CHAPTER - 1

DESCRIPTIVE STATISTICS

1.1 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 judgement 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 distinct 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.

1.2 STATISTICS

Def. : 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. Carefull 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 organisation.
3. **Presentation of Data :** The collected data must be edited very carefully so that irrelavent answers and wrogn computations must be corrected or adjusted.

   The collected data must be classified and tabulated before they can be analysed.
4. **Analysis of Data :** After presentation of the data the next step is to analyse the presented data. Analysis includes condensation, summarisation, conclusion etc., through means of Measures of Central Tendencies. Dispersion, Skewness, Kurtosis, correlation and Regression etc.
5. **Interpretation of Data :** Valid conslusions must be drawn on the basis of analysis. Correct interpretation leads to valid conclusion.

   The statistical generalisation provides the estimates of the characterstic behaviour 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 indispersible in all phases of human endeavour.

1.3 COLLECTION OF DATA

The basic problem of statistical enquiry is to collect facts and figures relating to a particular phenominan 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 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 muncipal 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 univercities 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.

1.4 POPULATION VS. SAMPLE

In Statistical enquiry, all the items, which fall within the perview of enquiry are known as **Universe** or **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 people.

1. **Finite and Infinite Population :**

Population can be either finite population or inifinite 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 inifinite, it is known as infinite population. For example, the number of stars in the sky is infinite population.

2. Hypothetical and Existant Population :

Universes can be classified as extent and hypothetical. An Universe containing persons or objects is known as existant 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 therotical 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 charecteristics 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.

a) 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.

1.5 TYPES OF VARIABLES

All 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 characterised 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 Variables: 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°C and 30°C is the same as 30°C to 40°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°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 metres is twice the distance of 5 metres.

1.6 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.

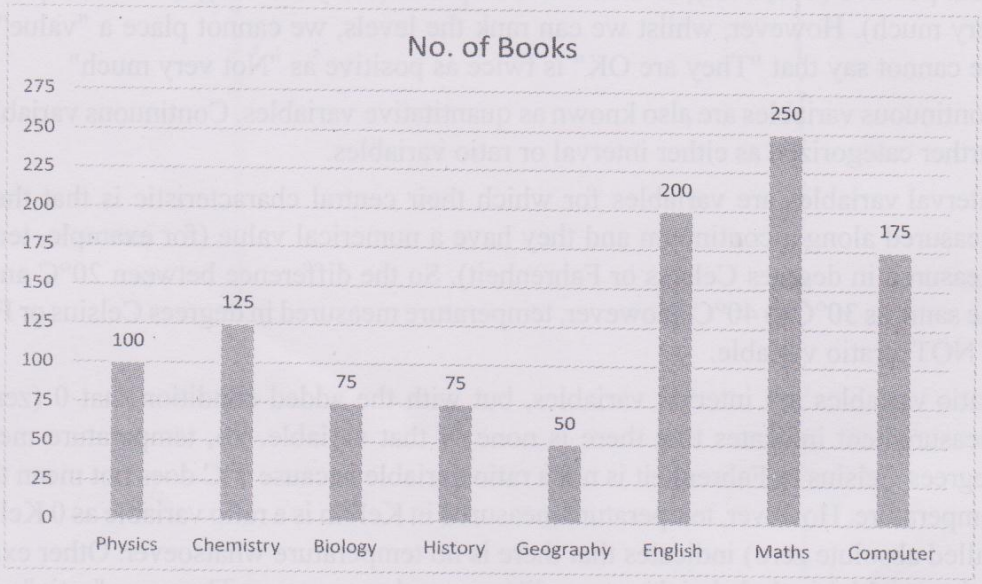
1. Bar Chart
2. Line chart
3. Scatter plot
4. Sparkline
5. Pie Chart

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.

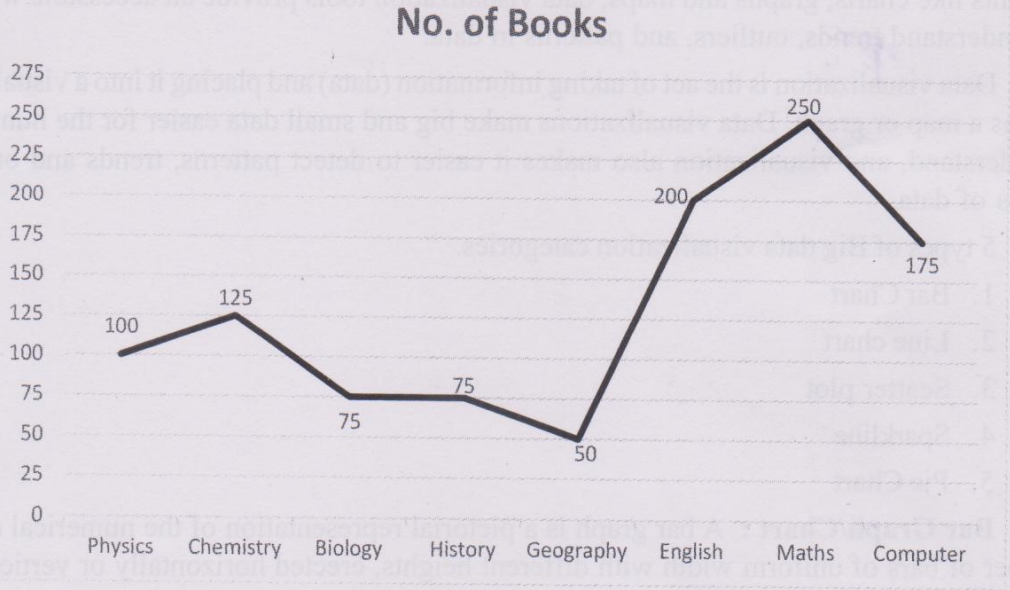
Ex. The following table shows the number of books of different subjects in a library.

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

Solution. 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 1/2 cm.

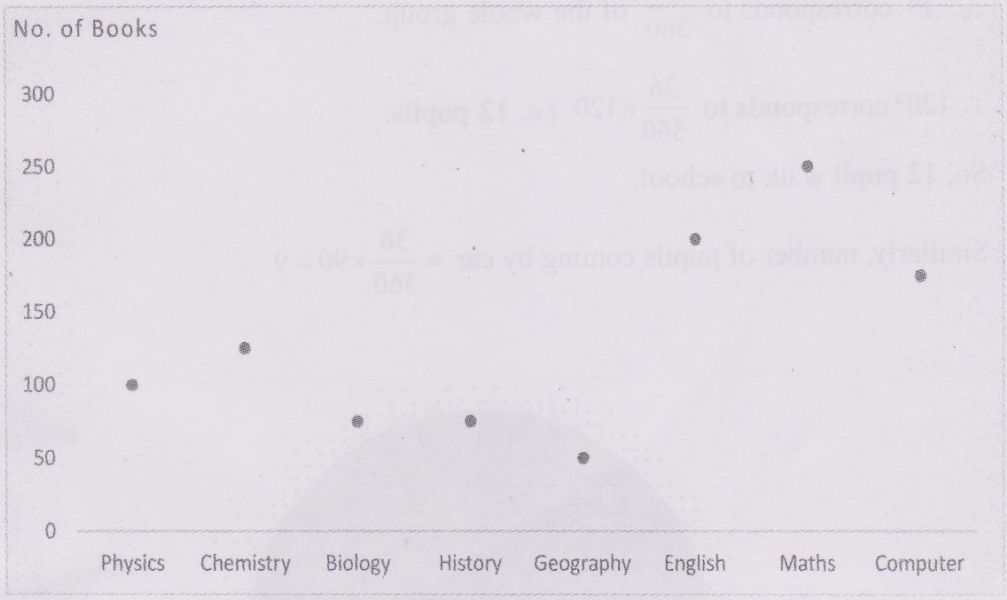


2. Line Graph/Chart : A line chart is a type of chart which displays information as a series of data points called markers connected by straight line segment.

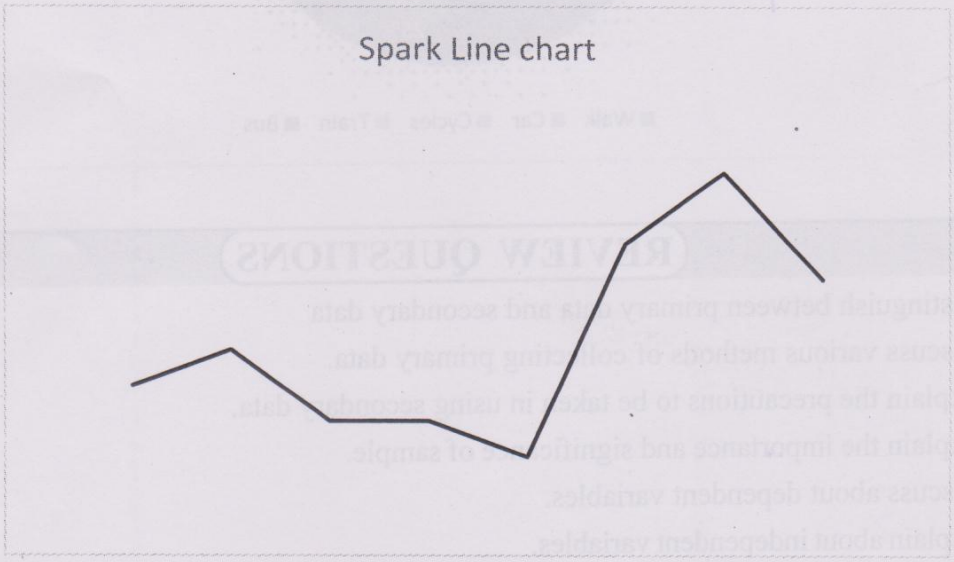


3. Scatter Graph/Chart : A scatter plot is a type of plot or mathematical diagram using cartesian coordinates to display values for two variables for a set of data. With scatter plots we often talk about how the variables relate to each other. This is called correlation.

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



4. Sparkline : A sparkline is a very small line chart drawn without axes or coordinates. It present the general shape of the variation in some measurement, such as temparature 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 $360°$. The angles at the centre are in proportion to each category. Thus, the angle of $120°$ at the centre corresponds to the walk group.

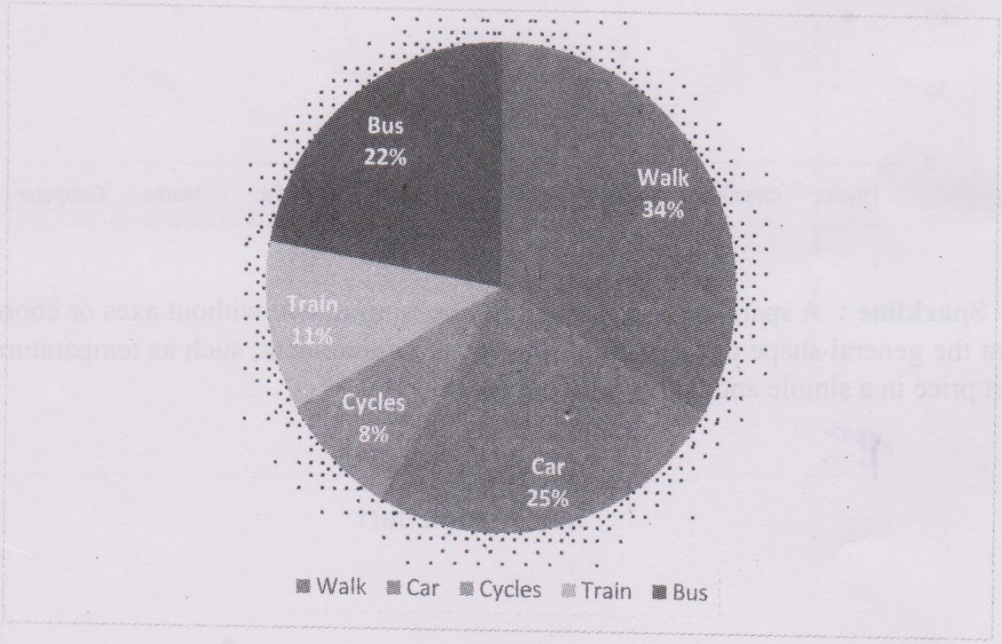
Since angle of $360°$ at the centre corresponds to the whole group of 36 pupils.

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

$\therefore$ $120°$ corresponds to $\frac{36}{360}\times 120$ *i.e.* 12 pupils.

So, 12 pupil walk to school.

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



REVIEW QUESTIONS

1. Distinguish between primary data and secondary data
2. Discuss various methods of collecting primary data.
3. Explain the precautions to be taken in using secondary data.
4. Explain the importance and significance of sample.
5. Discuss about dependent variables.
6. Explain about independent variables.
7. Discuss the catagorical variables.
8. Explain about continuous variables.
9. Discuss about Pie diagram.
10. Explain bar chart with an example.

CHAPTER - 2

MEASURES OF CENTRAL TENDENCY AND VARIABILITY

2.1. INTRODUCTION.

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.

2.2. MEASURES OF CENTRAL TENDENCY

The importance of statistical analysis is to find a number which represents in some definite way the entire data. Such a representative number is called the central value or an average. The value of an average lies somewhere in between the two extreme items possibly in the centre where most of the items concentrate. Hence an average constitutes a measure of the central tendency of the series. Measures of central tendency enable us to compare different groups of data.

The following are some important measures of central tendency in common use.

1. *Arithmetic Mean* (A.M.)
2. *Median*
3. *Mode*
4. *Geometric Mean* (G.M)
5. *Harmonic Mean* (H.M.)

An out line of formulae for the calculation of these measures of central tendency is given below.

2.3. ARITHMETIC MEAN

The arithmetic mean of a series of values of a variate is defined as the quantity obtained by dividing the sum of the value of variates by their number.

Arithmetic Mean of an Ungrouped Data

Let $x_1, x_2, ..., x_n$ be the n values of a variate x. Then the arithmetic mean is given by

$$\text{A.M.} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Arithmetic Mean of a Grouped Data

Let $x_1, x_2, ..., x_n$ be the midvalues of n classes of a frequency distribution with frequencies $f_1, f_2, ..., f_n$ respectively.

Since the value x_1 occurs f_1 times, the x_2 occurs f_2 times ... the value x_n occurs f_n times, the arithmetic mean by definition is given by

$$\text{A.M.} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i}$$

$$\text{A.M.} = \frac{1}{\text{N}} = \sum_{i=1}^{n} f_ix_i \quad \text{where } \text{N} = \sum_{i=1}^{n} f_i = \text{Total frequency.}$$

Arithmetic Mean of the Combined Data

The arithmetic means of two sets of data containing m, n items are $\bar{x}$ and $\bar{y}$ respectively. If the two sets of data are combined to get one set of $(m+n)$ items, then the arithmetic mean $\bar{z}$ of this set is given by $\bar{z} = \frac{m\bar{x} + n\bar{y}}{m+n}$.

2.4. MEDIAN

Median of a distribution is defined as the value of the variate which divides it into two equal parts when the variates are arranged in ascending or descending order of magnitude.

Median of an Ungrouped Data

Arrange the variates in ascending or descending order of magnitude. Determine the total number n of variates. If n is odd, then the median is the value of $\left(\frac{n+1}{2}\right)$ th variate. If n is even, the median is the average value of $\frac{n}{2}$ th and $\left(\frac{n}{2}+1\right)$ th variates.

i.e., If n is odd, Median $= \dfrac{\left(\frac{n}{2}\right)^{th} \text{ term} + \left(\frac{n}{2}+1\right)^{th} \text{ term}}{2}$

Median of a Grouped Data

In a continuous frequency distribution find the value of $\frac{N}{2}$ where N is the total frequency. Calculate the less than cumulative frequencies of the classes.

The class corresponding to the cumulative frequency just greater than $\frac{N}{2}$ is called the median class. The value of the median is calculated by the following formula :

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

where l = Lower limit of the median class

m = cumulative frequency of the class preceding the median class

N = Total frequency

f = Frequency of the median class

c = Width of the median class.

2.5. MODE

Mode or modal value of the distribution is defined as that value of the variate for which frequency is maximum.

Mode of an Ungrouped Data

When individual items are given, mode is the value of the item which occurs most frequently. However it is better to convert these individual items into a frequency table to calculate mode.

Mode of a Grouped Data

In a grouped data the class interval having the maximum frequency is called tha modal class. Even if the frequency distribution has class intervals of unequal magnitude the mode can be calculated provided the modal class and the classes preceding and succeeding it are of the same magnitude.

Some frequency distributions have two or more classes having the highest frequency. They are said to be *multimodal* distribution. The mode is not a good measure in such cases. Sometimes the midpoint of the class having the maximum frequency is taken as mode. In such cases it is called *crude* mode.

The mode of a grouped data is computed by the following formula:

$$\textbf{Mode} = l + \frac{f - f_1}{2f - f_1 - f_2} \times c$$

where l = Lower limit of the modal class

f = Frequency of the modal class

f_1 = Frequency of the class preceding the modal class

f_2 = Frequency of the class succeeding the modal class

c = Width of modal class.

2.6. GEOMETRIC MEAN

Geometric mean is defined as the nth root of the product of the values of a distribution none of them being zero.

Geometric Mean of an Ungrouped Data

Let $x_1, x_2, ..., x_n$ be the n values of the variate x none of them being zero. Then the geometric mean G is defined by $G = \sqrt[n]{x_1 x_2 ... x_n} = (x_1 x_2 ... x_n)^{1/n}$

The calculations can be easily made by taking logarithms both sides.

$$\log G = \log(x_1 x_2 ... x_n)^{1/n}$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n) = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$\therefore G = \text{Antilog}\left[\frac{1}{n}\sum_{i=1}^{n} \log x_i\right]$$

Geometric Mean of a Grouped Data

Let $x_1, x_2, ..., x_n$ be the mid-values of n classes of a frequency distribution with frequencies $f_1, f_2, ..., f_n$. Let $\sum_{i=1}^{n} f_i = N$. Then the geometric mean is defined as the Nth root of the product of the mid-values of the classes raised to the powers of their respective frequencies.

$$\therefore \text{ G.M.} = \sqrt[N]{x_1^{f_1} x_2^{f_2} ... x_n^{f_n}} = (x_1^{f_1} . x_2^{f_2} ... x_n^{f_n})^{(1/N)}$$

Geometric Mean of the Combined Data

If m items have a geometric mean x and n items have a geometric mean y, then the geometric mean G of the combined data is given by

$$\log G = \frac{m \log x + n \log y}{m+n}$$

2.7. HARMONIC MEAN

The harmonic mean (H.M.) of a series of values of defined as the reciprocal of the arithmetic mean of the reciprocal of individual items of which no item is equal to zero.

Harmonic mean of an ungrouped data

Let $x_1, x_2, ..., x_n$ be n terms of the data.

Their reciprocals are $\frac{1}{x}, \frac{1}{x_2}, ..., \frac{1}{x_n}$.

$$\therefore \text{ A.M. of these reciprocals } = \frac{1/x_1 + 1/x_2 + ... + 1/x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}$$

$\therefore$ Harmonic Mean = Reciprocal of the A.M. of the reciprocals

$$i.e., \; H.M. = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

2.8 HARMONIC MEAN OF A GROUPED DATA

Let $x_1, x_2, ..., x_n$ be the non-zero mid-values of n classes with frequencies $f_1, f_2, ..., f_n$.

Let $\sum_{i=1}^{n} f_i = N$.

Then $H.M.$ is calculated by the following formula.

$$H.M. = \frac{f_1 + f_2 + ... + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + ... + \frac{f_n}{x_n}} = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{x_i}}.$$

SOLVED EXAMPLES

Example 1 : Calculate the A.M. of the following data

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

Solution :

Roll No.s	Marks (X)
1	40
2	50
3	55
4	78
5	58
6	60
7	73
8	35
9	43
10	48
$n = 10$	$\sum X = 540$

Arithemetic Mean $= \bar{X} = \frac{\sum X}{n} = \frac{540}{10} = 54$

Example 2 : From the following, find the mean profits

Profits per shop (Rs)	120 – 200	200 – 300	300 – 400	400 – 500	500 – 600	600 – 700	700 – 800
Number of Shops	10	18	20	26	30	28	18

Solution : Calculation of Mean

Profit per shop	Mid value	$d = m - 450$	f	fd
100-200	150	– 300	10	–3000
200-300	250	–200	18	–3600
300-400	350	–100	20	–2000
400-500	450	0	26	0
500-600	550	100	30	3000
600-700	650	200	28	5600
700-800	750	300	18	5400
			$\sum f = 150$	$\sum fd = 5400$

Mean $= \bar{X} = A + \frac{\sum fd}{N}$

$= 450 + \frac{5400}{150} = 450 + 36 = 486$

$\therefore$ Average profit is Rs. 486

Example 3 : Find the median from the following : 57, 58, 61, 42, 38, 65, 72, 66

Solution : Arranging in ascending order, we get

38, 42, 57, 58, 61, 65, 66, 72. Here $n = 8$ (even)

$$\therefore \text{Median} = \frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{\text{th}} \text{term}}{2} = \frac{4^{\text{th}} \text{term} + 5^{\text{th}} \text{term}}{2} = \frac{58+61}{2} = \frac{119}{2} = 59.5$$

Example 4 : Calculate the 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

Solution : We will tabulate the values as follows :

Computation of the Median

Marks	Frequency	Cumulative Frequency
10-25	6	6
25-40	20	26
40-55	44	70
55-70	26	96
70-85	3	99
85-100	1	100

Here $\frac{N}{2} = \frac{100}{2} = 50$. Here the class just greater than $\frac{N}{2}$ is $40 - 55$.

$\therefore$ Median class is $40 - 55$.

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

where l = Lower end of the Median class $= 40$

$\frac{N}{2} = \frac{100}{2} = 50$

m = cumulative frequency of the class preceding the median class $= 26$

c = length of the class interval $= 15$

$$\therefore \text{Median} = 40 + \frac{50-26}{44} \times 15 = 40 + 8.18 = 48.18$$

Example 5 : Find the mode of the following salaries :

(*i*) 850, 750,600, 825, 850, 725, 600, 850, 640, 530

(*ii*) 40, 45, 48, 57, 78

(*iii*) 45, 55, 50, 45, 40, 55, 45, 55

Solution : (*i*) 85 is repeated 3 times. Therefore, modal salary = Rs. 850

(*ii*) No value is repeated. Thus there is no mode.

(*iii*) 45 is repeated 3 times. 55 is also repeated 3 times. Therefore there are two modes.

Mode (*i*) = 45 , mode (*ii*) = 55

Note :

(*i*) When we calculate mode from the data, if there is only one mode, then the series is called unimodal.

(*ii*) If there are two modes it is called bimodal

(*iii*) If there are 3 modes it is called trimodal.

(*iv*) If there are more than 3 modes it is called multimodal.

Example 6 : Find the mode of the following distribution.

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

Solution : Here the maximum frequency is 28.
Thus 40-50 is the model class

We use the formula, Mode $= l + \dfrac{(f - f_1)}{2f - f_1 - f_2} \times c$

where, $l =$ lower end of the model class $= 40$

$c =$ length of the class interval $= 10$

$f =$ frequency of the modal class $= 28$

$f_1 =$ frequency of the class precedings modal class $= 12$

$f_2 =$ frequency of the class succeding the modal class $= 20$

$\therefore$ Mode $= 40 + \dfrac{10(28 - 12)}{2 \times 28 - 12 - 30} = 40 + 6.666 = 46.666$

Example 7 : Find the geometric mean of the 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
No. of farms	5	9	19	23	7	4	1

Solution : The G.M. is calculated using the formula

G. M. $= \text{Antilog}\left(\dfrac{\Sigma f \log m}{N}\right)$ where m is the mid-value of the class interval.

We tabulate the values as follows :

Mid value (m)	$\log m$	f	$f \log m$
9	0.9542	5	4.7710
12	1.0792	9	9.7128
15	1.1761	19	22.3459
18	1.2553	23	28.8719
21	1.3222	7	9.2554
24	1.3802	4	8.5208
27	1.4314	1	1.4314
		$N = 68$	$\Sigma f \log m = 81.9092$

$$\therefore \text{ G. M. } = \text{Anti log}\left(\frac{\Sigma f \log m}{N}\right) = \text{Antilog}\left(\frac{81.9092}{68}\right)$$

$$= \text{Antilog } (1.2045) = 16.02 \text{ kgs.}$$

Example 8 : Calculate the H. M. of the following data

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

Solution : We tabulate the values as follows :

Size of item (X)	Frequency	$\frac{1}{X}$	$f\left(\frac{1}{X}\right)$
6	4	0.1667	0.6668
7	6	0.1429	0.8574
8	9	0.1250	1.1250
9	5	0.1111	0.5555
10	2	0.1000	0.2000
11	8	0.0909	0.7292
	$N = \Sigma f_i = 34$		$\Sigma\left(\frac{f}{X}\right) = 4.13$

$$\therefore \text{ H. M. } = \frac{N}{\Sigma \frac{f_i}{x_i}} = \frac{34}{4.1319} = 8.23$$

Example 9 : Calculate the H.M. of the following data

Marks	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100
Freq.	15	13	8	6	15	7	6

Solution : Calculation of H. M.

Marks	Mid value (m)	Freq.	$\frac{1}{m}$	$f\left(\frac{1}{m}\right)$
30-40	35	15	0.02857	0.42855
40-50	45	13	0.02222	0.28886
50-60	55	8	0.01818	0.14544
60-70	65	6	0.01534	0.09204
70-80	75	15	0.01333	0.19995
80-90	85	7	0.01176	0.08232
90-100	95	6	0.01053	0.06318
		$N = \Sigma f_i = 70$		$\Sigma\left(\frac{f}{m}\right) = 1.30034$

$$\text{H. M. } = \frac{N}{\Sigma\left(\frac{f_i}{m_i}\right)} = \frac{70}{1.30034} = 53.83$$

2.9. 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 distributions 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.

2.10 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
$\bar{X}_A = 5000$	$\bar{X}_B = 5000$

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

2.11 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.

2.12. MEAN DEVIATION FROM MEAN OF AN UNGROUPED DATA

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

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

SOLVED EXAMPLES

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

Solution : A.M. $= \bar{x} = \dfrac{40+62+54+68+76}{5} = \dfrac{300}{5} = 60$

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

$$= \frac{|40-60|+|62-60|+|54-60|+|68-60|+|76-60|}{5} = \frac{52}{5} = 10.4.$$

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

Solution : Mean $= \bar{x} = \dfrac{38+70+48+40+42+55+63+46+54+44}{10} = \dfrac{500}{10} = 50.$

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

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

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

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

Solution : The arithmetic mean of the given data is

(*a*) $\bar{x} = \dfrac{6+7+10+12+13+4+12+16}{8} = \dfrac{80}{8} = 10$

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

$\therefore$ Mean deviation from mean $= \dfrac{\sum |x_i - \bar{x}|}{n}$

$= \dfrac{4+3+0+2+3+6+2+6}{8} = \dfrac{26}{8} = 3.25$

(*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 $= \dfrac{10+12}{2} = \dfrac{22}{2} = 11$

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

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

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

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

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

Now Median $= \dfrac{\left(\dfrac{n}{2}\right)^{th} \text{ term} + \left(\dfrac{n}{2}+1\right)^{th} \text{ term}}{2} = \dfrac{42+44}{2} = 43$.

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 the median $= \dfrac{\sum |x_i - \text{Median}|}{n} = \dfrac{\sum |x_i - 43|}{n} = \dfrac{87}{10} = 8.7$.

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

Then the mean deviation about the arithmetic mean

$$= \frac{|x_i - \bar{x}| f_1 + |x_i - \bar{x}| f_2 + ... + |x_i - \bar{x}| f_n}{f_1 + f_2 + ... + f_n}$$

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

SOLVED EXAMPLES

Example 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

Solution : We will tabulate the values as follows :

x_i	f_i	$f_i x_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
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. $= \bar{x} = \frac{\sum f_i x_i}{\sum f} = \frac{320}{40} = 8$

$\therefore$ Mean deviation $= \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i} = \frac{140}{40} = 3.5$

Example 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

Solution : 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 is equal to 13.

Now we tabulate the absolute values of the deviations.

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

Thus $\sum f_i \lvert x_i - \text{med} \rvert = 149$

$\therefore$ Mean deviation from median $= \dfrac{\sum f_i \lvert x_i - \text{med} \rvert}{\sum f_i} = \dfrac{149}{30} = 4.97$

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

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

Solution : We tabulate the values as follows

size	freq.	cumulative freq.	deviations from Median $\lvert x_i - \text{med} \rvert$	$f_i \lvert x_i - \text{med.} \rvert$
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 $= \dfrac{\left(\frac{n}{2}\right)^{\text{th}} \text{ term} + 1}{2} = \dfrac{7+1}{2} = 4^{\text{th}} \text{ term} = 9.$

Mean deviation from median $= \dfrac{\sum f_i \lvert x_i - \text{med.} \rvert}{\sum f_i} = \dfrac{60}{48} = 1.25$

Example 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

Solution : **Calculations of Mean Deviation about Mean**

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

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

Mean deviation from mean $= \frac{\sum f_i \mid x_i - \bar{x} \mid}{\sum f_i} = \frac{158}{25} = 6.32$

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

Example 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

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

Sales	Number of companies f_i	Midpoint of the class x_i	$f_i x_i$	$\mid x_i - \bar{x} \mid$	$f_i \mid x_i - \bar{x} \mid$
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
	$\sum f_i = N = 100$		$\sum f_i x_i = 7100$		$\sum f_i \mid x_i - \bar{x} \mid = 1040$

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

$\therefore$ Mean Deviation from mean $= \frac{\sum f_i \mid x_i - \bar{x} \mid}{\sum f_i} = \frac{1040}{100} = 10.4$

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

Class interval	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20	20 – 24
Freq.	8	12	35	25	13	7

Class Interval	Mid value x_i	Freq. f_i	$f_i x_i$	$\lvert x_i - \bar{x} \rvert$ $= \lvert x_i - 11.76 \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
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
		$\Sigma f_i = N$ $= 100$	$\Sigma f_i x_i$ $= 1176$		$\Sigma f_i = \lvert x_i - \bar{x} \rvert$ $= 417.60$

Here $\bar{x} = \dfrac{\Sigma f_i x_i}{\Sigma f_i} = \dfrac{1176}{100} = 11.76$

$\therefore$ Mean deviation $= \dfrac{1}{N} \cdot \Sigma f_i \lvert x_i - \bar{x} \rvert = \dfrac{417.60}{100} = 4.176$

Step Deviation Method (Short Cut method)

Suppose in the given data the midpoints of the class intervals x_i and their associated frequences are numerically large. Then the computations become tedious. To avoid large caluculations, 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 devide 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, $d_i = \dfrac{x_i - a}{h}$.

Then A. M. $= \bar{x} = \left(\dfrac{\Sigma f_i d_i}{N} \right) h$

We illustrate the simplified procedure with some examples.

SOLVED EXAMPLES

Example 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
Freq.	4	8	9	10	7	5	4	3

Solution : We tabulate the data as follows :

classes	Mid values (x_i)	d_i	frequencies (f_i)	$f_i d_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
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
			50	4		7896

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

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

$\therefore$ Mean deviation from mean $= \frac{\sum f_i \lvert x_i - \bar{x} \rvert}{\sum f_i} = \frac{7896}{50} = 157.92$

Example 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	5	8	15	16	6

Solution : We form the following table

classes	Mid values (x_i)	freq. (f_i)	d_i	$f_i d_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
0 – 10	5	5	–2	–10	22	110
10 – 20	15	8	–1	–8	12	96
20 – 30	25	15	0	0	2	30
30 – 40	35	16	1	16	8	128
40 – 50	40	5	2	12	18	108
		$\sum f_i = N = 50$		$\sum f_i d_i = 10$		$\sum f_i \lvert x_i - \bar{x} \rvert = 472$

Here $\bar{x} = a + \frac{\sum f_i d_i}{N} \times c = 25 + \frac{10}{50} \times 10 = 27$ and $d_i = \frac{x_i - \bar{x}}{h} = \frac{x_i - 27}{10}$

$\therefore$ Mean deviation from mean $= \frac{f_i \lvert x_i - \bar{x} \rvert}{N} = \frac{472}{50} = 9.44$

Example 3 : Find Mean deviation from mean of the following data, using the step deviation method.

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of students	6	5	8	15	7	6	3

Solution : We will construct the following table to calculate the required values

classes	Mid point x_i	No. of students (f_i)	$d_i = \frac{x_i - 35}{10}$	$f_i d_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
0 – 10	5	6	–3	–18	28.4	170.4
10 – 20	15	5	–2	–10	18.4	92
20 – 30	25	8	–1	–8	8.4	67.2
30 – 40	35	15	0	0	1.6	24.0
40 – 50	45	7	1	7	11.6	81.2
50 – 60	55	6	2	12	21.6	129.6
60 – 70	65	3	3	9	31.6	94.8

$$\text{Mean } (\bar{x}) = A + h\frac{\sum f_i d_i}{N} = 35 + \frac{10(-8)}{50} = 33.4$$

$$\text{Mean deviation from mean} = \frac{1}{N}\sum f_i \lvert x_i - \bar{x} \rvert = \frac{659.2}{50} = 13.18$$

Example 4 : Find Mean deviation from median for the following data

Age of workers	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50	50 – 55	55 – 60
No. of workers	120	125	175	160	150	140	100	30

Solution : We form the following table for the given data

class interval	Midpoint x_i	Freq. f_i	Cumulative freq. $(c.f)$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
20 – 25	22.5	120	120	$\lvert 22.5 - 37.5 \rvert = 15$	1800
25 – 30	27.5	125	245	$\lvert 27.5 - 37.5 \rvert = 10$	1250
30 – 35	32.5	175	420	$\lvert 32.5 - 37.5 \rvert = 5$	875
35 – 40	37.5	160	580	$\lvert 37.5 - 37.5 \rvert = 0$	0
40 – 45	42.5	150	730	$\lvert 42.5 - 37.5 \rvert = 5$	750
45 – 50	47.5	140	870	$\lvert 47.5 - 37.5 \rvert = 10$	1400
50 – 55	52.5	100	970	$\lvert 52.5 - 37.5 \rvert = 15$	1500
55 – 60	57.5	30	1000	$\lvert 57.5 - 37.5 \rvert = 20$	600
		$N = 1000$			8175

Here $\frac{N}{2} = \frac{1000}{2} = 500$. The c.f. just greater than $\frac{N}{2}$ is 580.

The corresponding class interval is $35-40$. This is the median class.

$$\therefore \text{Median} = l + \left(\frac{\frac{N}{2} - m}{f}\right) \times c = 35 + \left(\frac{500-420}{160}\right) \times 5 = 35 + \frac{400}{160} = 35 + 2.5 = 37.5$$

and Mean deviation about the Median $= \frac{\Sigma f_i \,|x_i - \text{med.}|}{\Sigma f_i} = \frac{8175}{1000} = 8.175$

Example 5 : Calculate the mean deviation from the median of the following data

Wages / week (Rs.)	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
No. of workers	4	6	10	20	10	6	4

Solution :

wages / week (in Rs.)	Mid values(x_i)	Freq. f_i	Cumulative freq. $c.f$	$\lvert x_i - 45 \rvert$	$f_i \lvert x_i - 45 \rvert$
10 – 20	15	4	4	30	120
20 – 30	25	6	10	20	120
30 – 40	35	10	20	10	100
40 – 50	45	20	40	0	0
50 – 60	55	10	50	10	100
60 – 70	65	6	56	20	120
70 – 80	75	4	60	30	120
		$N = \Sigma f_i = 60$			$\Sigma f_i \lvert x_i - 45 \rvert = 680$

Here $N = 60$; $\frac{N}{2} = 30$

The cumulative frequencies just greater than $\frac{N}{2} = 30$ is 40 and the corresponding class is $40-50$. So $40-50$ is the median class.

Here $l = 40, f = 20, h = 10, \text{P.C.F} = 20, N = 60$

$$\therefore \text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c = 40 + \frac{30-20}{20} \times 10 = 45$$

and Mean deviation from Median $= \frac{680}{60} = 11.33$

2.14. 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 $x_1, x_2, ..., x_n$ are n observations and $\bar{x}$ is their mean , we conisder $\Sigma(x_i - \bar{x})^2$.

The following cases may arise :

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

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

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

If we take the mean of the squared deviations from the mean *i.e.*, $\frac{1}{n}\Sigma(x_i - \bar{x})^2$, 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.

$$\text{Variance} = \sigma^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$$

SOLVED EXAMPLES

Example 1 : Find the variance and standard deviation of the following data :
5, 12, 3, 18, 6, 8, 2, 10

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

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

Here $\Sigma(x_i - \bar{x})^2 = 194$

$$\text{Variance } (\sigma^2) = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{194}{8} = 24.25$$

$$\text{Standard deviation} = \sigma = \sqrt{24.75} = 4.95 \text{ (approx.)}$$

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

Solution : Mean $\bar{x} = \frac{45+60+62+60+50+65+58+68+49+48}{10} = \frac{560}{10} = 56$

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

$\therefore$ Variance, $\sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{662}{10} = 66.2$

Standard deviation $(\sigma) = \sqrt{66.2} = 8.136$

Variance and Standard Deviation for a Discrete Frequency Distribution.

Example 3 : Calculate the standard deviation for the following distribution

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

Solution : We costruct following table for computing the required values.

x_i	f_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i \lvert x_i - \bar{x} \rvert^2$
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
	$\Sigma f_i = N = 30$	$\Sigma f_i x_i = 420$			$\Sigma f_i (x_i - \bar{x})^2 = 1374$

Here $N = 30, \bar{x} = AM = \frac{\Sigma f_i x}{\Sigma f_i} = \frac{420}{30} = 14$

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

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

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

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

Solution : We have

x_i	f_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i \mid x_i - \bar{x} \mid^2$
6	2	12	–13	169	338
10	4	40	– 9	81	324
14	7	95	–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
	$\sum f_i = N$ $= 40$	$\sum f_i x_i$ $= 760$			$\sum f_i (x_i - \bar{x})^2$ $= 1736$

Here $N = \sum f_i = 40$, $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i} = \dfrac{760}{40} = 19$ and $\sum f_i (x_i - \bar{x})^2 = 1736$

$\therefore$ Variance $= \sigma^2 = \dfrac{\sum f_i (x_i - \bar{x})^2}{N} = \dfrac{1736}{40} = 43.4$

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

Variance and Standard deviation of a Continuious Frequency Distribution

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

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

ANOTHER METHOD :

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

We know that variance $\sigma^2 = \dfrac{1}{N} \sum_{i=1}^{n} f_i (x_i - \bar{x})^2$

$$= \frac{1}{N} \sum f_i (x_i^2 + \bar{x}^2 - 2 x_i \bar{x})$$

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

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

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 + \bar{x} - 2\bar{x}^2$$

$$= \frac{1}{N}\Sigma f_i x_i^2 - (\bar{x}^2) = \frac{1}{N}\Sigma f_i x_i^2 - \left(\frac{\Sigma f_i x_i}{N}\right)^2$$

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

Step Deviation Method (Short Cut Method) :

If the midvalues 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 $y_i = \frac{x_i - A}{h}, i = 1, 2, ..., n$. Then $x_i = A + hy_i$

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{\Sigma f_i (A + hy_i)}{N}$$

$$= \frac{1}{N}\left[\sum_{i=1}^{n} Af_i + \sum_{i=1}^{n} hf_i y_i\right] = \frac{1}{N}\left[A\sum_{i=1}^{n} f_i + h\sum_{i=1}^{n} f_i y_i\right]$$

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

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

$$= \frac{1}{N}\sum_{i=1}^{n} f_i h^2 (y_i - \bar{y})^2 = h^2\left[\frac{1}{N}\Sigma f_i (y_i - \bar{y})^2\right]$$

$$= h^2\sigma_y^2 \text{ (or) } \sigma_x = h\,\sigma_y$$

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

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

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

Example 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
Freq.	3	7	12	15	8	3	2

Solution : We tabulate the following way.

class interval (C.I.)	Freq. (f_i)	Mid point (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	8	2	4	6	12
90 – 100	2	95	3	9	6	18
	$N = \Sigma f_i = 50$				$\Sigma f_i y_i = -15$	$\Sigma f_i y_i^2 = 105$

Assumed mean $= A = 65$,

length of the class interval $= h = 10$

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

$$\text{Variance } \sigma^2 = \frac{h^2}{N^2}\left[N\Sigma f_i y_i^2 - (\Sigma f_i y_i)^2\right]$$

$$= \frac{100}{2500}[50(105) - (-15)^2] = 201$$

$\therefore$ Standard deviation $= 14.18$

Example 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

Solution : 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)	Frequencies (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
	$\Sigma f_i = N = 72$			$\Sigma f_i y_i = -31$	$\Sigma f_i y_i^2 = 239$

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

$$\text{Variance } \sigma^2 = \frac{h^2}{N^2}\left[N \Sigma f_i y_i^2 - (\Sigma f_i y_i)^2\right]$$

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

$\therefore\ \sigma_x = 88.52$

Example 7 : Find the mean deviation from the mean and standard deviation of the series $a, a + d, a + 2d, \ldots, a + 2nd$.

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

$$\bar{x} = A.M. = \frac{a + (a+d) + (a+2d) + \ldots + (a+2nd)}{2n+1}$$

$$= \frac{2n+1}{2}\left[2a + (2n+1-1)d\right] = a + nd$$

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

$$\text{Mean deviation} = \frac{\Sigma |x_i - \bar{x}|}{2n+1}$$

$$= \frac{nd + (n-1)d + (n-2)d + \ldots\ldots + d + 0 + d + \ldots\ldots + (n-1)d + nd}{2n+1}$$

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

$$= \frac{2d(1 + 2 + \ldots\ldots + n)}{2n+1} = \frac{2d\ n(n+1)}{2} \cdot \frac{1}{2n+1} = \frac{n(n+1)}{2} d\ .$$

Let the assumed mean be A. Then

$$\sigma^2 = \frac{1}{n}\Sigma(x_i - A)^2 - \left[\frac{\Sigma(x_i - A)}{n}\right]^2$$

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

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

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

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

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

Solution : We have $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$

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

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

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

Hence the result.

Example 9 : The variance of 20 observations is 5. If each of the observations is multiplied by 2, find the variance of the resulting observation.

Solution : Here $a = 2$ and $\sigma^2 = 5$

$\therefore$ Variance of the resulting observations $= a^2\sigma^2 = 2^2 \times 5 = 20$, using Example 8

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

Solution : Let $\bar{x}$ be the mean of $x_1, x_2, ..., x_n$.

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

Let the new observations be $y_i = x_i + k$ where k is + ve or − ve number.

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

$$= \frac{1}{n}\Sigma x_i + \frac{1}{n}\Sigma k = \bar{x} + \frac{1}{n}(kn) = \bar{x} + k$$

The variance of the new observations $= \sigma_2^2$

$$= \frac{1}{n}\Sigma(y_i - \bar{y})^2 = \frac{1}{n}\Sigma\left[(x_i + k) - (\bar{x} + k)\right]^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2 = \sigma_1^2$$

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

2.15 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 $\bar{x}$ and it is often expressed as a percentage.

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

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 $\bar{x}_1 = \bar{x}_2 = \bar{x}$ but different standard deviations σ_1 and σ_2 respectively. Then coefficient of variations are given by $\left(\frac{\sigma_1}{\bar{x}} \times 100\right)$ and $\left(\frac{\sigma_2}{\bar{x}} \times 100\right)$. Thus the C.V.'s can be compared using σ_1 and σ_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

Example 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

Solution : 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 $\bar{x}_1 = \bar{x}_2 = \bar{x} = 45$.

Hence section B has greater variability in the performance.

Example 2 : Goals scored by two teams A and B in foot ball 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.

Solution :

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		$\Sigma f_i d_i = -44$	$\Sigma f_i d_i^2 = 126$		50		– 44	140

Mean $= \bar{x} = a + \dfrac{f_i d_i}{f_i} = 2 - \dfrac{44}{50} = 1.12$

and $\sigma = \sqrt{\dfrac{1}{N}\Sigma f_i d_i^2 - \left(\dfrac{1}{N}\Sigma f_i d_i\right)^2}$

$= \sqrt{\dfrac{126}{50} - \left(\dfrac{-44}{50}\right)^2} = \sqrt{2.52 - 0.7774} = 1.32$

Mean $= \bar{x} = a + \dfrac{f_i d_i}{f_i} = 2 - \dfrac{44}{50} = 1.12$

and $\sigma = \sqrt{\dfrac{1}{N}\Sigma f_i d_i^2 - \left(\dfrac{1}{N}\Sigma f_i d_i\right)^2}$

$= \sqrt{2.8 - 0.7744} = 1.42$

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

Hence team A is more consistent.

Example 3 : Lives of two models of refrigerators A and B, are given below. Which refrigeration 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

Solution :

Class interval	Mid point x_i	x_i^2	Model A			Model B		
			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
			$\Sigma f_i = N = 46$	212	1221	$N = 49$	297	2025

For model A, $\bar{x}_A = \dfrac{\Sigma f_i x_i}{\Sigma f_i} = \dfrac{212}{46} = 4.6$

and $\sigma_A = \sqrt{\dfrac{f_i x_i^2}{N} - \left(\dfrac{f_i x_i}{N}\right)^2} = \sqrt{5.38} = 2.319$

For model B, $\bar{x}_B = \dfrac{\Sigma f_i x_i}{\Sigma f_i} = \dfrac{297}{49} = 6.06$

and $\sigma_B = \sqrt{\dfrac{2025}{49} - \left(\dfrac{297}{49}\right)^2} = \sqrt{4.61} = 2.147$

Coefficient of variation for model A $= \dfrac{\sigma_A \times 100}{\bar{x}_A} = \dfrac{2.319}{4.6} \times 100 = 50.41$

Coefficient of variation for model B $= \dfrac{\sigma_B \times 100}{\bar{x}_B} = \dfrac{2.147}{6.06} \times 100 = 35.43$

Hence the model A is suggested to purchase.

Example 4 : In example 6 find the coefficient of variance.

Solution : We have mean $\bar{x} = 278.47$ and standard deviation $= \sigma_x = 88.52$

$\therefore$ Coefficient of variation $= \dfrac{88.52}{278.47} \times 100 = 31.79$

Example 5 : The following data gives the analysis of month wise wages paid to the workers of two firms A and B belonging to the same industry

(*i*) which firm has greater variability in individual works ?

(*ii*) which firm has larger wage bill ?

	Firm A	Firm B
Number of workers	500	600
Average daily wage	186	175
Variance of distribution of wages	81	100

Solution : (*i*) For the firm A variance of distribution of wages, $\sigma_1^2 = 81 \Rightarrow \sigma_1 = 9$

Similarly, for B, $\sigma_2^2 = 100 \Rightarrow \sigma_2 = 10$

$\therefore$ C. V. of the distribution of wages of firm A $= \frac{\sigma_1}{\bar{x}_1} \times 100 = \frac{9}{180} \times 100 = 4.84$

and C. V. of the distribution of wages of firm B $= \frac{\sigma_2}{\bar{x}_2} \times 100 = \frac{10}{175} \times 100 = 5.71$

Since C. V. of firm B > C. V of firm A, we conclude that the firm B has greater variability in individual wages.

(*ii*) Number of workers for firm A $(x_1) = 500$

Daily average wage $= \bar{x}_1 =$ Rs.186

Total daily wages paid to workers of A $= n_1 x_1 = 500 \times 186 =$ Rs. 93,000

For the firm B, number of wage earners $= n_2 = 60$

Average daily wage $= \bar{x}_2 =$ Rs.175

Total daily wages paid to workers of firm B $= n_2 \bar{x}_2 = 600 \times 175 =$ Rs.1,05,000

$\therefore$ Firm B has larger wage bill.

Example 6 : The scores of two cricketers A and B in 10 innings are given here. Find who is a better run getter and who is more a consistent player.

Scores of A x_i	40	25	19	80	38	8	67	121	66	76
Scores of B y_i	28	70	31	0	14	111	66	31	25	4

Solution : For cricketer A, average $\bar{x} = \frac{540}{10} = 54$

For cricketer B, average $\bar{y} = \frac{380}{10} = 38$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
40	–14	196	28	–10	100
25	29	841	70	32	1024
19	–35	1225	31	–7	49
80	26	676	0	–38	1444
38	–16	256	14	–24	576
8	– 46	2116	111	73	5329
67	13	169	66	28	784
121	67	4489	31	–7	49
66	12	144	25	–13	169
70	22	484	4	–34	1156
$\Sigma x_i = 540$		10596	$\Sigma y_i = 380$		10,674

Standard deviation of A $= \sigma_x = \sqrt{\frac{1}{n}(x_i - \bar{x})^2} = \sqrt{\frac{10596}{10}} = 32.55$

Standard deviation of B $= \sigma_y = \sqrt{\frac{1}{n}(y_i - \bar{y})^2} = \sqrt{\frac{10674}{10}} = \sqrt{1067.4} = 32.67$

C. V. of A $= \frac{\sigma_x}{\bar{x}} \times 100 = \frac{32.55}{54} \times 100 = 60.28$

C. V. of B $= \frac{\sigma_y}{\bar{y}} \times 100 = \frac{32.67}{38} \times 100 = 86$

Since $\bar{x} > \bar{y}$, cricketer A is a better rungetter

But C. V of A < C. V. of B, $\therefore$ cricketer A is more consistent player.

EXERCISE

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

(*i*) 6, 7, 10, 12, 13, 4, 8, 12 (*ii*) 3, 6, 10, 4, 9, 10

2. Find the mean deviation about the median for the following data :

(*i*) 22, 24, 30, 27, 29, 31, 25, 28, 41, 42 (*ii*) 4, 6, 9, 3, 10, 13, 2

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

(*i*)

x_i	10	30	50	70	90
f_i	4	24	28	16	8

(*ii*)

x_i	10	11	12	13
f_i	3	12	18	12

4. Calculate the mean deviation from the median.

(*i*)

classes	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Frequences	5	10	20	5	10

(*ii*)

Marks obtained	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
No. of boys	5	8	7	12	28	20	10	10

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

(*i*)

Classes	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80
Freq.	2	3	8	14	8	3	2

(*ii*)

Height (in cms)	95–105	105–115	115–125	125–135	135–145	145–155
No. of boys	9	13	26	30	12	10

6. Find the variance for the discrete data given below

(*i*) 4,5,2,8,7 (*ii*) 6,7,10,12,13,4,8,12

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

(*i*)

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

(*ii*)

x_i	60	61	62	63	64	65	66	67	68
f_i	2	1	12	29	25	12	10	4	5

8. (*i*) 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

(*ii*)

Classes	0–30	30–60	60–90	90–120	120–150	150–180	180–210
Freq.	2	3	5	10	3	5	2

9. From the prices of shares X and Y given below, for 10 days of trading, which share is more stable?

X	35	54	52	53	56	58	52	50	51	49
Y	108	107	105	105	106	107	104	103	104	101

10. The coefficients of variations of two distributions are 60 and 70 and their standard deviations are 21 and 16 respectively. Find their arithmetic means.

11. The mean of 5 observations is 4.4. Their variance is 8.24. If three of the observations are 1,2,6. Find the other two observations.

12. The arithmetic mean and standard deviation of a set of 9 items are 43 and 5 respectively. If an item of value 63 is added to that set find the mean and standard deviation of 10 items set given.

ANSWERS

1. (*i*) 2.79 (*ii*) 2.67 **2.** (*i*) 4.7 (*ii*) 3.28 **3.** (*i*) 16 (*ii*) 0.71

4. (*i*) 9 (*ii*) 170.58 **5.** (*i*) 14.284 (*ii*) 11.288 **6.** (*i*) 4.72, 2.172

(*ii*) 10.125, 3.181 **7.** (*i*) 43.4, 6.59 (*ii*) 2.86, 1.69 **8.** (*i*) 140.89, 11.86

(*ii*) 227.61, 47.708 **11.** 4, 9 **12.** 58.5, 7.648

2.16 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 characterise 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 or 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 $(S_{k_p}) = \frac{\bar{X} - \text{mode}}{\sigma}$

In case mode is illdefined.

The coefficient of skewness $(S_{k_p}) = \frac{3\,(\text{Mean} - \text{Median})}{\sigma} = \frac{3(\bar{X} - M)}{\sigma}$

SOLVED EXAMPLES

Example 1 : Calculate Karl Pearson's coefficient of skewness for the following data : 25, 15, 23, 40, 27, 25, 23, 25, 20.

Solution :

Computation table of Mean and Standard Deviation

Size	Deviation from A = 25 $(d) = \text{size} - A$	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
	$\Sigma d = -2$	$\Sigma d^2 = 362$

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

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

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

Example 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

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

Variable X	Mid value (m)	Freq. (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
		$\Sigma f = N = 75$		$\Sigma fd' = -9$		$\Sigma fd'^2 = 193$

Here c = Class interval $= 5$

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

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

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

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

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

2.17 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 μ_4 = fourth moment, μ_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

Example 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
Freq.	1	3	4	2

Solution :

Computation of Moments, Skewness and Kurtosis

Income	Mid value (m)	Frequency (f)	$d=\frac{m-15}{10}$	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			$\Sigma fd = 7$	$\Sigma fd^2 = 13$	$\Sigma fd^3 = 19$	$\Sigma fd^4 = 37$

(*i*) Moments about mean .

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

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

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

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

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

(*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.

REVIEW QUESTIONS

1. (*i*) Explain Skewness. (*ii*) Define Skewness
2. Distinguish between positive and negative Skewness
3. What are the tests of Skewness.
4. (*i*) Define Kurtosis (*ii*) What is Kurtosis. How does it differ from skewness.
5. Explain the terms : (*i*) Lepto Kurtic (*ii*) Meso Kurtic (*iii*) Platy Kurtic