	16	15	9	6

MEDIAN:

Median is the value of the variable which divides the group into two equal parts. One part containing all the values greater than median and the other values less than median.

COMPUTATION OF MEDIAN:

Median M **(Find c.f. just greater than N/2, the corresponding class is Median class)**	**Individual series** **X:X_1, X_2, X_3, ….X_n**	**Discrete series** **X :** **f :**	**Continuous series** **C.I. :** **f :**
	$M = \left(\frac{n+1}{2}\right)^{th} value$	Sec the c.f., just greater than $\frac{N+1}{2}$ The corresponding value of the variable is median $N = \sum f$	Median $M = l + \frac{\frac{N}{2}-m}{f} \times c$ l = lower limit of the median class, m = cumulative frequency just before median class. f = the frequency of the median class c = is the class interval of the median class.

A) INDIVIDUAL SERIES:

Arrange the given values in ascending (increasing) order or descending order. If the number of observation is odd, median is the middle value. If the number of observations are even, median is the average of two middle values.

Median $M = \left(\frac{n+1}{2}\right)^{th}$ value for odd number of observations.

For even number of observations $M = \left\{\frac{\left(\frac{n}{2}\right)+\left(\frac{n}{2}+1\right)}{2}\right\}^{th}$ value.

B) DISCRETE SERIES:

In case of frequency distribution where the variable takes the values x_1, x_2, x_3, ……, x_n with respective frequencies f_1, f_2, f_3 …..f_n with $\sum f = N$, total frequency, median is the size of the $\left(\frac{n+1}{2}\right)^{th}$ item facilitates the calculations. The steps involved are:

i) Prepare the less than cumulative frequency.

ii) Find $\frac{N+1}{2}$

iii) Sec the c.f., just greater than $\frac{N+1}{2}$

iv) The corresponding value of the variable is median

C) CONTINUOUS SERIES:

The following steps involving for calculation of median value.

i) Prepare less than cumulative frequency.

ii) Find N/2 value.

iii) See c.f. just greater than N/2.

iv) The corresponding class contains the median value and is called median class.

v) The value of median is now obtained by using the formula.

Median $M = l + \frac{\frac{N}{2} - m}{f} \times c,$ where l is the lower limit of the median class.

m is the cumulative frequency just before median class (or) preceding class cumulative frequency of median class.

f is the frequency of the median class.

c is the class interval of the median class.

CALCULATING OF MEDIAN:

INDIVIDUAL SERIES:

Case i) : For odd number of observations:

Ex: Calculate median for the following variable.

6, 15, 25, 30, 8, 16, 18, 21, 24

Sol: Here n = 9 (odd number)

Arrange the given variable in ascending order i.e., 6, 8, 15, 16, 18, 21, 24, 25, 30

Median $M = \frac{n+1}{2} th\ value$

$= \frac{9+1}{2} = \frac{10}{2} = 5^{th}\ value$

5^{th} value is 18, so median m = 18.

Case ii) : Calculate median for the following values

20, 18, 16, 35, 40, 48, 28, 12, 42, 26

Sol: Here n = 10 (even number)

Arrange the given values in ascending order

i.e., 12, 16, 18, 20, 26, 28, 35, 40, 42, 48

Median, $M = \frac{\frac{n}{2}+\left(\frac{n}{2}+1\right)}{2}\ th\ value = \frac{\frac{10}{2}+\left(\frac{10}{2}+1\right)}{2} = \frac{5th\ value+(5+1)th\ value}{2}$

$M = \frac{26+28}{2} = \frac{54}{2} = 27$

DISCRETE SERIES:

Ex: Calculate median for the following distribution:

X :	12	18	23	28	34	38	45	60
f :	3	7	12	28	32	18	12	8

Sol:

X	f	Cumulative Frequency
12	3	3
18	7	3+7=10
23	12	10+12=22
28	28	22+28=50
34	32	50+32=82
38	18	82+18=100
45	12	100+12=112
60	8	112+8=120
	N=120	

$M = \left(\frac{N+1}{2}\right) th\ value = \frac{120+1}{2} = \frac{121}{2} = 60.5$ value will lie in c.f. 82. So the corresponding X value 34 is median. i.e., M = 34.

Ex: The frequency distribution of weight in grams of mangoes of a given variety is given below. Calculate the median.

Wt in gms:	410 – 419	420 – 429	430 – 439	440 – 449	450 – 459	460 – 469	470 – 479
No. of mangoes:	14	20	42	54	45	18	7

Sol: For calculation median in continuous series we change the class intervals are exclusive class.

Class	f	Cumulative Frequency
409.5 – 419.5	14	14
419.5 – 429.5	20	34
429.5 – 439.5	42	76
439.5 – 449.5	54	130

449.5 – 459.5	45	175
459.5 – 469.5	18	193
469.5 – 479.5	7	200
	200	

$$M = l + \frac{\frac{N}{2} - m}{f} \times c$$

Here, $\frac{N}{2} = \frac{200}{2} = 100$

Then $\frac{N}{2} = 100$ value will lie in c.f. 130, the corresponding class 439.5 – 449.5 in median class.

L = lower limit of the median class = 439.5.

m = preceeding class c.f. of the median class = 76

f = frequency of median class = 54

c = class interval = 10

$$m = 439.5 + \frac{100-76}{54} \times 10 = 439.5 + \frac{240}{54} = 439.5 + 4.44 = 443.9\ grams$$

1. Calculate median Q_1 and Q_3 from the following data

X:	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20	20 – 24	24 - 28	28 – 32
f :	5	25	40	70	90	40	20	10

Sol:

Class	**f**	**c.f.**
0 – 4	5	5
4 – 8	25	30
8 – 12	40	70
12 – 16	70	140
16 – 20	90	230
20 – 24	40	270
24 – 28	20	290
28 – 32	10	300
	300	

$$Median\ m = l + \frac{\frac{N}{2} - m}{f} \times c$$

Here, $\frac{N}{2} = \frac{300}{2} = 150$

Then $\frac{N}{2} = 150$ value will lie in c.f. 230, the corresponding class 16 – 20 in median class.

L = lower limit of the median class = 16.

m = preceding class c.f. of the median class = 140

f = frequency of median class = 90

c = class interval = 4

$m = 16 + \frac{150-140}{90} \times 4 = 16 + \frac{10}{90} \times 4 = 16 + \frac{40}{90} = 16 + 0.44 = 16.44$

$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$

$\frac{N}{4} = \frac{300}{4} = 75$ will lie in c.f. 140, the corresponding class 12 – 16 is 1st Quartile class.

$l = 12, m = 70, f = 70, c = 4$

$Q_1 = 12 + \frac{75-79}{70} \times 4 = 12 + \frac{5}{70} \times 4 = 12 + \frac{20}{70} = 12 + 0.2857 = 12.2857$

$Q_3 = l_3 + \frac{\frac{3N}{4} - m_3}{f_3} \times c_3$

$\frac{3N}{4} = \frac{3(300)}{4} = \frac{900}{4} = 225$ will lie in c.f. 230, the corresponding class 16 – 20 is 3rd Quartile class.

$l_3 = 16, m_3 = 140, f_3 = 90, c_3 = 4$

$Q_3 = 16 + \frac{225-140}{90} \times 4 = 16 + \frac{85}{90} \times 4 = 16 + \frac{340}{90} = 16 + 3.7778 = 19.7778.$

MODE:

Mode is the value which occurs most frequently in a set of observations and around which the other items of the set cluster density.

According A.M. Tuttle 'mode is the value which has the greatest frequency density in its immediate neighborhood'.

COMPUTATION OF MODE:

A) INDIVIDUAL SERIES:

More number of repeated value is mode.

Ex: 3, 15, 20, 22, 28, 20, 32, 20, 32, 20, 22

In the above example the value 20 is more no. of (4) times repeated so mode Z = 20.

B) DISCRETE SERIES:

In this series, the highest frequency and the corresponding variable X value is mode.

Ex: Calculate mode from the following table:

X:	3	10	15	22	28	32	41	55
f :	2	8	20	32	50	30	26	10

Sol: In the above problem the highest frequency in 50, the corresponding X value 28 is mode.

Therefore mode Z = 28.

LOCATION OF MODE BY GROUPING METHOD:

The following steps involving to locate the mode.

1. In 1st column we write down the given frequencies.

2. Column 2 is obtained by adding the frequencies two by two.

3. Leave the 1st frequency and adding the remaining frequencies two by two and write in column 3.

4. Column 4 is obtained by adding the frequencies three by three.

5. Eliminate the 1st frequency and add by three and write in column 5.

6. Eliminate 1st and 2nd frequencies and adding the remaining frequencies three by three and write in column 6.

7. Mark the highest frequency in each column.

8. Then from analysis table to find the model class.

After finding the model class, use the formula for calculating mode value.

$Z = l + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times c$

$where\ l = lower\ limit\ of\ the\ model\ class\ ;\ f_0 = frequency\ of\ the\ model\ class$

$f_1 = preceding\ class\ frequency\ of\ the\ model\ class$

Ex: Calculate mode from the following data (Grouping table):

Class	**f_1**	**II**	**III**	**IV**	**V**	**VI**
0 – 10	4					
10 – 20	6	10				
20 – 30	20		26	30		
30 – 40	32	52			58	
40 – 50	33		65	82		85
50 – 60	17	50			58	
60 – 70	8		25			27
70 – 80	2	10				

Sol: Analysis table

Columns	**0 – 10**	**10 – 20**	**20 – 30**	**30 – 40**	**40 – 50**	**50 – 60**	**60 – 70**	**70 – 80**
I					√			
II			√	√				
III				√	√			
IV			√	√	√			
V		√	√	√	√	√	√	
VI			√	√	√			
		1	**4**	**5**	**5**	**1**	**1**	

In analysis table the class 30 – 40 and 40 – 50 maximum (5) times, in this situation the formula for mode I is not useful. Then this situation we use the empirical formula for calculating mode.

Mode Z = 3 median – 2 mean.

$Z = 3m - 2\overline{X}$

Calculation of $\overline{X}$ *and median*:

Class	f	Mv (x)	fx	Cumulative Frequency
0 – 10	4	5	20	4
10 – 20	6	15	90	10
20 – 30	20	25	500	30
30 – 40	32	35	1120	62
40 – 50	33	45	1485	95
50 – 60	17	55	935	112
60 – 70	8	65	520	120
70 – 80	2	75	150	122
	122		**4820**	

Median $m = l + \frac{\frac{N}{2} - m}{f} \times c$

$\frac{N}{2} = \frac{122}{2} = 61$ will lie in c.f. 62, the corresponding class 30 – 40 is median class.

$l = 30, m = 30, f = 32, c = 10$

$m = 30 + \frac{61-30}{32} \times c = 30 + \frac{31}{2} \times 10 = 30 + 9.6875 = 39.6875$

$\overline{X} = \frac{\Sigma fx}{N} = \frac{4820}{122} = 39.508$

$Z = 3m - 2\overline{X} = 3(39.6875) - 2(39.508) = 119.0625 - 79.016 = 40.0465.$

Ex: Calculate mode from the following data:

Class:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency:	4	12	20	45	30	8	4

Sol:

X	f_1	II	III	IV	V	VI
0 – 10	4	16				
10 – 20	12		32	36		
20 – 30	20	25			77	95
30 – 40	45		75	83		
40 – 50	30	38			42	
50 – 60	8		12			
60 – 70	4					

Sol: Analysis table

Columns	**0 – 10**	**10 – 20**	**20 – 30**	**30 – 40**	**40 – 50**	**50 – 60**	**60 – 70**
I				√			
II			√	√			
III				√	√		
IV				√	√	√	
V		√	√	√			
VI			√	√	√		
		1	**3**	**6**	**3**	**1**	

Mode $Z = l + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times c$;

$l = 30, f_0 = 45, f_1 = 20, f_2 = 30, c = 10$

$Z = 30 + \frac{45-20}{2(45)-20-30} \times 10 = 30 + \frac{25}{90-50} \times 10 = 30 + \frac{25}{40} \times 10 = 30 + 6.25 = 36.25.$

Ex: Calculate Mode from the following data:

X:	10 – 20	20 – 30	30 – 40	40 - 50	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100	100 – 110
f :	4	6	5	10	20	22	24	6	2	1

Sol: Grouping Method:

X	**f_1**	**II**	**III**	**IV**	**V**	**VI**
10 – 20	4					
20 – 30	6	10				
30 – 40	5		11			
40 – 50	10	15		15		
50 – 60	20		30		21	
60 – 70	22	42		52		35
70 – 80	24		46		66	
80 – 90	6	30		32		52
90 – 100	2		8		9	
100 – 110	1	3				

MERITS:

1. Mode is easy to calculate and easy to understand.
2. Mode is not affected by extreme values
3. It can be conveniently obtained in the case of open end.

DEMERITS:

1. Mode is not rigidly defined.
2. It is not based on all the observations.

3. It is the grouping method, also gives two values of mode the distribution is called bi-model distribution.
4. Mode is not suitable for further mathematical treatments.

TRY YOURSELF:

1. Calculate mode from the following table:

   15, 32, 29, 17, 35, 32, 35, 36, 32, 34, 33, 32

2. Calculate mode from the following data:

X:	5	10	15	200	25	30	35	40
F :	4	7	21	47	53	24	10	3

3. Calculate mode for the following data:

Class:	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45
Frequency:	5	16	36	40	48	23	12

NOTE:

1. Relation between A.M, G.M and H.M.

   A.M. $\geq$ G.M. $\geq$ H.M.

2. Relation between A.M., Median and Mode

   A.M > Median > Mode

   A.M. < Median < Mode

3. If a and b are two observations then

   i) The A.M $= \frac{a+b}{2}$

   ii) G.M $= \sqrt{a \times b}$

   iii) H.M $= \frac{2ab}{a+b}$

   iv) $GM^2 = (AM)(HM)$

DISPERSION:

The averages (or) the measures of central tendency gives us an idea to locate the center of the distribution, but they do not say how the observations and spares. Out on the center value. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the observations are not equal. There is difference among the values. The dispersion is large it indicate the less uniformity, the dispersion is small if indicate the high uniformity of observations.

MEASURES OF DISPERSION:

INTRODUCTION:

If the items within a distribution differ from one another in magnitude the term dispersion or scatteredness is used to indicate the difference. The distribution differ from one another in respect of two main characteristics.

1. They may differ in measures of central tendency.

2. They may have the same measure of central tendency but have wide disparities in the formation of distributions.

Consider the two series 3, 4, 5, 6, 7 and 12, 13, 14, 15, 16. The arithmetic means of the two series are 5 and 14. Although the means are different the items in the two series are scattered in the same way around the means. Next consider two other series 5, 8, 10, 4, 3 and 6, 15, 0, 7, 2. These two series have the same arithmetic mean 6 but the scatteredness of the various items in the two series about their mean is different. From the above two examples we infer that the average fails to give us an idea how the various items are scattered and how the distributions are constituted. It is therefore necessary to have measures of scatteredness or dispersion in order to study the distribution fully. Measures of dispersion enable us to compare the variability of two or more frequency distributions.

The measures of dispersion in common use are:

1. Range
2. Mean deviation
3. Standard deviation

RANGE:

Range for an ungrouped data is defined as the difference between the greatest and the least values of the variate.

For a grouped data range is defined as the difference between the upper limit of the largest class and lower limit of the smallest class.

Ex: 1. Find the range of marks of students in a class given as

60, 72, 96, 28, 35, 10, 40, 9, 85, 25.

Range = Largest value – Smallest value = 96 – 9 = 87.

Ex: 2. The following table gives the daily sales (Rs.) of two firms A and B for five days.

Firm A	**Firm B**
5050	4900
5025	3100
4950	2200
4835	1800
5140	13000
$\boldsymbol{\overline{X_A} = 5000}$	$\boldsymbol{\overline{X_B} = 5000}$

The sales of both the firms in average is same but distribution pattern is not similar. There is a great amount of variation in the daily sales of the firm B than that of the firm A.

Range of sales of firm A = Greatest value – Smallest value = 5140 – 4835 = 305

Range of sales of firm B = Greatest value – Smallest value = 13000 – 1800 = 11200

MEAN DEVIATION:

Mean deviation is defined as arithmetic average of absolute values of the deviations of the variates measured from an average (median, mode or mean).

The absolute value of the deviation denoted by $|deviation|$ is the numerical value of the deviation with positive sign.

NOTE:

Mean deviation can be similarly calculated by taking deviations from the median or mode.

MEAN DEVIATION FROM MEAN OF AN UNGROUPED DATA:

Let $x_1, x_2, x_3 \ldots\ldots x_n$ be the values of n variates and $\overline{x}$ be their arithmetic mean. Let $|x_i - \overline{x}|$ be the absolute value of the deviation of the variate $x_i \; from \; \overline{x}$.

$\therefore$ Mean deviation $= \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n}$

SOLVED EXAMPLES:

1. Calculate the mean deviation of the variates 40, 62, 54, 68, 76 from A.M.

Sol: A.M. $= \overline{x} = \frac{40+62+54+68+76}{5} = \frac{300}{5} = 60$

Mean deviation $= \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n} = \frac{|40-60|+|62-60|+|54-60|+|68-60|+|76-60|}{5} = \frac{52}{5} = 10.4$

2. Find the mean deviation from the mean for the following data: 38, 70, 48, 40, 42, 55, 63, 46, 54, 44.

Sol: Mean $= \overline{x} = \frac{38+70+48+40+42+55+63+46+54+44}{10} = \frac{500}{10} = 50$

The deviations of the given observations from $\overline{x}$:

x_i	38	70	48	40	42	55	63	46	54	44
$x_i - \overline{x}$	38 – 50	70 – 50	48 – 50	40 – 50	42 – 50	55 – 50	63 – 50	46 – 50	54 – 50	44 – 50

Mean deviation from the mean $= \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n} = \frac{84}{10} = 8.4.$

3. Find the main deviation about the a) mean b) median for the following distribute data 6, 7, 10, 12, 13, 4, 12, 16.

Sol: The arithmetic mean of the given data is

a) $\overline{x} = \frac{6+7+10+12+13+4+12+16}{8} = \frac{80}{8} = 10$

The absolute values of deviation from A.M. are

$(6-10), (7-10), (10-10), (12-10), (13-10), (14-10), (12-10), (16-10)$

$i.e., 4, 3, 0, 2, 3, 6, 2, 6$

Mean deviation from the mean $= \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n} = \frac{4+3+0+2+3+6+2+6}{8} = \frac{26}{8} = 3.25.$

b) Writing the data in ascending order magnitude, we get 4, 6, 7, 10, 12, 12, 13, 16

The median b of these observations is $\frac{10+12}{2} = \frac{22}{2} = 11$

The absolute values of the deviations from the median are

$|4-11|, |6-11|, |7-11|, |10-11|, |12-11|, |12-11|, |13-11|, |16-11|$

$i.e., 7, 5, 4, 1, 11, 2, 5$

Mean deviation from median $= \frac{\sum |x_i - b|}{n} = \frac{26}{8} = 3.25$

4. Find the mean deviation from the median for the data 34, 66, 30, 38, 44, 50, 40, 60, 42, 51.

Sol: Arranging the data in ascending order, we have:

0, 34, 38, 40, 42, 44, 50, 51, 60, 66. (n = 10 terms)

Now median $= \frac{\left(\frac{n}{2}\right)^{th} term + \left(\frac{n}{2}+1\right)^{th} term}{2} = \frac{42+44}{2} = 43$

x_i	30	34	38	40	42	44	50	51	60	66
$x_i - Med$	30 – 43	34 – 43	38 – 43	40 – 43	42 – 43	55 – 43	63 – 43	46 – 43	54 – 43	44 – 43

$\therefore$ Mean deviation from median $= \frac{\sum |x_i - Median|}{n} = \frac{\sum |x_i - 43|}{n} = \frac{87}{10} = 8.7$

MEAN DEVIATION FOR A GROUPED DATA:

We know that data can be arranged as a frequency distribution in two ways

i) Discrete frequency distribution and ii) Continuous frequency distribution.

MEAN DEVIATION ABOUT MEAN OF A GROUPED DATA WITH DISCRETE FREQUENCY DISTRIBUTION:

Let $x_1, x_2, \ldots..x_n$ be the midvalues of n class intervals with frequencies $f_1, f_2, \ldots\ldots, f_n$ of a frequency distribution. Let $\overline{x}$ be the arithmetic mean of the distribution. Let $|x_i - \overline{x}|$ be the absolute value of the deviation of the midvalue x_i from the arithmetic mean $\overline{x}$.

Then the mean deviation about the arithmetic mean $= \frac{|x_i - \overline{x}| f_1 + |x_i - \overline{x}| f_2 + \cdots.. + |x_i - \overline{x}| f_n}{f_1 + f_2 + \cdots.. + f_n}$

$= \frac{\sum_{i=1}^{n} f_i |x_i - \overline{x}|}{\sum_{i=1}^{n} f_i} = \frac{1}{n} \sum_{i=1}^{n} f_i |x_i - \overline{x}|$ $where$ $\sum_{i=1}^{n} f_i = N$

SOLVED EXAMPLES:

1. Find the mean deviation about the mean for the following data.

x_i	2	5	7	8	10	35
f_i	6	8	10	6	8	2

Sol: We will tabulate the values as follows:

x_i	f_i	$f_i x_i$	$\lvert x_i - \overline{x} \rvert$	$f_i \lvert x_i - \overline{x} \rvert$
2	6	12	6	36
5	8	40	3	24
7	10	70	1	10
8	6	48	0	0
10	8	80	2	16
35	2	10	27	54
	$\sum f_i = N = 40$	$\sum f_i x_i = 320$		**140**

Thus A.M $\overline{\boldsymbol{x}} = \frac{\sum \boldsymbol{f_i x_i}}{\sum \boldsymbol{f_i}} = \frac{\boldsymbol{320}}{\boldsymbol{40}} = \boldsymbol{8}$

$\therefore$ Mean deviation $= \frac{\sum_{i=1}^{n} f_i |x_i - \overline{x}|}{\sum_{i=1}^{n} f_i} = \frac{140}{40} = 3.5$

2. Find the mean deviation from the median for the following data.

x_i	6	9	3	12	15	13	21	22
f_i	4	5	3	2	5	4	4	3

Sol: We write observations in ascending order to get the table as follows:

x_i	3	6	9	12	13	15	21	22
f_i	3	4	5	2	4	5	4	3

Here $N = \sum f_i = 30$

$\therefore$ Median is the mean of 15th and 16th observations which are equal to 13. Now we tabulate the absolute values of the deviations.

$\lvert x_i - \boldsymbol{Med}\rvert = \lvert x_i - \boldsymbol{13}\rvert$	10	7	4	1	0	2	8	9
f_i	3	4	5	2	4	5	4	3
$f_i \lvert x_i - \boldsymbol{Med}\rvert$	30	28	20	2	0	10	32	27

Thus $\sum f_i |x_i - Med| = 149$

$\therefore$ Mean deviation from median $= \frac{\sum f_i |x_i - Median|}{\sum f_i} = \frac{149}{30} = 4.97$

3. Calculate the mean deviation from median from the following data:

Size of item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Sol: We tabulate the values as follows:

Size	**Frequency**	**Cumulative frequency**	**Deviations from Median** $\lvert x_i - \boldsymbol{Med}\rvert$	$f_i \lvert x_i - \boldsymbol{med}\rvert$
6	3	3	3	9
7	6	9	2	12
8	9	18	1	9
9	13	31	0	0
10	8	39	1	8
11	5	44	2	10
12	4	48	3	12

Here no. of values, n = 7 (odd), $\therefore$ Median $= \frac{\left(\frac{n}{2}\right)^{th} term + 1}{2} = \frac{7+1}{2} = 4^{th} term = 9$

Mean deviation from median $= \frac{\sum f_i |x_i - Median|}{\sum f_i} = \frac{60}{48} = 1.25$

4. Find the mean deviation from the mean for the following data:

x_i	5	10	15	20	25
f_i	7	4	6	3	5

Sol: Calculations of mean deviations about mean

x_i	f_i	$f_i x_i$	$\|x_i - \overline{x}\|$	$f_i\|x_i - \overline{x}\|$
5	7	35	$\|5 - 14\| = 9$	63
10	4	40	$\|10 - 14\| = 4$	16
15	6	90	$\|15 - 14\| = 1$	6
20	3	60	$\|20 - 14\| = 6$	18
25	5	125	$\|25 - 14\| = 11$	5
	$\sum f_i = 25$	$\sum f_i x_i = 350$		**158**

Mean, $\overline{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{350}{25} = 14$

$\therefore$ Mean deviation from mean $= \frac{\sum_{i=1}^{n} f_i|x_i - \overline{x}|}{\sum_{i=1}^{n} f_i} = \frac{158}{25} = 6.32.$

MEAN DEVIATION FROM THE MEAN FOR A CONTINUOUS FREQUENCY DISTRIBUTION:

A continuous frequency distribution is a series in which the data is classified into different class intervals along with respective frequency. We calculate the A.M. of a continuous frequency distribute, we take x_i as the mid value of the class interval.

SOLVED EXAMPLES:

1. The following table gives the sales of 100 companies. Find the mean deviation from the mean.

Sales in thousands	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90	90 - 100
Number of companies	5	15	25	30	20	5

Sol: We shall construct the following table for the given data.

Sales	**Number of companies (f_i)**	**Mid point of the class (x_i)**	$f_i x_i$	$\|x_i - \overline{x}\|$	$f_i\|x_i - \overline{x}\|$
40 – 50	5	45	225	26	130
50 – 60	15	55	825	16	240
60 – 70	25	65	1625	6	150
70 – 80	30	75	2250	4	120
80 – 90	20	85	1700	14	280
90 – 100	5	95	475	24	120
	$\sum f_i = N = 100$		$\sum f_i x_i = 7100$		$\sum f_i\|x_i - \overline{x}\| = 1040$

Now $\overline{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{7100}{100} = 71$

$\therefore$ Mean deviation from mean $= \frac{\sum_{i=1}^{n} f_i |x_i - \overline{x}|}{\sum_{i=1}^{n} f_i} = \frac{1040}{100} = 10.4.$

2. Find the mean deviation of the following frequency distribution.

Class Interval	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20	20 – 24
Frequency	8	12	35	25	13	7

Sol:

Class Interval	**Mid value (x_i)**	**Frequency (f_i)**	$f_i x_i$	$\lvert x_i - \overline{x} \rvert$	$f_i \lvert x_i - \overline{x} \rvert$
0 – 4	2	8	16	9.76	78.08
4 – 8	6	12	72	5.76	69.12
8 – 12	10	35	350	1.76	61.60
12 – 16	14	25	350	2.24	56.00
16 – 20	18	13	234	6.24	81.12
20 – 24	22	7	154	10.24	71.68
		$\sum f_i = N = 100$	$\sum f_i x_i = 1176$		$\sum f_i \lvert x_i - \overline{x} \rvert = 417.60$

Here, $\overline{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1176}{100} = 11.76$

$\therefore$ Mean deviation $= \frac{1}{N} \sum f_i |x_i - \overline{x}| = \frac{417.60}{100} = 4.176$

STEP DEVIATION METHOD (SHORT CUT METHOD):

Suppose in the given data the midpoints of the class intervals x_i and their associated frequencies are numerically large. Then the computations becomes tedious. To avoid large calculations, we take an assumed mean a which lies in the middle or close to it in the data and take the deviations of the mid points x_i from this assumed mean. This is equal to shifting the origin from 0 to assumed mean on the number line.

Again, if there is a common factor of all the deviations, we divide them by their common factor (h) to further simplify the deviations. These are known as **Step Deviations**. This amounts to change of scale on the number line.

With the assumed mean a and a common factor h, we define a new variable, $d_i = \frac{x_i - a}{h}$

Then A.M. $= \overline{x} = \left(\frac{\sum f_i d_i}{N} \right) h$

We illustrate the simplified procedure with some examples.

SOLVED EXAMPLES:

1. Find the mean deviation from the mean for the following data:

Classes	0 – 100	100 – 200	200 – 300	300 – 400	400 – 500	500 – 600	600 – 700	700 – 800
Frequency	4	8	9	10	7	5	4	3

Sol: We tabulate the data as follows:

Classes	**Mid value (x_i)**	d_i	**Frequency (f_i)**	$f_i d_i$	$\|x_i - \bar{x}\|$	$f_i\|x_i - \bar{x}\|$
0 – 100	50	-3	4	-12	308	1232
100 – 200	150	-2	8	-16	208	1664
200 – 300	250	-1	9	-9	108	972
300 – 400	350	0	10	0	8	80
400 – 500	450	1	7	7	92	644
500 – 600	550	2	5	10	192	960
600 – 700	650	3	4	12	290	1168
700 – 800	750	4	3	12	392	1176
			$\sum f_i = N = 50$	$\sum f_i x_i = 4$		$\sum f_i\|x_i - \bar{x}\| = 7896$

$d_i = \frac{x_i - assumed\ mean}{class\ size} = \frac{x_i - 350}{100}$

Now $\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i} \times class\ size = 350 + \frac{4}{50} \times 100 = 358$

∴ Mean deviation from mean $\frac{\sum_{i=1}^{n} f_i|x_i - \bar{x}|}{\sum_{i=1}^{n} f_i} = \frac{7896}{50} = 157.92$

2. Find the mean deviation about the mean finite following data:

Marks obtained	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students	4	8	9	10	7

Sol: We form the following table:

Classes	**Mid value (x_i)**	d_i	**Frequency (f_i)**	$f_i d_i$	$\|x_i - \bar{x}\|$	$f_i\|x_i - \bar{x}\|$
0 – 10	5	-2	5	-10	22	110
10 – 20	15	-1	8	-8	12	96
20 – 30	25	0	15	0	2	30
30 – 40	35	1	16	16	8	128
40 – 50	45	2	5	12	18	108
			$\sum f_i = N = 50$	$\sum f_i d_i = 10$		$\sum f_i\|x_i - \bar{x}\| = 472$

$d_i = \frac{x_i - \bar{x}}{h} = \frac{x_i - 27}{10}$

Now $\bar{x} = a + \frac{\sum f_i d_i}{N} \times c = 25 + \frac{10}{50} \times 10 = 27$

∴ Mean deviation from mean $\frac{\sum_{i=1}^{n} f_i|x_i - \bar{x}|}{N} = \frac{472}{50} = 9.44$

VARIANCE AND STANDARD DEVIATION OF UNGROUPED / GROUPED DATA:

Earlier in calculating the mean deviation about mean or median, we were taking the absolute values of the deviations, so that they may not cancel among themselves. Again we adopt another method to avoid the difficulty that arise due to the signs of the deviations. We consider the squares of the deviations to make them non-negative. Thus if $x_1, x_2, \ldots x_n$ are n observations and $\overline{x}$ is their mean, we consider $\sum(x_i - \overline{x})^2$.

The following cases may arise:

* Case (i) : If $\sum(x_i - \overline{x})^2 = 0, then\ each\ (x_i - \overline{x}) = 0$ which implies that all observations are equal to the mean $\overline{x}$ and there is no dispersion.

* Case (ii) : If $\sum(x_i - \overline{x})^2\ is\ small,$ then it shows that each observation x_i is very close to the mean $\overline{x}$ and hence the degree of dispersion is very low.

* Case (iii) : If $\sum(x_i - \overline{x})^2\ is\ large$, then it indicates that a higher degree of dispersion of observations from the mean $\overline{x}$.

If we take the mean of the squared deviations from the mean i.e., $\frac{1}{n}\sum(x_i - \overline{x})^2$, then it is found that this number leads to a proper measure of dispersion. The number is called variance and is denoted by σ^2. Then σ the standard deviation is given by the positive square root of variance.

Variance $\sigma^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$

Standard deviation $= \sigma = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$

SOLVED EXAMPLES:

1. Find the variance and standard deviation of the following data:

   5, 12, 3, 18, 6, 8, 2, 10.

Sol: We construct the following table to calculate variance and standard deviation.

x_i	5	12	3	18	6	8	2	10
$(x_i - \overline{x})$	-3	4	-5	10	-2	0	-6	2
$(x_i - \overline{x})^2$	9	16	25	100	4	0	36	4

Here $\sum(x_i - \overline{x})^2 = 194$

Variance $\sigma^2 = \frac{1}{n}\sum(x_i - \overline{x})^2 = \frac{194}{8} = 24.25$

Standard deviation $\sigma = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2} = \sqrt{24.75} = 4.95\ (approx.)$

2. Find the variance and standard deviation for the following data:
   45, 60, 62, 60, 50, 65, 58, 68, 44, 48

Sol: Mean $\overline{x} = \frac{45+60+62+60+50+65+58+68+44+48}{10} = \frac{560}{10} = 56$

x_i	45	60	62	60	50	65	58	68	44	48
$(x_i - \overline{x})$	-11	4	6	4	-6	9	2	12	-12	-8
$(x_i - \overline{x})^2$	121	16	36	16	36	81	4	144	144	64

Here $\sum(x_i - \overline{x})^2 = 662$

Variance $\sigma^2 = \frac{1}{n}\sum(x_i - \overline{x})^2 = \frac{662}{10} = 66.2$

Standard deviation $= \sigma = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2} = \sqrt{66.2} = 8.136$

VARIANCE AND STANDARD DEVIATION FOR A DISCRETE FREQUENCY DISTRIBUTION :

3. Calculate the standard deviation for the following distribution.

x_i	6	10	14	18	24	28	30
f_i	2	4	7	12	8	4	3

Sol: We have

x_i	f_i	$f_i x_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_i(x_i - \overline{x})^2$
6	2	12	-13	169	338
10	4	40	-9	81	324
14	7	98	-5	25	175
18	12	216	-1	1	12
24	8	192	5	25	200
28	4	112	9	81	324
30	3	90	11	121	363
	$\sum f_i = \mathbf{40}$	$\sum f_i x_i = \mathbf{760}$			**1736**

Here, N = 40, Mean, $\overline{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{760}{40} = 19$ *and* $\sum f_i(x_i - \overline{x})^2 = 1736$

Variance $\sigma^2 = \frac{\sum f_i(x_i - \overline{x})^2}{\sum f_i} = \frac{1736}{40} = 43.4$

Standard deviation $\sigma = \sqrt{43.4} = 6.59$

4. Find the mean standard deviation of the following frequency distribution.

x_i	4	8	11	17	20	24	32
f_i	3	5	9	5	4	3	1

Sol: We construct following table for computing the required values.

x_i	f_i	$f_i x_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_i\lvert x_i - \overline{x}\rvert$
4	3	12	-10	100	300
8	5	40	-6	36	180
11	9	99	-3	9	81
17	5	85	3	9	45
20	4	80	6	36	144
24	3	72	10	100	300
32	1	32	18	324	324
	$\sum f_i = \mathbf{30}$	$\sum f_i x_i = \mathbf{420}$			**1374**

Here, N = 30, Mean, $\overline{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{420}{30} = 14$

Variance $\sigma^2 = \frac{\sum f_i (x_i - \overline{x})^2}{\sum f_i} = \frac{1374}{30} = 45.8$

Standard deviation $\sigma = \sqrt{45.8} = 6.77$

VARIANCE AND STANDARD DEVIATION OF A CONTINUOUS FREQUENCY DISTRIBUTION:

If there are n classes in given distribution, each class represented by its mid point x_i and corresponding frequency f_i, then we calculate standard deviation using the formula.

$\sigma = \sqrt{\frac{\sum f_i (x_i - \overline{x})^2}{\sum f_i}}$ where $N = \sum f_i \ and \ \overline{x}$ is the mean of the distribution.

ANOTHER METHOD:

To avoid the tediousness of calculation and to simplify the calculation, we adopt the following alternative method.

We know that variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2 = \frac{1}{N}\sum f_i \left(x_i^2 + \overline{x}^2 - 2x_i\overline{x}\right)$

$= \frac{1}{N}\left[\sum_{i=1}^{n} f_1 x_i^2 + \sum_{i=1}^{n} f_i \overline{x}^2 - \sum_{i=1}^{n} 2\overline{x} f_i x_i\right] = \frac{1}{N}\left[\sum_{i=1}^{n} f_i x_i^2 + \overline{x}^2 . N - 2\overline{x} N x\right]$

$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 + \overline{x} - 2\overline{x}^2 = \frac{1}{N}\sum f_i x_i^2 - \left(\overline{x}^2\right) = \frac{1}{N}\sum f_i x_i^2 - \left(\frac{\sum f_i x_i}{N}\right)^2$

$\therefore$ Standard Deviation $\sigma = \sqrt{\frac{1}{N}\sum f_i x_i^2 - \left(\frac{\sum f_i x_i}{N}\right)^2}$

STEP DEVIATION METHOD (SHORT CUT METHOD):

If the mid values x_i in the continuous distribution are large, the calculation of mean are variance becomes difficult. In such cases we apply the step deviation method, as described below.

Let h be the length of the class interval and A is the assumed mean.

We define $y_i = \frac{x_i - A}{h}, i = 1, 2, \ldots\ldots, n. Then \ x_i = A + hy_i$

$\overline{x} = \frac{\sum_{i=1}^{n} f_1 x_i}{N} = \frac{\sum f_i (A + hy_i)}{N} = \frac{1}{N}\left[\sum_{i=1}^{n} A f_i + \sum_{i=1}^{n} h f_i y_i\right] = \frac{1}{N}\left[A\sum_{i=1}^{n} f_i + h\sum_{i=1}^{n} f_i y_i\right]$

$= A + h\sum_{i=1}^{n} f_i y_i = A + h\overline{y}$

$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \overline{x})^2 = \frac{1}{N}\sum f_i (A + hy_i - A - h\overline{y})^2 = \frac{1}{N}\sum_{i=1}^{n} f_i h^2 (y_i - \overline{y})^2$

$= h^2\left[\frac{1}{N}\sum f_i (y_i - \overline{y})^2\right] = h^2\sigma_y^2 \ (or) \sigma_x = h\sigma_y$

But we know that, S.D. $\sigma_x = \frac{1}{N}\sqrt{N\sum f_i x_i^2 - (\sum f_i x_i)^2}$

$= \frac{1}{N}\sqrt{N\sum f_i y_i^2 - (\sum f_i y_i)^2}$

$\therefore \sigma_x = \frac{h}{N}\sqrt{N\sum_{i=1}^{n} f_i y_i^2 - (\sum f_i y_i)^2}$

5. Calculate the variance and standard deviation of the following continuous frequency distribution.

Class Interval	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100
Frequency	3	7	12	15	8	3	2

Sol: We tabulate the following way:

Classes	**Frequency** (f_i)	**Mid value** (x_i)	$y_i = \frac{x_i - A}{h}$ $A = 65, h = 10$	y_i^2	$f_i y_i$	$f_i y_i^2$
30 – 40	3	35	-3	9	-9	27
40 – 50	7	45	-2	4	-14	25
50 – 60	12	55	-1	1	-12	12
60 – 70	15	65	0	0	0	0
70 – 80	8	75	1	1	8	8
80 – 90	3	85	2	4	6	12
90 – 100	2	95	3	9	6	18
	$\sum f_i = N = 50$				$\sum f_i y_i = -15$	$\sum f_i y_i^2 = 105$

Assumed mean = A = 65,

Length of the class interval = h = 10

Mean $= \bar{x} = A + \frac{\sum f_i y_i}{N} \times h = 65 - \left(\frac{15}{50} \times 10\right) = 62$

Variance $\sigma^2 = \frac{h^2}{N^2}[N\sum_{i=1}^{n} f_i y_i^2 - (\sum f_i y_i)^2] = \frac{100}{2500}[50(105) - (-15)^2] = 201$

∴ Standard deviation = 14.18.

6. Find the standard deviation of the following data (use the step deviation method):

Wages (Rs.)	125 – 175	175 – 225	225 – 275	275 – 325	325 – 375	375 – 425	425 – 475	475 - 525	525 - 575
No. of workers	2	22	19	14	3	4	6	1	1

Sol: Length of the class interval h = 50, assumed mean = a = 300, $y_i = \frac{x_i - a}{h}$

Mid points of C.I (x_i)	**Frequency** (f_i)	y_i	y_i^2	$f_i y_i$	$f_i y_i^2$
150	2	-3	9	-6	18
200	22	-2	4	-44	88
250	19	-1	1	-19	19
300	14	0	0	0	0
350	3	1	1	3	3
400	4	2	4	8	16

450	6	3	9	18	54
500	1	4	16	4	16
550	1	5	25	5	25
	$\sum \boldsymbol{f_i = N} = \mathbf{72}$			$\sum \boldsymbol{f_i y_i} = \mathbf{-31}$	$\sum f_i \boldsymbol{y_i^2} = 239$

Mean $= \overline{x} = A + \frac{\sum f_i y_i}{N} \times h = 300 + \left(\frac{-31}{72} \times 50\right) = 300 - \frac{1550}{72} = 278.47$

Variance $\sigma^2 = \frac{h^2}{N^2}[N\sum_{i=1}^{n} f_i y_i^2 - (\sum f_i y_i)^2] = \frac{2500}{72^2}[72(239) - (31)^2]$

$= 2500\left[\frac{239}{72} - \left(\frac{31}{72}\right)^2\right]$

$\therefore$ Standard deviation $\sigma_x = 88.52$

7. Find the mean deviation from the mean and standard deviation of the series a, a+d, a+2d, ….,a+2nd.

Sol: Number of terms in the series = 2n + 1

$\overline{x} = A.M. = \frac{a+(a+d)+(a+2d)+\cdots.+(a+2nd)}{2n+1} = \frac{2n+1}{2}[2a + (2n+1-1)d] = a + nd$

Series is $a, a+d, a+2d, \ldots., a+(n-1)d, a+(n+1)d, \ldots..a+(2n-1)d, a+2nd.$

Mean deviation $= \frac{\sum|x_i - \overline{x}|}{2n+1} = \frac{nd+(n-1)d+(n-2)d+\cdots+d+0+d+\cdots..+(n-1)d+nd}{2n+1}$

$= \frac{2[d+\cdots..+(n-2)d+(n-1)d+nd+0+d+\cdots.+(n-1)d+nd]}{2n+1} = \frac{2d(1+2+\cdots.+n)}{2n+1} = \frac{2dn(n+1)}{2}.\frac{1}{2n+1} = \frac{n(n+1)}{2}d$

Let the assumed mean be A. Then

$\sigma^2 = \frac{1}{n}\sum(x_i - A)^2 - \left[\frac{\sum(x_i - A)}{n}\right]^2 = \frac{d^2+2^2d^2+\cdots.+(2n)^2d^2}{2n+1} - \left[\frac{d+2d+\cdots+2nd}{2n+1}\right]^2$

$= \frac{d^2}{2n+1}\frac{2n(2n+1)(4n+1)}{6} - d^2\left[\frac{2n(2n+1)}{2(2n+1)}\right]^2 = \frac{d^2n(4n+1)}{3} - d^2n^2 = nd^2\left(\frac{4n+1}{3} - n\right)$

$= \frac{nd^2(4n+1-3n)}{3} = \frac{n(n+1)}{3}d^2$

$\therefore \sigma = d\sqrt{\frac{n(n+1)}{3}}$

8. Given that $\overline{x}$ is the mean and σ^2 is the variance of n observations $x_1, x_2, \ldots..,x_n$. Prove that the mean and variance of the observations $ax_1, ax_2, \ldots..,ax_n$ are $a\overline{x}$ & $a^2\sigma^2 respectively\ (a \neq 0)$

Sol: We have $\overline{x} = \frac{x_1+x_2+\cdots.+x_n}{n}$

Mean of $ax_1, ax_2, \ldots.., ax_n = \frac{ax_1+ax_2+\cdots.+ax_n}{n} = \frac{a(x_1+x_2+\cdots.+x_n)}{n} = a\overline{x}$

Also we have $\sigma^2 = \frac{\sum(x_i-\overline{x})^2}{n}\ and$

$\therefore$ Variance of $ax_1, ax_2, \ldots.., ax_n = \frac{\sum(ax_i - a\overline{x})^2}{n} = a^2\frac{\sum(x_i-\overline{x})^2}{n} = a^2\sigma^2$

Hence the result.

9. The variance of 20 observations is 5. If each of the observations is multiplied by 2. Find the variance of the resulting observation.

Sol: Here a = 2, and $\sigma^2 = 5$

$\therefore$ Variance of the resulting observations $= a^2\sigma^2 = 2^2 \times 5 = 20,$

10. If each of the observations $x_1, x_2, \ldots, x_n$ is increased by k, k is a positive or negative number, then the variance remains unchanged.

Sol: Let $\overline{x}$ be the mean of $x_1, x_2, \ldots, x_n$

Variance $\sigma_1^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$

Let the new observations be $y_i = x_i + k \; where \; k \; is + ve \; or - ve \; number$

Then, mean of the new observations $= \overline{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum(x_i + k) = \frac{1}{n}[\sum x_i + \sum k]$

$= \frac{1}{n}\sum x_i + \frac{1}{n}\sum k = \overline{x} + \frac{1}{n}(kn) = \overline{x} + k$

The variance of the new observations $= \sigma_2^2 = \frac{1}{n}\sum(y_i - \overline{y})^2 = \frac{1}{n}\sum[(x_i + k) - (\overline{x} + k)]^2$

$= \frac{1}{n}\sum(x_i - \overline{x})^2 = \sigma_1^2$

NOTE:

Adding (or subtracting) a positive number to (or from) each of given set of observations do not effect the variance.

COEFFICIENT OF VARIATION:

A measure of dispersion is expressed in the same units as the variate in question. For example standard deviation of weights expressed in grams is also expressed in kilograms. So it becomes difficult to compare the variability of two distributions whose variates are expressed in different units. Hence, it becomes necessary to find out a relative measure of dispersion which is purely a number and independent of units of measurement. Coefficient of variation is one such relative measure.

Coefficient of variation (C.V.) is defined as the ratio of the standard deviation σ to the arithmetic mean $\overline{x}$ and it is often expressed as a percentage.

$\therefore$ Coefficient of variation $= \frac{\sigma}{\overline{x}} \times 100 \; where \; \overline{x} \neq 0$

NOTE:

The distribution having greater coefficient of variation is said to be more variable than the other. The distribution having lesser coefficient of variation is said to be more consistent than the other.

ANALYSIS / COMPARISON OF TWO FREQUENCY DISTRIBUTIONS WITH EQUAL MEANS:

Suppose two distributions are having same mean $\overline{x_1} = \overline{x_2} = \overline{x}$ but different standard deviations $\sigma_1 \; and \; \sigma_2$ respectively. Then coefficient of variations are given by $\left(\frac{\sigma_1}{\overline{x}} \times 100\right)$ & $\left(\frac{\sigma_2}{\overline{x}} \times 100\right)$. Thus the C.V.s can be compared using $\sigma_1 \; and \; \sigma_2 \; only$. Here, the series with lower value of standard deviation is said to be more consistent than the other series with greater standard deviation. The series with greater standard deviation is called more dispersed than other.

SOLVED EXAMPLES:

1. Students of two sections A and B of a class show the following performance in a test (for 100 marks). Which section of students has greater variability in performance.

	Section A	**Section B**
No. of students	50	60
Average marks in test	45	45
Variance of distributions of marks	64	81

Sol: Given variances are 64 and 81.
$\therefore$ Standard deviations are 8 and 9 i.e., $\sigma_1 = 8 \; and \; \sigma_2 = 9$
They are having same mean. Thus $\overline{x_1} = \overline{x_2} = \overline{x} = 45$
Hence section B has greater variability in the performance.

2. Goals scored by two teams A and B in football season are as follows:

Number of goals scored in match	**Number of matches**	
	Team A	**Team B**
0	24	25
1	9	9
2	8	6
3	5	5
4	4	5

By calculating the standard deviation in each case find which team be consider more consistent.

Sol:

Team A					**Team B**				
x_i	f_i	$d_i = x_i - a = x_i - 2$	$f_i d_i$	$f_i d_i^2$	x_i	f_i	$d_i = x_i - a = x_i - 2$	$f_i d_i$	$f_i d_i^2$
0	24	-2	-48	96	0	25	-2	-50	100
1	9	-1	-9	9	1	9	-1	-9	9
2	8	0	0	0	2	6	0	0	6
3	5	1	5	5	3	5	1	5	5
4	4	2	8	16	4	5	2	10	20
	50		**-44**	**126**		**50**		**-44**	**140**
Mean = $\overline{x} = a + \frac{f_i d_i}{f_i} = 2 - \frac{44}{50} = 1.12$ and $\sigma = \sqrt{\frac{1}{N}\sum f_i d_i^2 - \left(\frac{\sum f_i d_i}{N}\right)^2}$ $= \sqrt{\frac{126}{50} - \left(-\frac{44}{50}\right)^2} = \sqrt{2.52 - 0.7774}$ $= 1.32$					Mean = $\overline{x} = a + \frac{f_i d_i}{f_i} = 2 - \frac{44}{50} = 1.12$ and $\sigma = \sqrt{\frac{1}{N}\sum f_i d_i^2 - \left(\frac{\sum f_i d_i}{N}\right)^2}$ $= \sqrt{2.8 - 0.7744} = 1.42$				

Here, we find means are equal but S.D. of team A < S.D of team B. Hence team A is more consistent.

3. Lives of two models of refrigerators A and B are given below, Which refrigerator model would you suggest to purchase?

Life in years	Model A	Model B
0 – 2	5	2
2 – 4	16	7
4 – 6	13	12
6 – 8	7	19
8 – 10	5	9

Sol:

Class Interval	Mid point		Model A			Model B		
	x_i	x_i^2	f_i	$f_i x_i$	$f_i x_i^2$	f_i	$f_i x_i$	$f_i x_i^2$
0 – 2	1	1	5	5	5	2	2	2
2 – 4	3	9	16	48	144	7	21	63
4 – 6	5	25	13	65	325	12	60	300
6 – 8	7	49	7	49	343	19	133	931
8 – 10	9	81	5	45	405	9	81	729
			46	**212**	**1221**	**49**	**297**	**2025**

For model A, $\overline{x_A} = \frac{\sum f_i x_i}{\sum f_i} = \frac{212}{46} = 4.6 \; and \; \sigma_A = \sqrt{\frac{f_i x_i^2}{N} - \left(\frac{f_i x_i}{N}\right)^2} = \sqrt{5.38} = 2.319$

For model B, $\overline{x_B} = \frac{\sum f_i x_i}{\sum f_i} = \frac{297}{49} = 6.06 \; and \; \sigma_B = \sqrt{\frac{2025}{49} - \left(\frac{297}{49}\right)^2} = \sqrt{4.61} = 2.147$

Coefficient of variation for model A $= \frac{\sigma_A \times 100}{\overline{x_A}} = \frac{2.319}{4.6} \times 100 = 50.41$

Coefficient of variation for model B $= \frac{\sigma_B \times 100}{\overline{x_B}} = \frac{2.147}{6.06} \times 100 = 35.43$

Hence the model A is suggested to purchase.

SKEWNESS:

The measures of Central Tendency and Dispersion do not indicate whether the distribution is symmetric or not. Measures of skewness gives the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness. Thus skewness is the lack of symmetry. The measures of central tendency and dispersion are inadequate to characterize a distribution completely. They may be supported by two more measures Skewness and Kurtosis.

A distribution which is not symmetrical is called a skewed distribution. In such distributions the mean, the mode and median will not coincide. The values are pulled apart.

TEST OF SKEWNESS:

The absence of asymmetry or skewness can be stated under the following conditions.

If the distribution is symmetric, the following conditions are observed.

1) The values of mean, mode, median coincide (the values are equal(

2) Q_3 – Median = Median – Q_1

3) The sum of positive deviations = The sum of negative deviations.

4) The frequencies on either side of the mode are equal.

Similarly, a skewed distribution will have following characteristics.

1) Mean ≠ Median ≠ Mode

2) Q_3 – Median ≠ Median – Q_3

3) The sum of positive deviations ≠ The sum of negative deviations.

MEASURES OF SKEWNESS:

Absolute skewness = Mean – Mode = ('+' positive skewness) = ('-' negative skewness)

If the value of Mean is greater than Mode, then the skewness is positive.

If the value of Mode is greater than Mean, the skewness is negative.

The absolute measure of skewness will not be proper measure for comparison. Hence, in each series a relative measure of coefficient of skewness will have to be computed.

There are three important measures of relative skewness.

1. Karl Pearson's coefficient of skewness

2. Bowley's coefficient of skewness

3. Kelly's coefficient of skewness

Generally Karl Pearson method is widely used.

Karl Pearson coefficient of skewness $\left(S_{k_p}\right) = \frac{\overline{X}-mode}{\sigma}$

In case mode is illdefined.

The coefficient of skewness $\left(S_{k_p}\right) = \frac{3(Mean-Median)}{\sigma} = \frac{3(\overline{X}-M)}{\sigma}$

SOLVED EXAMPLES:

1. Calculate Karl Pearson's coefficient of skewness for the following data:

   25, 15, 23, 40, 27, 25, 23, 25, 20.`

Sol: Computation table of Mean and Standard Deviation

Size	Deviation from A = 25 (d) = size – A	Square of deviation (d^2)
25	0	0
15	-10	100
23	-2	4

40	15	225
27	2	4
25	0	0
23	-2	4
25	0	0
20	-5	25
	$\sum d = -2$	$\sum d^2 = 362$

Here $n = 9, Mean = A \pm \frac{\sum d}{n} = 25 - \frac{2}{9} = 24.78,\ Mode = 25\ and$

$S.D. = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\frac{362}{9} - \left(\frac{-2}{9}\right)^2} = 6.3$

Karl Pearson coefficient of skewness $= \frac{Mean - Mode}{S.D.} = \frac{24.78 - 25}{6.3} = -0.03$

2. Calculate Karl Pearson's coefficient of skewness for the following data:

Variable	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40
Frequency	2	5	7	13	21	16	8	3

Sol: Computation of mean and standard deviation. Take A = 22.5.

Variable X	**Mid value (m)**	**Frequency (f)**	**Deviation** $d' = \frac{(m - 22.5)}{5}$	fd'	d'^2	fd'^2
0 – 5	2.5	2	-4	-8	16	32
5 – 10	7.5	5	-3	-15	9	45
10 – 15	12.5	7	-2	-14	4	28
15 – 20 l	17.5	13 f_1	-1	-13	1	13
20 – 25	22.5	21 f	0	0	0	0
25 – 30	27.5	16 f_2	1	16	1	16
30 – 35	32.5	8	2	16	4	32
35 – 40	37.5	3	3	9	9	27
		$\sum f = N = 75$		$\sum fd' = -9$		$\sum fd'^2 = 193$

Here c = class interval = 5

Mean $\overline{X} = A \pm \frac{\sum fd'}{N} \times c = 22.5 + \frac{-9}{75} \times 5 = 21.9$

S.D. $= \sigma = \left(\sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2}\right) \times c = \left(\sqrt{\frac{193}{75} - \left(\frac{-9}{75}\right)^2}\right) \times 5 = \sqrt{2.573 - 0.014} \times 5$

$= \sqrt{2.558} \times 5 = 1.599 \times 5 = 7.9950 = 8$

Mode $= l + \frac{f-f_1}{2f-f_1-f_2} \times c = 20 + \frac{21-13}{2\times21-13-16} \times 5 = 20 + \frac{40}{13} = 23.08$

$\therefore$ Pearson's coefficient of skewness $= \frac{Mean-Mode}{S.D} = \frac{21.9-23.08}{8} = -\frac{1.18}{8} = -0.148$

KURTOSIS:

The expression Kurtosis is used to describe the peakedness of curve. As far as the measurement of a shape is concerned, we have two characteristics. Skewness which refers to asymmetry of a series and Kurtosis which measures the peakedness of a normal curve. All the frequency curves expose different degrees of flatness or peakedness.

This characteristic of frequency curve is termed as Kurtosis. Measures of Kurtosis denote the shape of the top of a frequency curve.

MEASURES OF KURTOSIS:

The measures of Kurtosis of a frequency distribution are based upon the fourth moment about the mean of the distribution.

Symmetrically, $\beta_2 = \frac{\mu_4}{\mu_2^2}\ where\ \mu_4 = fourth\ moment, \mu_2 = second\ moment$

If $\beta_2 = 3$, the distribution is said to be normal (neither flat nor peaked) and the curve is **normal curve (mesokurtic).**

If $\beta_2 > 3$, the distribution is said to be more peaked and the curve is **lepokurtic.**

If $\beta_2 < 3$, the distribution is said to be flatter than normal curve and the curve is **platykurtic.**

SOLVED EXAMPLES:

1. From the following distribution, calculate

i) First 4 moments about the mean ii) Skewness based on moments iii) Kurtosis

Income (Rs)	0 – 10	10 – 20	20 – 30	30 – 40
Frequency	1	3	4	2

Sol: Computation of Moments, Skewness and Kurtosis

Income	**Mid value (m)**	**Frequency** (f)	**Deviation** $d = \frac{(m-15)}{10}$	fd	fd^2	fd^3	fd^4
0 – 10	5	1	-1	-1	1	-1	1
10 – 20	15	3	0	0	0	0	0
20 – 30	25	4	1	4	4	4	4
30 – 40	35	2	2	8	8	16	32
	N = 10			$\sum fd = 7$	$\sum fd^2 = 13$	$\sum fd^3 = 19$	$\sum fd^4 = 37$

i) Moments about mean

$\mu_1' = \frac{\sum fd}{N} \times c = \frac{7}{10} \times 10 = 7$

$\mu_2' = \frac{\sum fd^2}{N} \times c^2 = \frac{13}{10} \times 10^2 = 130$

$\mu_3' = \frac{\sum fd^3}{N} \times c^3 = \frac{19}{10} \times 10^3 = 1900$

$\mu_4' = \frac{\sum fd^4}{N} \times c^4 = \frac{37}{10} \times 10^4 = 37000$

First moment about mean, $\mu_1 = \mu_1' - \mu_1' = 7 - 7 = 0$

Second moment about mean, $\mu_2 = \mu_2' - (\mu_2')^2 = 130 - 7^2 = 81$

Third moment about mean, $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1^3 = 1900 - 3(130)(7) + 2(7)^3 = -144$

Fourth moment about mean, $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_2^2 - 3\mu_1^4 = 37000 - 53200 + 38200 - 7203 = 14817$

ii) Skewness based on moments is studied by β_1

$\therefore \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-144)^2}{(81)^3} = 0.039$

iii) Kurtosis is studied by β_2

$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{14817}{(81)^2} = 2.26$

Since $\beta_2 < 3$, the curve is more peaked and is platykurtic.

CORRELATION:

* Correlation refers relation between two or more variables.
* Correlation is a statistical analysis which measures and analyzes two variables, how they fluctuate with reference to each other.

TYPES OF CORRELATION:

* If two variables tend to move in same direction is called positive or direct correlation.
* If two variables tend to move in opposite direction is called negative or inverse correlation.
* Study of only two variables, the relationship is described is called simple correlation.
* Study of more than two variables simultaneously is called multiple correlation.
* Study of two variables excluding some other variables is called partial correlation. (price, demand, eliminating supply). Study of all the variables is called total correlation.
* The ratio of change between two variables is uniform then there will be linear correlation between them otherwise non – linear correlation.

METHOD OF STUDYING CORRELATION:

1. GRAPHIC METHOD

a) Scatter diagram (or) Scatergram

b) Simple graph

2. MATHEMATICAL METHOD

a) Karl Pearson's Coefficient of Correlation

b) Spearman's Rank Coefficient of Correlation

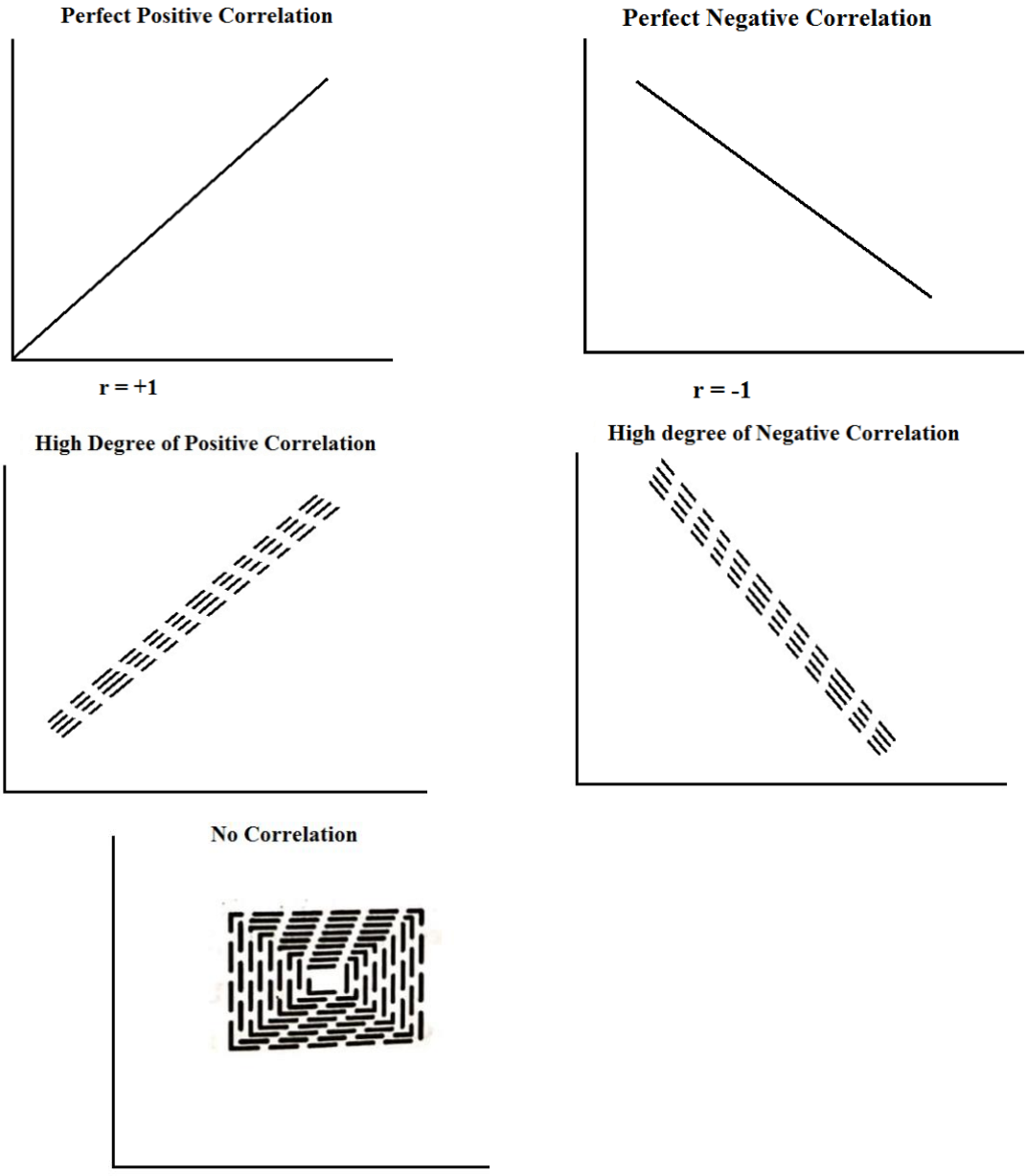
c) Coefficient of Concurrent deviation

d) Method of Least Squares

1. GRAPHIC METHOD

A) SCATER DIAGRAMS:

Scater diagram is a chart obtained by plotting two variables to find out whether there is any relationship between them. In this diagram X variables are plotted on horizontal axis and Y variables are plotted on vertical axis.



* Scater diagram is simple attractive method to find out nature of correlation.

* It is easy to understand.

* A rough idea is got at a glance whether it is +'ve or –'ve .

B) SIMPLE GRAPH:

In this method two variables are plotted on a graph paper. We get two curves. By comparing the curves we can decide relation between the variables.

2. MATHEMATICAL METHOD:

In this method, basing on value of correlation coefficient we can decide the relation between variables.

A) COEFFICIENT OF CORRELATION:

KARL PEARSON'S COEFFICIENT OF CORRELATION:

Karl Pearson, a British statistician suggested a mathematical method for measuring the magnitude of linear relationship between the variables. This is known as Pearsonian coefficient of correlation. This is denoted by r. There are several formulas for r.

$(1)\ r = \frac{co\ variance\ of\ xy}{\sigma_x \sigma_y}$ $(2)\ r = \frac{\sum xy}{N\sigma_x\sigma_y}$ $(3)\ r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$

$where\ X = (x - \bar{X})\quad Y = (y - \bar{Y}),\ \bar{X}, \bar{Y}$ are the means of series x and y.

$\sigma_x = S.D.of\ series\ x$

$\sigma_y = S.D.of\ series\ y$

PROPERTIES OF CORRELATION COEFFICIENT:

1. Limits of correlation coefficient are $-1 \le r \le 1$
2. If r = 1, correlation is perfect and +ve.

   If r = -1, correlation is perfect and –ve.

   If r = 0, there is no relationship between variables.
3. Two independent variables are un correlated i.e., x and y are independent then r(xy) = 0

PROBLEMS:

1. Find if there is any significant correlation between the heights and weights given below:

Height in inches	57	59	62	63	64	65	55	58	57
Weight	113	117	126	126	130	129	111	116	112

Sol: Coefficient of correlation $r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$

$\bar{X} = \frac{540}{9} = 60;\quad \bar{Y} = \frac{1080}{9} = 120$

Height in inches (x)	**Deviation from mean $X = (x - \bar{X})$**	**X^2**	**Weights (y)**	**Deviation from mean $Y = y - \bar{Y}$**	**Y^2**	**XY**
57	-3	9	113	-7	49	21
59	-1	1	117	-3	9	3
62	2	4	126	6	36	12
63	3	9	126	6	36	18

64	4	16	130	10	100	40
65	5	25	129	9	81	45
55	-5	25	111	-9	81	45
58	-2	4	116	-4	16	8
57	-3	9	112	-8	64	24
540	**0**	**102**	**1080**	**0**	**472**	**216**

$$r = \frac{216}{\sqrt{102 \times 472}} = 0.98$$

2. Find Karl Pearson's coefficient of correlation from the following data:

Wages	100	101	102	102	100	99	97	98	96	95
Cost of living	98	99	99	97	95	92	95	94	90	91

Sol: $\bar{X} = \frac{990}{10} = 99; \quad \bar{Y} = \frac{950}{10} = 95$

Wages (x)	**Deviation from mean $X = (x - \bar{X})$**	**X^2**	**Cost of living (y)**	**Deviation from mean $Y = y - \bar{Y}$**	**Y^2**	**XY**
100	1	1	98	3	9	3
101	2	4	99	4	16	8
102	3	9	99	4	16	12
102	3	9	97	2	4	6
100	1	1	95	0	0	0
99	0	0	92	-3	9	0
97	-2	4	95	0	0	0
98	-1	1	94	-1	1	1
96	-3	9	90	-5	25	15
95	-4	16	91	-4	16	16
990	**0**	**54**	**950**	**0**	**92**	**61**

$$r = \frac{61}{\sqrt{54 \times 92}} = 0.847$$

PRACTICE PROBLEM:

3. Calculate the coefficient of correlation between age of cars and annual maintenance cost and comment.

Age of cars (years)	2	4	6	7	8	10	12
Annual maintenance cost	1600	1500	1800	1900	1700	2100	2000

Hint: $\bar{X} = \frac{49}{7} = 7; \quad \bar{Y} = \frac{12600}{7} = 1800$

$$r = \frac{\sum XY - \frac{\sum XY}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}} = \frac{\sum XY \times N - \frac{\sum XY}{N}}{\sqrt{[\sum X^2 N - (\sum X)^2][\sum Y^2 N - (\sum Y)^2]}}$$

4. With the following data in 6 cities, calculate the coefficient of correlation by Pearson's method between the density of population and death rate.

Cities	Area in Sq.km	Population (000)	No. of deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

Sol: $Density\ of\ population = \frac{Population}{Area}$; $Death\ rate = \frac{No.of\ deaths}{Population}$

Density (x)	200	500	400	700	600	300
Death rate (y)	10	16	14	20	17	13

$\bar{X} = \frac{2700}{6} = 450; \quad \bar{Y} = \frac{90}{6} = 15$

(x)	$X = (x - \bar{X})$	X^2	(y)	$Y = y - \bar{Y}$	Y^2	XY
200	-250	62500	10	-5	25	1250
500	50	2500	16	1	1	50
400	-50	2500	14	-1	1	50
700	250	62500	20	5	25	1250
600	150	22500	17	2	4	300
300	-150	22500	13	-2	4	300
2700	**0**	**175000**	**90**	**0**	**60**	**3200**

$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} = \frac{3200}{\sqrt{175000 \times 60}} = 0.988$

PRACTICE PROBLEM:

5. Calculate coefficient of correlation for the following data.

X	12	9	8	10	11	13	17
Y	14	8	6	9	11	12	3

$r = \frac{\sum XY \times N - \sum X \sum Y}{\sqrt{[\sum X^2 N - (\sum X)^2][\sum Y^2 N - (\sum Y)^2]}} = 0.95$

WHEN DEVIATION ARE TAKEN FROM AN ASSUMED MEAN:

When actual mean is not a whole number but a fraction or when the series is large the calculation by direct method will involve a lot of time. To avoid such tedious calculation we can use assumed mean method.

Formula $r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$ $(or)\ r = \frac{\sum XY \times N - \sum X \sum Y}{\sqrt{[\sum X^2 N - (\sum X)^2][\sum Y^2 N - (\sum Y)^2]}}$

1. Calculate Karl Pearson's correlation coefficient for the period

X	28	41	40	38	35	33	40	32	36	33
Y	23	34	33	34	30	26	28	31	36	38

$\bar{X} = \frac{355}{10} = 35.5 = 35; \quad \bar{Y} = \frac{313}{10} = 31.3 = 31$

(x)	$X = (x - \bar{X})$	X^2	**(y)**	$Y = y - \bar{Y}$	Y^2	**XY**
28	-7	49	23	-8	64	56
41	6	36	34	3	9	18
40	5	25	33	2	4	10
38	3	9	34	3	9	9
35	0	0	30	1	1	0
33	-2	4	26	-5	25	10
40	5	25	28	-3	9	-15
32	-3	9	31	0	0	0
36	1	1	36	5	25	5
33	-2	4	38	7	49	-14
355	**6**	**162**		**3**	**195**	**79**

$$r = \frac{\sum XY \times N - \sum X \sum Y}{\sqrt{[\sum X^2 N - (\sum X)^2][\sum Y^2 N - (\sum Y)^2]}} = \frac{79 \times 10 - 6 \times 3}{\sqrt{[162 \times 10 - 36][195 \times 10 - 9]}} = \frac{772}{\sqrt{1584 \times 1941}} = 0.45$$

PRACTICE PROBLEMS:

2. Find suitable coefficient of correlation for the following data:

Fertilizers used	15	18	20	24	30	35	40	50
Productivity	85	93	95	105	130	130	150	160

$\bar{X} = \frac{232}{8} = 29; \quad \bar{Y} = \frac{938}{8} = 119 \; r = 0.99$

3. Find out coefficient of correlation in the following case:

Height of father in inches	65	66	67	67	68	69	71	73
Height of son in inches	67	68	64	68	72	70	69	70

$r = 0.472$

4. Find the correlation coefficient for the following data:

X	65	66	67	68	69	70	72
Y	67	68	65	72	72	69	71

$r = 0.603$

5.

X	1	2	3	4	5	6	7	8	9
Y	12	11	13	15	14	17	16	19	18

$r = 0.933$

6.

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

$r = 0.95$

RANK CORRELATION COEFFICIENT:

A British Psychologist Charles Adward Spearman found out the method of finding the coefficient of correlation by ranks. This method is based on rank and is useful in dealing with qualitative characteristics such as Morality, Character, Intelligence and Beauty. It cannot be measured quantitatively as in the case of Pearson's coefficient correlation. It is based on the ranks given to the observations. Rank correlation is applicable only to the individual observations. The formula for Spearman's Rank Correlation is given by

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

Where $\rho = Rank\ coefficient\ of\ correlation$

$D^2 = Sum\ of\ squares\ of\ differences\ of\ two\ ranks$

$N = Number\ of\ paired\ observation$

PROPERTIES OF RANK CORRELATION COEFFICIENT:

1. The value of q lies between -1 and 1 i.e., $-1 \leq \rho \leq 1$
2. If $\rho = 1$ there is complete agreement in the order if the ranks and the direction of the rank is same.
3. If $\rho = -1$ then there is complete disagreement in the order of the ranks and they are in opposite directions.

PROCEDURE TO SOLVE PROBLEMS:

1. When the ranks are given

   Step 1: Compute the differences of two ranks and denote it by D.

   Step 2: Square D and get D^2.

   Step 3: Obtain ρ by substituting figures in formula.

2. When the ranks are not given but actual data are given then we must give ranks. We can give ranks by taking the highest as 1 or the lowest value as 1 next to the highest (lowest) as 2 and follow same procedure for both the variables.

PROBLEMS:

1. Following are ranks obtained by 10 students in two subjects, Statistics and Mathematics. To what extent the knowledge of the students in two subjects is related.

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

(x)	(y)	D = x – y	D^2
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			40

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6\times 40}{10(10^2-1)} = 1 - \frac{240}{990} = 1 - 0.3 = 0.7$$

2. A random sample of 5 college students is selected and their grades in mathematics and statistics are found to be

Mathematics	85	60	73	40	90
Statistics	93	75	65	50	80

Calculate Spearman's rank coefficient

Sol:

Marks in Maths (X)	Rank (x)	Marks in Stats (Y)	Rank (y)	Rank Difference D = x – y	D^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	-1	1
40	5	50	5	0	0
90	1	80	2	-1	1
					4

$N = 5, \sum D^2 = 4$

Pearman's Rank Correlation

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6\times 4}{5(5^2-1)} = 1 - \frac{24}{120} = 1 - 0.2 = 0.8$$

3. Ten competitors in a musical test were ranked by their judges A, B and C in the following order.

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common liking in music.

Sol: Here N = 10

Ranks by A (X)	Ranks by B (Y)	Ranks by C (Z)	Rank Difference $D_1 = X - Y$	Rank Difference $D_2 = X - Z$	Rank Difference $D_3 = Y - Z$	D_1^2	D_2^2	D_3^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
						200	**60**	**214**

$$\rho_1(X,Y) = 1 - \frac{6\sum D_1^2}{N(N^2-1)} = 1 - \frac{6\times 200}{10\times 99} = -\frac{7}{33}$$

$$\rho_2(X,Z) = 1 - \frac{6\sum D_2^2}{N(N^2-1)} = \frac{7}{11}$$

$$\rho_3(Y,Z) = 1 - \frac{6\sum D_3^2}{N(N^2-1)} = -\frac{49}{165}$$

Since $\rho_2(X,Z)$ is maximum we conclude that the pair of judges A and C has the nearest approach to common likings in music.

EQUAL OR REPEATED RANKS:

If any two or more persons are equal in any classification or if there is more than one item with same value in the series then the Spearman's formula for calculating the rank correlation coefficient break down. In this case common ranks are given to repeated items. The common rank is the average of ranks which there items would have assumed, if they were different from each other and the next item will get the rank next to ranks already assumed.

For example: If two individuals are placed in 7th place each of them are given the rank 7.5 next rank will be 9. Similarly if 3 are ranked equal at the 7th place then they are given the rank $\frac{7+8+9}{3} = 8$ which is common rank assigned to each, and the next rank will be 10.

In this case, $\rho = 1 - 6\left[\frac{\sum D^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m)....}{N^3-N}\right]$

Where $m = the\ number\ of\ items\ whose\ ranks\ are\ common.$

1. From the following data calculate the rank correlation coefficient after making adjustment for tied ranks.

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

Sol: First we have to assign ranks to the variables.

(X)	**Rank (x)**	**(Y)**	**Rank (y)**	**Rank Difference D = x – y**	**D^2**
48		13			
33		13			
40		24			
9		6			
16		15			
16		4			
65		20			
24		9			
16		6			
57		19			

16 is repeated 3 times in X items hence m = 3.

Since 13 and 6 are repeated twice in Y items m = 2

$$\rho = 1 - 6\left[\frac{\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m).}{N^3 - N}\right] = 0.7332.$$

Obtain rank correlation coefficient for the following data:

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

$\rho = 0.545$

REGRESSION:

* A statistical method which helps to estimate the unknown value of one variable from the known value of related variable is called regression.

* The line described in the average relationship between two variables is known as line of regression.

USES:

1. It is used to estimate the relation between two economic variables like income expenditure.
2. It is highly valuable tool in economic and business.
3. Widely used for prediction purpose
4. We can calculate coefficient of correlation and coefficient of determination with the help of regression coefficient.
5. It is useful in statistical estimation of demand curves, supply curves, production function, cost function consumption function etc.

COMPARISON BETWEEN CORRELATION AND REGRESSION:

The correlation coefficient is a measure of comparability between two variables, while the regression establishes a functional relation between dependent and independent variables. In correlation both values x and y are random variable whereas in regression x is random variable and y is fixed variable. The coefficient of correlation is relative measure whereas regression coefficient is an absolute figure.

METHOD OF STUDYING REGRESSION:

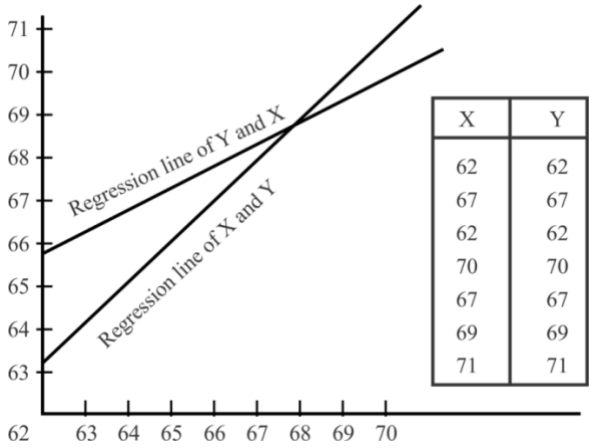
There are two methods to study regression.

1. Graphic method 2. Algebraic method

1. GRAPHIC METHOD:

In this method, the points representing the pair of values of the variables are plotted on a graph. These points form a scatter diagram. A regression line is drawn between these points by free hand.

FIT A REGRESSION LINE ON THE SCATER DIAGRAM FOR THE FOLLOWING DATA:



X	Y
62	62
67	67
62	62
70	70
67	67
69	69
71	71

ALGEBRAIC METHOD:

REGRESSION LINE:

A regression line is a straight line fitted to data by the method of least squares. It indicates best possible mean value of one variable corresponding to mean value of the other. These are always two regression lines constructed for the relationship between two variables X and Y.

REGRESSION EQUATION:

Regression equation is an algebraic expression of the regression line.

The standard form of regression equation is $Y = a + bX,\ a, b\ are\ constants$. 'a' indicates value of Y when X = 0. It is called Y – intercept. 'b' indicates value of slope of regression line. It is also called regression coefficient Y on X. If we know the value of a and b we can easily compute value of Y for given value of X. The values of a, b are found with help of normal equation.

For $Y = a + bX$ $(Regression\ equation\ of\ Y\ on\ X)$

Normal equations

$\sum Y = Na + b\sum X$

$\sum XY = a\sum X + b\sum X^2$

For $X = a + bY$ $(Regression\ equation\ X\ on\ Y)$

$\sum X = Na + b\sum Y$

$\sum XY = a\sum Y + b\sum Y^2$

Determine equation of a straight line which best fits the data:

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

$a = 0.82,\ b = 1.56$

DEVIATION TAKEN FROM ARITHMETIC MEAN OF X AND Y:

This method is easier and simpler than the previous method to find values of a and b.

We can find out deviation of X and Y series from their respective means.

Regression equation of X on Y

$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$

The regression coefficient of X on Y $= r\frac{\sigma_x}{\sigma_y} = \frac{\sum XY}{\sum Y^2} = b_{XY}$

The regression coefficient of Y on X $= r\frac{\sigma_y}{\sigma_x} \Rightarrow\ r^2 = b_{XY}.b_{YX}$

REGRESSION:

EXAMPLE PROBLEMS:

1. Determine the equation of a straight line which best fits the data or find regression line of y on x.

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

Sol: Let the required straight lines Y = a + bX

The two normal equations are

$\sum Y = b\sum X + Na$

$\sum XY = b\sum X^2 + a\sum X$

$\boldsymbol{X}$	$\boldsymbol{X^2}$	$\boldsymbol{Y}$	$\boldsymbol{XY}$
10	100	10	100
12	144	22	264
13	169	24	312
16	256	27	432
17	289	29	493
20	400	33	660
25	625	37	925

$\sum X = 113$	$\sum X^2 = 1938$	$\sum Y = 182$	$\sum XY = 3186$

Substituting the values

$\sum Y = b\sum X + Na$

$\sum Y = 182;\ \sum X = 113; N = 7$

$113b + 7a = 182$ ……(1)

$\sum XY = 3186;\ \sum X^2 = 1983;\ \sum X = 113$

$1983b + 113a = 3186$ …….(2)

Multiplying (1) by 113;

$12769b + 791a = 20566$ …….(3)

Multiplying (2) by 7;

$13881b + 791a = 22302$ ……. (4)

Subtracting (4) from (3); $b = \frac{1736}{1112} = 1.56 \Rightarrow a = 0.82$

The equation of straight line is

$Y = a + bX$

$a = 0.82;\ \ b = 1.56$

$Y = 0.82 + 1.56X$

∴The equation of the required straight line is $Y = 0.82 + 1.56X$

This is called regression equation of Y on X.

2. Calculate the regression equations of Y on X from the data given below, taking deviations from actual means of X and Y.

Price (Rs.)	10	12	13	12	16	15
Amount demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs. 20.

Sol: Calculation of Regression equation.

x	$(x - 13) = X$	X^2	y	$(y - 41) = Y$	Y^2	XY
10	-3	9	40	-1	1	3
12	-1	1	38	-3	9	3
13	0	0	43	2	4	0
12	-1	1	45	4	16	-4
16	3	9	37	-4	16	-12
15	2	4	43	2	4	4

Regression equation of Y on X is

$Y - \bar{Y} = r\frac{\sigma y}{\sigma x}(X - \bar{X})$

$r\frac{\sigma y}{\sigma x} = \frac{\sum XY}{\sum X^2} = -0.25$

$Y - 41 = -0.25(X - 13) \Rightarrow Y = -0.25X + 44.25$

When X is 20, Y = 39.25

When the price is Rs 20, the likely demand is 39.25.

We have $r = \sqrt{b_{yx} \times b_{xy}}$

$\therefore r = \sqrt{\frac{-a_1}{b_1} \times \frac{-b_2}{a_2}} = \sqrt{\frac{a_1 b_2}{a_2 b_1}} < 1 \Rightarrow a_1 b_2 < a_2 b_1$

3. For the following data, find equations of the two regression lines.

X	1	2	3	4	5
Y	15	25	35	45	55

Sol: Calculation of regression coefficients

X	Y	$x = X - \bar{X}$	x^2	$y = Y - \bar{Y}$	Y^2	xy
1	15	-2	4	-20	400	40
2	25	-1	1	-10	100	10
3	35	0	0	0	0	0
4	45	1	1	10	100	10
5	55	2	4	20	400	40
15	175		10		1000	100

We have $\bar{X} = \frac{\sum X}{5} = \frac{15}{5} = 3 \; and \; \bar{Y} = \frac{\sum Y}{5} = \frac{175}{5} = 35$

Now $b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{100}{10} = 10$

∴ Regression equation of y on x is given by,

$y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y - 35 = 10(x - 3)$

Now $b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{100}{1000} = \frac{1}{10}$

∴ Regression equation of x on y is given by

$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 3 = \frac{1}{10}(y - 35)$

4. In the following table S is weight of Potassium bromide which will dissolve in 100 gms of water at V°C. Fit an equation of the form S = mT + b by the method of least squares. Use this relation to estimate S when T = 50°.

T	0	20	40	60	80
S	54	65	75	85	96

Sol:

T	$d_T = \frac{T - 40}{10}$	d_T^2	S	$d_S = \frac{S - 75}{10}$	d_S^2	$d_T d_S$
0	-4	16	54	-2.1	4.41	8.4
20	-2	4	65	-1.0	1.00	2.0
40	0	0	75	0	0.00	3.0
60	+2	4	85	+1.0	1.00	2.0
80	+4	16	96	+2.1	4.41	8.4
200	0	40	375	0	10.82	20.8

Now, $m = \frac{\sum d_T d_S}{\sum d_{T^2}} = \frac{20.8}{40} = 0.52$

And b is given by the equation $\sum S = m \sum T + Nb$

$\therefore 375 = (0.52)(200) + 5b \Rightarrow b = 54.2$

When $T = 50^o C, S = 0.52 \times 50 + 54.2 = 80.2$

5. A panel of two judges P and Q graded seven dramatic performances by independently awarding marks as follows:

Performance	1	2	3	4	5	6	7
Marks by P	46	42	44	40	43	41	45
Marks by Q	40	38	36	35	39	37	41

The eight performance, which judge Q would not attend, was awarded 37 marks by judge P. If judge Q had also been present, how many marks would be expected to have been awarded by him to the eight performance.

Sol: Calculation of Regression coefficients.

Performance	**X (Marks by P)**	$x = X - \bar{X}$	x^2	**Y (Marks by Q)**	$y = Y - \bar{Y}$	xy
1	46	+3	9	40	+2	+6
2	42	-1	1	38	0	0
3	44	+1	1	36	-2	2
4	40	-3	1	35	-3	+9
5	43	0	0	39	+1	0
6	41	-2	4	37	-1	+2
7	45	+2	4	41	+3	+6
	301		28	266		21

$\bar{X} = \frac{301}{7} = 43;\ \bar{Y} = \frac{266}{7} = 38$

Regression equation of Y on X is given by

$Y - \bar{Y} = b_{yx}(X - \bar{X})$ …… (1)

Now $b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{21}{28} = 0.75$

Substituting the values in equation (1), we get

$Y - 38 = 0.75(X - 43) \Rightarrow Y = 0.75X + 5.75$

If $X = 37, then\ Y = 0.75(37) + 5.75 = 33.5$

Hence if the judge Q would have been present, he would have awarded 33.5 marks to the eight performance.

6. Find the most likely production corresponding to a rainfall 40 from the following data.

	Rainfall (X)	Production (Y)
Average	30	500 Kgs
Standard deviation	5	100 Kgs
Coefficient of correlation	0.8	

Sol: We have to calculate the value of Y when X = 40.

So we have to find the regression equation of Y on X

Mean of X series, $\bar{X} = 30$; Mean of Y series, $\bar{Y} = 500$

$\sigma\ of\ X\ series, \sigma_x = 5; \quad \sigma\ of\ Y\ series, \sigma_y = 100$

Regression of Y on X is

$Y - \bar{Y} = r.\frac{\sigma_x}{\sigma_y}(X - \bar{X}) \Rightarrow Y - 500 = (0.8)\frac{5}{100}(X - 30)$

When $X = 40, Y - 500 = \frac{4}{100}(40 - 30) = \frac{40}{100} \Rightarrow Y = 500 + \frac{4}{10} = 500.4$

Hence the expected value of Y is 500.4 kg.

7. From a sample of 200 pairs of observation the following quantities were calculated.

$\sum X = 11.34, \sum Y = 20.78,\ \sum X^2 = 12.16, \sum Y^2 = 84.96,\ \sum XY = 22.13$

From the above data show how to compute the coefficients of the equation Y = a + bX

Sol: We can compute the coefficients of the equation Y = a + bX by solving the normal equations:

$\sum Y = na + b\sum X \ \ and\ \sum XY = a\sum X + a\sum X^2$

Substituting the values

20.72 = 200a + 11.34b $\Rightarrow a = \frac{20.72 - 11.34b}{200}$ and =0.1036-0.0567b...(1) =0.1036-0.0567(1.82) $\therefore$ a = 0.0005	34b and 22.13 = 11.13a + 12.16b $\Rightarrow$ 22.13=11.34(0.1036-0.567b)+12.16b, using (1) $\Rightarrow$ 22.13=1.175-0.643b+12.16b $\Rightarrow$ 20.955=11.517b $\therefore b = \frac{20.955}{11.517} = 1.82$

8. Criticise the following:
Regression coefficient of Y on X is 0.7 and that of X on Y is 3.2

Sol: $r = \sqrt{b_{yx}.\, b_{xy}} = \sqrt{3.3 \times 0.7} = \sqrt{2.24} = 1.5$ (approximately)
But correlation coefficient cannot exceed 1.
Hence there is some inconsistency in the information given.

9. Write the relation between correlation and regression coefficients. Is it possible to have two variables x and y with regression coefficient as 2.8 and -0.5? Explain.

Sol: We have correlation coefficient, $r = \sqrt{\sigma_{xy} \times \sigma_{yx}} = \sqrt{2.8 \times -(0.5)} = \sqrt{-1.4}$ which is not a real number.
Thus it is not possible.

DEVIATIONS TAKEN FROM THE ASSUMED MEAN:

If the actual mean is fraction this method is used.
In this method we take deviations from the assumed mean instead of Arithmetic Mean.

$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$

We can find out the value of $r\frac{\sigma_x}{\sigma_y}$ by applying the following formula.

$r\frac{\sigma_x}{\sigma_y} = \frac{\sum dx\, dy - \frac{\sum dx \times \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}, where\ dx = X - A; dy = Y - A$

The regression equations of Y on X is $Y - \bar{Y} = r\frac{\sigma_x}{\sigma_y}(X - \bar{X})$

We can find out the value of $r\frac{\sigma_x}{\sigma_y}$ by applying the following formula:

$b_{yx} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum dx\, dy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$

SOLVED PROBLEMS:

1. The following data, based on 450 students, are given for marks in statistics, economics at a certain examination.
Mean marks in Statistics = 40
Mean marks in Economics = 8
S.D. of marks in Statistics = 12
Variance of marks (Economics) = 256
Sum of the products of deviations of marks from this respective mean 42075. Give the equations of the two lines of regression and estimate the average marks in Economics of candidates who obtained 50 marks in Statistics.

Sol: Given $\bar{X}$ = Mean of marks in Statistics = 40
$\bar{Y}$ = Mean of marks in Economics = 8
σ_Y = S.D. of marks in Statistics = 12
σ_Y = S.D. of marks in Economics = 16
Coefficient of correlation, $r = \frac{\sum XY}{N\sigma_X\sigma_Y} = \frac{42075}{450 \times 12 \times 16} = 0.49$
Regression equation of X on Y is $X - \bar{X} = \frac{r\sigma_X}{\sigma_Y}(Y - \bar{Y}) \Rightarrow X = 0.37Y + 22.24$
Regression equation of Y on X is $Y - \bar{Y} = \frac{r\sigma_X}{\sigma_Y}(X - \bar{X}) \Rightarrow Y = 0.65X + 22$
When $X = 50, Y = 54.5$

2. Price indices of cotton and wool are given below for the 12 months of a year. Obtain the equations of lines of regression between the indices.

Price index of cotton (X)	78	77	85	88	87	82	81	77	76	83	97	93
Price index of wool (Y)	84	82	82	85	89	90	88	92	83	89	98	99

Sol: Calculation of Regression Equation.

Price index of cotton (X)	$(X-84)=dx$	dx^2	**Price index of wool (Y)**	$(Y-88)=dy$	dy^2	$dxdy$
78	-6	36	84	-4	16	24
77	-7	49	82	-6	36	42
85	1	1	82	-6	36	-6
88	4	16	85	-3	9	-12
87	3	9	89	1	1	3
82	-2	4	90	2	4	-4
81	-3	9	88	0	0	0
77	-7	49	92	4	16	-28
76	-8	64	83	-5	25	40
83	-1	1	89	1	1	-1
97	13	169	98	10	100	130
93	9	81	99	11	121	99
1004	**-4**	**488**	**1061**	**5**	**365**	**287**

Now $b_{xy}=\frac{\sum dxdy-\left(\frac{\sum dx\times\sum dy}{N}\right)}{\sum dy^2-\frac{(\sum dy)^2}{N}}=\frac{287-\left(\frac{-4\times5}{12}\right)}{365-\frac{5^2}{12}}=0.795$

Now Regression equation of X on Y : $X-\bar{X}=b_{xy}(Y-\bar{Y})$

$\Rightarrow X-83.67=0.795(Y-88.42)\Rightarrow X=0.795Y+13.38$

Now $b_{yx}=\frac{\sum dxdy-\left(\frac{\sum dx\times\sum dy}{N}\right)}{\sum dx^2-\frac{(\sum dx)^2}{N}}=\frac{287-\left(\frac{-4\times5}{12}\right)}{488-\frac{(-4)^2}{12}}=0.59$

Now Regression equation of Y on X : $Y-\bar{Y}=b_{yx}(X-X)$

$\Rightarrow Y-88.42=0.59(X-83.67)\Rightarrow Y=0.59X+39.05$

3. From the following data, calculate (i) Correlation coefficient (ii) Standard deviation of Y(σ_Y)
$b_{xy}=0.85;\ b_{yx}=0.89;\ \sigma_x=3$

Sol: (i) Coefficient of correlation:

$r=\sqrt{b_{xy}\times b_{yx}}=\sqrt{0.85\times0.89}=0.87$

(ii) Standard deviation of Y:

$r\times\frac{\sigma_x}{\sigma_y}=0.85=0.87\times\frac{3}{\sigma_y}=0.85\Rightarrow\sigma_y=3.07$

4. Given the following data, calculate the expected value of Y when X = 12

	x	**y**
Average	7.6	14.8
Standard deviation	3.6	2.5
r = 0.99		

Sol: We have to calculate the expected value of Y when X = 12.
So we have to find out the regression equation of Y on X
Mean of X series, $\bar{X}=7.6$; Mean of Y series, $\bar{Y}=14.8$
$\sigma\ of\ X\ series, \sigma_x=3.6;\quad \sigma\ of\ Y\ series, \sigma_y=2.5$
Coefficient of correlation (r) = 0.99
Regression of Y on X is

$Y - \bar{Y} = r.\frac{\sigma_x}{\sigma_y}(X - \bar{X}) \Rightarrow Y - 14.8 = (0.99) \times \frac{2.5}{3.6}(X - 7.6) = 0.688X + 9.57$

When $X = 12, Y = 0.688(12) + 9.57 = 17.826$

Hence the expected value of Y is 17.83.

5. The heights of mothers and daughters are given in the following table. From the two tables of regression estimate the expected average height of daughter when the height of the mother is 64.5 inches.

Height of mother (inches)	62	63	64	64	65	66	68	70
Height of daughter (inches)	64	65	61	69	67	68	71	65

Sol: Let X = Height of the mother

And Y = Height of the daughter

Let $dx = X - 65 \; and \; dy = Y - 67$. Then

$\sum X = 522, \sum dx = 2, \sum dx^2 = 50, \sum Y = 530m \; \sum dy = -6, \sum dy^2 = 74, \; \sum dxdy = 20$

Now $\bar{X} = \frac{\sum X}{N} = \frac{522}{8} = 66.25, \; \bar{Y} = \frac{\sum Y}{N} = \frac{522}{8} = 65.25$

$\therefore b_{yx} = \frac{\sum dxdy - \left(\frac{\sum dx \times \sum dy}{N}\right)}{\sum dx^2 - \frac{(\sum dx)^2}{N}} = \frac{20 - \left(\frac{2 \times -6}{8}\right)}{50 - \frac{2}{8}} = 0.434$

Hence Regression equation of Y on X : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$\Rightarrow Y = 37.93 + 0.434Z \; when \; X = 64.5 \; then \; Y = 65.923.$

6. The following calculations have been made for prices of 12 stocks (X) in stock exchange, on a certain day along with the volume of the sales in thousands of shares (Y). From these calculations find the regression equation of prices of stocks, on the volume of the sales of shares.

$\sum X = 580, \; \sum Y = 370, \; \sum XY = 11499, \; \sum X^2 = 41658, \; \sum Y^2 = 17.206$

Sol: We have Mean $= \bar{X} = \frac{\sum X}{N} = \frac{580}{12} = 48.33 \; and \; \bar{Y} = \frac{\sum Y}{N} = \frac{370}{12} = 30.83$

$\therefore b_{xy} = \frac{\sum XY - N\bar{X}\bar{Y}}{\sum Y^2 - N(\bar{Y})^2} = \frac{11494 - 12 \times 48.33 \times 30.83}{17206 - 12(30.83)^2} = -1.101$

Thus regression equation of X on Y is

$X - 48.33 = (-1.101)(Y - 30.83)$

$\Rightarrow X = -1.101Y + 82.27.$

7. Given the following information regarding a distribution $N = 5, \bar{X} = 10, \bar{Y} = 20,$ $\sum (X - Y)^2 = 100, (Y - 10)^2 = 160.$ Find the regression coefficients and hence the coefficient of correlation.

Sol: Here $dx = X - 4, \; dy = Y - 10$

$\bar{X} = A + \frac{\sum dx}{N} \Rightarrow 10 = Y + \frac{\sum dx}{5} \Rightarrow \sum dx = 30 \; \; (here \; A = 4)$

Also $\bar{Y} = B + \frac{\sum dy}{N} \Rightarrow 20 = 10 + \frac{\sum dy}{5} \Rightarrow \sum dy = 50, \; (B = 10)$

We know that

$b_{YX} = \frac{\sum dxdy - \left(\frac{\sum dx \times \sum dy}{N}\right)}{\sum dx^2 - \frac{(\sum dx)^2}{N}} = \frac{80 - \left(\frac{30 \times 50}{5}\right)}{100 - \frac{(30)^2}{5}} = \frac{80 - 300}{100 - 180} = \frac{-220}{-80} = 2.75$

$b_{XY} = \frac{\sum dxdy - \left(\frac{\sum dx \times \sum dy}{N}\right)}{\sum dy^2 - \frac{(\sum dy)^2}{N}} = \frac{80 - \left(\frac{30 \times 50}{5}\right)}{160 - \frac{(50)^2}{5}} = \frac{-220}{-340} = 0.65$

Coefficient of correlation = $\pm\sqrt{b_{XY} + b_{YX}} = \sqrt{0.65 \times 2.75} = \sqrt{1.7875} = 1.337$

We have $b_{XY} \; and \; b_{YX}$ is positive, so r is also positive.

Here we get the coefficient of correlation as more than 1. The given data is inconsistent.

ANGLE BETWEEN TWO REGRESSION LINES:

Let the lines of regression of X on Y and Y on X are respectively given by

$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$ ……. (1) and $y - \bar{y} = r\frac{\sigma_x}{\sigma_y}(x - \bar{x})$ ……. (2)

Slope of the line (1) $= m_1 = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$; Slope of the line (2) $= m_2 = r\frac{\sigma_y}{\sigma_x}$

Let θ be the angle between two regression lines X on Y and Y on X. Then

$$tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2} = \frac{\frac{1}{r}\frac{\sigma_y}{\sigma_x} - r\frac{\sigma_y}{\sigma_x}}{1 + \left(\frac{1}{r}\frac{\sigma_y}{\sigma_x}\right)\left(r\frac{\sigma_y}{\sigma_x}\right)} = \frac{\frac{\sigma_y}{\sigma_x}\left(\frac{1}{r} - r\right)}{1 + \frac{\sigma_y^2}{\sigma_x^2}} = \frac{\sigma_y}{\sigma_x}\left(\frac{1-r^2}{r}\right)\frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}$$

NOTE:

1. If θ is acute, $tan\theta = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} (\because -1 \le y \le 1)$
2. If θ is obtuse, $tan\theta = \left(\frac{r^2-1}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}$
3. If r = 0 then ,$tan\theta = \infty \Rightarrow \theta = \pi/2$
   Thus if there is no relationship between the two variables (i.e., they are independent) then $tan\theta = \pi/2$
4. If r = $\pm$ 1 then $tan\theta = 0 \Rightarrow \theta = 0 \; or \; \pi$
   Hence the two regression lines are parallel or coincident. The correlation between two variables is perfect.

SOLVED PROBLEMS:

1. If θ is the angle between two regression lines and S.D. of Y is twice the S.D. of X and r = 0.25, find tanθ.

Sol: Given $\sigma_y = 2\sigma_x \; and \; r = 0.25$

If θ is the angle between two regression lines, then

$$tan\theta = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} = \left(\frac{1-(0.25)^2}{0.25}\right)\frac{\sigma_x(2\sigma_x)}{\sigma_x^2 + 4\sigma_x^2} = \frac{1-0.0625}{0.25}.\frac{2}{5} = 1.5$$

2. If $\sigma_x = \sigma_y = \sigma$ and the angle between the regression lines is $Tan^{-1}\left(\frac{4}{3}\right)$. Find r.

Sol: $tan\theta = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \Rightarrow \theta = \tan^{-1}\left[\frac{1-r^2}{r}.\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}\right]$

Here $\sigma_x = \sigma_y = \sigma$

$\theta = \tan^{-1}\left[\frac{1-r^2}{r}.\frac{\sigma^2}{2\sigma^2}\right] = \tan^{-1}\left(\frac{1-r^2}{2r}\right)$ …… (1)

By data, $\theta = \tan^{-1}\left[\frac{4}{3}\right]$ …… (2)

From (1) and (2), we have

$\therefore \frac{1-r^2}{2r} = \frac{4}{3} \Rightarrow 3 - 3r^2 - 8r = 0 \Rightarrow 3r^2 + 8r - 3 = 0$

$\Rightarrow (3r - 1)(r + 3) = 0 \Rightarrow r = \frac{1}{3} \; or \; r = -3$

Since $-1 \le r \le 1$, we cannot have r = -3. ∴ r = 1/3.

3. The tangent of the angle between two regression lines is 0.6 and if $\sigma_x = \frac{1}{2}\sigma_y$, find the correlation coefficient between x and y.

Sol: We know that $tan\theta = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \Rightarrow 0.6 = \left(\frac{1-r^2}{r}\right)\frac{0.5\sigma_y^2}{\frac{\sigma_y^2}{4}+\sigma_y^2} \Rightarrow 0.6 = \left(\frac{1-r^2}{r}\right)\frac{0.5\sigma_y^2}{\sigma_y^2(1.25)}$

$\Rightarrow \left(\frac{1-r^2}{r}\right) = 1.5 = \frac{3}{2} \Rightarrow 2r^2 + 3r - 2 = 0 \Rightarrow r = \frac{3\pm\sqrt{9+16}}{4} = \frac{-3}{4} \pm \frac{5}{4} = -2 \; or \; \frac{1}{2}$

$\therefore r = \frac{1}{2}.$ $(\because \; r = -2 \; is \; not \; possible).$

PRACTICE PROBLEMS:

MEASURE OF CENTRAL TENDENCY :

1. Calculate the A.M of the following data

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks(x)	40	50	55	78	58	60	73	35	43	48

[Ans : 54]

2. From the following data find the mean profits.

Profits for shop	100-200	200-300	300-400	400-500	500-600	600-700	700-800
Number of shops	10	18	20	26	30	28	18

[Ans : 486]

3. Calculate median from the following data

Marks	10-25	25-40	40-55	55-70	70-85	85-100
Frequency	6	20	44	26	3	1

[Ans : 48.18]

4. Find the mode of the following distribution

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	7	12	28	20	10	10

[Ans : 46.666]

5. Find the geometric mean of following data

Yield of wheat (kg)	7.5-10.5	10.5-13.5	13.5-16.5	16.5-19.5	19.5-22.5	22.5-25.5	25.5-28.5
Frequency	5	9	19	23	7	4	1

[Ans : 16.02kg]

6. Calculate the H.M of the following data

Size of items	6	7	8	9	10	11
Frequency	4	6	9	5	2	8

[Ans : 8.23]

MEASURE OF DISPERSION:

1. Calculate the mean deviation from the median.

Class	0-10	10-20	20-30	30-40	40-50
Frequencies	5	10	20	5	10

[Ans : 9]

2. Find the variance and standard deviation for the following frequency distribution.

x	6	10	14	18	24	28	30
f	2	4	7	12	8	4	3

[Ans : 43.4]

3. Find the mean and variance using step deviation method for the following data.

Age in years	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of numbers	3	61	132	153	140	51	2

[Ans : 140.89]

CORRELATION AND REGRESSION:

1. Find the coefficient of correlation between X and Y for the following data

X	1	2	3	4	5	6	7	8	9
Y	10	11	12	14	13	15	16	17	18

[Ans : 0.8833]

2. A Sample of 12 fathers and their elder sons gave the following data about their elder sons. Calculate coefficient of rank correlation.

Fathers	65	63	67	64	68	62	70	66	68	67	69	71
Sons	68	66	68	65	69	66	68	65	71	67	68	70

[Ans:0.722]

3. Calculate the Regression equations of Y on X from the data given below taking deviations from actual mean of X and Y.

Price(Rs)	10	12	13	12	16	15
Amount Demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs.20 [Ans : 29.15]

****