No. of lecture required	Name of Topic	Content of topics	Difficulty Level	Bloom's Taxonomy Level	Learning Outcomes (Learner will be able to)
1	**Introduction** 4.1	Reason for learning Statistics, Statistical thinking and analysis, Types of Statistical methods, Importance and scope of Statistics (Statistics in Engineering), Limitations			Explain the role and importance of statistics in engineering applications
1	**Collection of data** 4.2	Need for data, collection of primary data, collection of secondary data, case study method, advantages and limitations			Identify sources and methods of data collection
1	**Classification of data, Graphical representation of data** 4.3	Basis of classification (Geographical, chronological, quantitative and qualitative), method of data classification (exclusive and inclusive method), Bar chart, Pie chart			Classify and present data using appropriate graphical methods
1	**Measures of central tendency** 4.4	Requisites of a measure of central tendency, arithmetic mean, median			Compute measures of central tendency
1	4.4 **Measures of central tendency (Contd.)**	Mode, Geometric mean, Harmonic mean			Apply various measures of central tendency in problem-solving
1	**Measures of dispersions** 4.5	Range and coefficient of range, Quartile deviation, coefficient of quartile deviation			Calculate measures of dispersion for given data sets
1	**Measures of dispersions (Contd.)** 4.5 4.6.1	Mean deviation, coefficient of mean deviation, Standard deviation,			Interpret dispersion results and their implications
1	4.6 **Standard deviations of combination of two groups, Coefficient of variation**	Standard deviations of combination of two groups, Coefficient of variation, , **Applications**			Evaluate coefficient of variation and compare two data groups
1	**Moments** 4.7	Central moments, raw moments, moment about origin			Compute moments for given data
1	**Moments (Contd.)** 4.7	Relations between moments, Karl Pearson's coefficients			Relate moments to coefficients and interpret results
1	**Skewness** 4.8	Meaning and test of skewness, moment coefficients of skewness			Measure and interpret skewness in data sets
1	4.9 4.10 **Kurtosis**	Definition and measure of kurtosis Steps for computing kurtosis, **Applications**			Evaluate kurtosis and its implications for data distribution

UNIT - IV

STATISTICAL TECHNIQUES - I

4.1 :- INTRODUCTION :-

Why Statistics ?.

Statistics is the science of collecting, organizing, presenting, analyzing and interpreting data. We live in a data driven world where statistics helps us move from intuition to evidence-based decisions in business, engineering, science and government depend heavily on data.

- without statistics → decisions are based on in intuition or guesswork
- with statistics → decisions are based on facts and evidence

4.1.1 Reason for learning statistics.

1. **Decision-Making :-** statistics provides evidence-based insights to make better choices in engineering, business, medicine and daily life.

Ex:- A company analyzes survey results to determine the launch of new product.

2. **Research and Development :-** Almost every field (science, economics, AI, social studies etc.) relies on statistics to validate findings.

example:- Doctors use clinical trials data to check if a new drug works.

3. Understanding Uncertainty:- Statistics helps measure and manage uncertainty.

example:- Weather forecast (probability of rain).

4. Problem-Solving:- It helps identifying trends, test hypotheses, and find solutions.

example:- Engineer use statistical quality control to reduce defects in production.

5. Data-Driven World:- In today's digital age, we are surrounded by data. Statistics is the backbone of big data, machine learning, and artificial intelligence.

Table- Real life need of Statistics

Field	Use of Statistics	Example
Education	Performance and Evaluation	students exam score analysis
Engineering	Quality control	Defect rate in automobile manufacturing
Business	Forecasting	Stocks inventory, customer data to understand buying habits.
Government	Planning resources, budgets & Policies	census data helps to build hospital, road or school
Medicine	Testing new drugs	Covid-19 vaccine trials.

4.1.2: Statistical Thinking and Analysis

Statistical thinking can be defined as the thought process that focuses on ways to identify, control and reduce variations present in all phenomena or processes.

* A way of approaching problem using data, not guses
* Focus on variations
* Make evidence based decisions
* Consider uncertainty and probability

Flow chart:-

Problem → Data collection → Variation Analysis → Model Building → Interpretation → Action

Statistical Analysis

Statistical analysis applies methods to interpret data and support decisions.

Steps in Statistical Analysis:-

1. Define the problem/objective → Need to identify the objective
2. Collect data → collect relevant data (surveys, experiments, observations)
3. Organize and summarize data:→ Arrange data in tables, charts or graphs, Use mean, median, standard deviation

4. Analyze data → Apply statistical methods (regression, hypothesis testing, correlation etc.)

5. Interpret results → Draw conclusions from the analysis

6. Make decisions/Action → Use findings to make informed decisions or predictions

4.1.3 — Types of statistical Methods

(a) Descriptive Statistics

* Summarize and organize data into meaningful form.
* Purpose:- To present data in a simple, understandable form
* Tools:- Mean, Median, Mode, Standard deviation, tables, graphs, charts
* Example:- The average marks of 200 students is 65.

(b) Inferential statistics:-

* Used for estimation of population characteristics on the basis of sample results and testing of statistical hypothesis.
* Purpose:- Make conclusions or predictions about a population based on sample

* Tools:- Estimation, hypothesis testing, Regression, ANOVA, Chi square tests.

Example:- From a survey of 1000 voters elections results can be predicted

(c) Predictive statistics:-

* Uses past data and models to forecast future trends.

* Tools: Time series, Regression, Machine learning models

Example:- Using traffic data from past 10 years to predict future traffic flow on a highway and design proper lane capacity.

4.1.4 Importance and scope of statistics

* General Importance.

(i) In Business:- Market surveys, demand forecasting, cost analysis

(ii) Engineering and Technology:- Reliability testing, optimization, quality assurance

(iii) Medical and Biological Sciences:- clinical trials, genetics, epidemiology

(iv) Social Sciences:- Surveys, education, psychology, population studies

(v) Government & Administration :- Census, national income, budgeting, public welfare

(vi) Research and Innovation :- Designing experiments, testing hypotheses

4.1.5 :- Limitations :-

(i) Statistics does not deals with isolated measurement

* Statistics study groups or masses of data not individual

* Example :- Average income does not show the income of a particular person.

(ii) Depends upon the quality of data

* If the data collected is biased, incomplete or inaccurate, results will be misleading

(iii) Statistical results are true only on an average

* conclusions are not universally true

* Example :- If the average height of a student is 160m does not mean that every student's height is 160m

(iv) can be misused or mislead

* wrong methods, biased sampling or selective presentation can mislead people

4.2 COLLECTION OF DATA

4.2.1 NEED FOR DATA

Statistical data are the basic material method to make an effective decision in a particular situation. The main reasons for collecting data are as listed below.

(a) To provide necessary inputs to a given phenomenon or situation under study.
(b) To measure performance in an ongoing process such as production, service & so on.
(c) To enhance the quality of decision making by enumerating alternative courses of action in a decision making process & selecting an appropriate one.
(d) To satisfy the desire to understand an unknown phenomenon.
(e) To assist in guessing the causes and probable effects of certain characteristics in given situations.

In order to design an experiment or conduct a survey one must understand the different types of data and their measurement levels.

4.2.2 COLLECTION OF PRIMARY DATA

The methods which may be used for primary data collection are briefly discussed below:

OBSERVATION In observational studies, the investigators does not ask questions to seek clarifications on certain

issues. Instead he records the behaviour, as it occurs, of an event in which he is interested. Sometimes mechanical devices are also used to record the desired data.

→ Studies based on observations are best suited for researchers requiring non-self report descriptive data.

Interviewing Interviews can be conducted either face to face or over telephone. Such interviews provide an opportunity to establish a report with interviewer & help to extract valuable information.

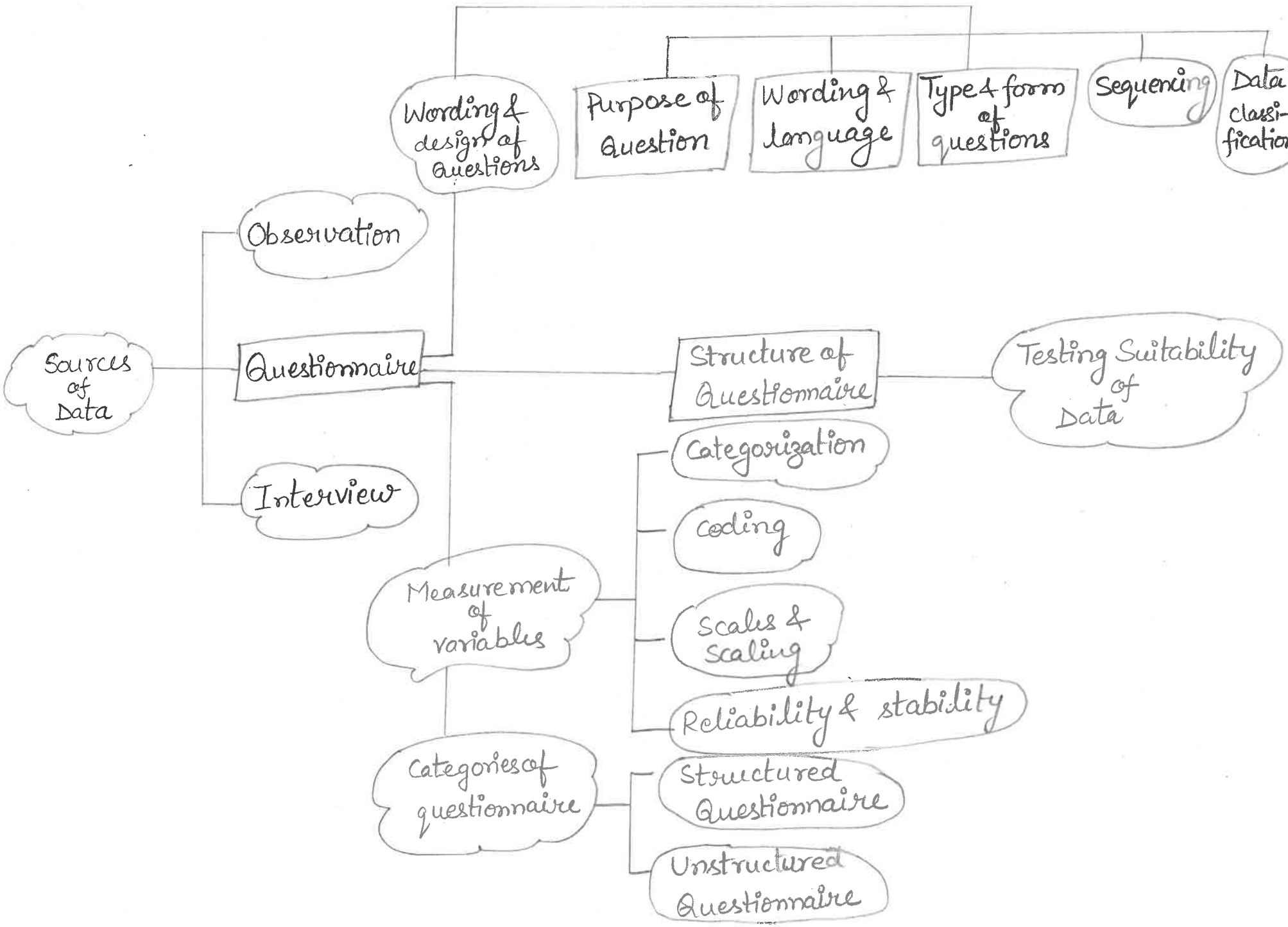
Questionnaire It is a formalized set of questions for extracting information from the target respondents. The form of the questions should correspond to the form of the required information.

The three general form of questions are: dichotomous (yes/no), multiple choice & open ended.

It is an efficient method of collecting primary data when the investigator knows what exactly is required & how to measure such variables of interest as:

- Behaviour – past, present or intended.
- Demographic characteristics – age, sex, income & occupation
- Level of knowledge
- Attitude & opinions

However, general principles of questionnaire design based on numerous studies & experiences of survey researchers.


Sources of Data
Observation
Questionnaire
Interview
Wording & design of questions
Purpose of Question
Wording & language
Type & form of questions
Sequencing
Data classification
Structure of Questionnaire
Testing Suitability of Data
Measurement of variables
Categorization
Coding
Scales & scaling
Reliability & stability
Categories of questionnaire
Structured Questionnaire
Unstructured Questionnaire

4.2.3. COLLECTION OF SECONDARY DATA

Secondary data means data that are already available, i.e. they refer to the data which have already been collected and analysed by someone else. When the researchers utilises secondary data, then they have to look into various sources from where they can obtain them.

Secondary data may either be published or unpublished data. Usually published data are available in:

(a) various publications of the central, state or local governments
(b) various publications of foreign governments or of international bodies
(c) technical & trade journals
(d) books, magazines & newspapers
(e) Report & publications of various associations connected with banks, business, industry, stock exchange etc.
(f) reports prepared by research scholars in different fields.
(g) Public records & statistics, historical documents & other sources of published information.

Before using secondary data, researcher must see that they possess following characteristics:

- Reliability of data
- Suitability of data
- Adequacy of data

PRIMARY DATA

It is the information collected by first hand by a researcher for a specific purpose or project. It is original & has not been used or published before.

→ Collected directly from the source
→ More accurate for intended research

For example :- lab experiment results, sensor readings, prototype test results, performance benchmarks, survey responses, field observations, Interview records, etc.

SECONDARY DATA

It is the information that has already been collected and recorded by someone else and is used for a purpose different from its original collection.

→ Obtained from existing sources (books, reports, research papers, database.)
→ May not perfectly match your exact research needs.

For example :- Published research papers, Government or industry reports, Online database, company documentation, market analysis reports etc.

Comparison Table :-

Aspect	Primary Data	Secondary Data
Source	Collected first hand by researcher	Collected by someone else, used from existing source.
Originality	Original & unique	Already existing and possibly used before
Purpose	Collected for a specific research objective	Collected for a different purpose, reused for your study.
Cost & Time	Usually costly and time consuming	Usually cheaper and quicker.
Accuracy	High	May vary, depends on original source quality.

4.2.4 CASE STUDY METHOD

The case study is essentially an intensive investigation of the particular unit under consideration. The object of the case study method is to locate the factors that account for the behaviour-patterns of the given unit as an integrated totality.

Characteristics:

1. Under this method the researcher can take one single social unit or more of such units for his study purpose.
2. Here the selected unit is studied intensively i.e. it is studied in minute details.
3. In the context of this method we make complete study of social unit covering all facets.
4. In respect of the case study method, an effort is made to know the mutual inter-relationship of casual factors.
5. Under case study method the behaviour pattern of the concerning unit is studied directly and not by an indirect or abstract approach.

4.2.5 ADVANTAGES AND LIMITATIONS

☆ There are several advantages of case study method:

1. Through case study a researcher can obtain a real & enlightened record of personal experiences which would reveal one's inner strivings, tensions & motivations that drive him to action along with the forces that direct him

to adopt a certain pattern of behaviour.

2. This method enables the researcher to trace out the natural history of the social unit and its relationship with the social factors and the forces involved in its surrounding environment.

3. It helps in formulating relevant hypothesis along with the data which may be helpful in testing them.

4. Case study method enhances the experiences of student and thus in turn increases his analysing ability & skills.

5. Information collected under the case study method helps a lot to the student in the task of constructing the appropriate questionnaire or schedule for the said task requires thorough knowledge of the concerning universe.

★ Important limitations of case study method may as well be highlighted:

1. Read Bain does not consider the case data as significant scientific data since they do not provide knowledge of the "impersonal, universal, non-ethical, non-practical, repetitive aspects of phenomena." Real information is often not collected because the subjectivity of the researcher does not enter in the collection of the information in a case study.

2. The danger of false generalisation is always there in view of the fact that no set rules are followed in collection of the information and only few units are studied.

3. It consumes more time and requires a lot of expenditure. More time is needed under the case study method since one studies the natural history cycles of social unit & that too minutely.

4. Case study method is based on several assumptions which may not be very realistic at times, and as such the usefulness of case data is always subject to doubt.

5. Case study method can only be used in a limited sphere, it is not possible to use it in case of a big society. Sampling is also not possible under a case study method.

Besides, case studies, in modern time can be conducted in such a manner that the data are amenable to quantification and statistical treatment. Possibly, this is also the reason why case studies are becoming popular day by day.

4.3 Classification of data

4.3.1 Basis of Classification :- Statistical data are classified after taking into account the nature, scope and purpose of an investigation. Generally, data are classified on the basis of the following four bases :-

(a) Geographical Classification :- In this classification, data are classified on the basis of geographical or locational differences such as - cities, districts, or villages between various elements of the data set. Example :-

City	: Mumbai	Kolkata	Delhi	Chennai
Population density (per square km)	: 654	685	423	205

Such a classification is also known as spatial classification. These are generally listed in alphabetical order. Elements in the data set are also listed by the frequency size.

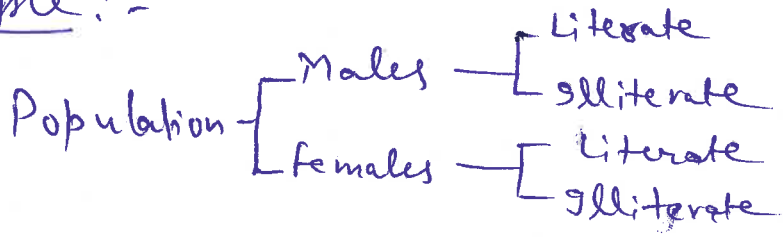
(b) Chronological Classification :- When data are classified on the basis of time, the classification is known as chronological classification. Such classification are also called time series. Example :-

Year	: 1941	1951	1961	1971	1981	1991	2001
Population (crore)	: 31.9	36.9	43.9	54.7	75.6	85.9	98.6

(c) Qualitative Classification :- In this classification, data are classified on the basis of descriptive characteristics or on the basis of attributes like sex, literacy, region, caste, or education, which can not be quantified. This is done in 2 ways :-

(i) Simple classification :- Here, each class is subdivided into 2 sub-classes and only one attribute is studied such as : male and female, blind and not blind, educated and uneducated and so on.

(ii) Manifold Classification :- Here, a class is subdivided into more than 2 sub-classes which may be sub-divided further. Example :-



(d) Quantitative Classification :- In this classification, data are classified on the basis of some characteristics which can be measured such as height, weight, income, expenditure, production, or sales.

Quantitative variables can be divided into the following 2 types.

(i) Continuous Variable is the one that can take any value within the range of numbers. Thus the height or weight of individuals can be of any value with in the limits.

(ii) Discrete (discontinuous) Variable is the one whose values change by steps or jumps and can not assume a fractional value. The number of children in a family, number of workers (or employees), no. of students in a class, are few examples of a discrete variable.

4.3.2 Method of ~~Data~~ Classification :- There are 2 ways in which observations in the data set are classified on the basis of class interval, namely (i) Exclusive Method and (ii) Inclusive method

(i) Exclusive Method :- when data are classified in such a way that the upper limit of a class interval is the lower limit of the succeeding class interval then it is said to be the exclusive method of classifying data.

Example :-

Dividends Declared in percent (class interval)	Number of Companies (frequencies)
0 - 10	5
10 - 20	7
20 - 30	15
30 - 40	10

To avoid confusion data are displayed in a slightly different manner as follows ;-

Dividends in percent	No. of Companies
0 but less than 10	5
10 but " " 20	7
20 " " " 30	15
30 " " " 40	10

(ii) Inclusive Method :- When the data are classified in such a way that both lower and upper limits of a class interval are included in the interval itself, then it is said to be inclusive method.

Example :-

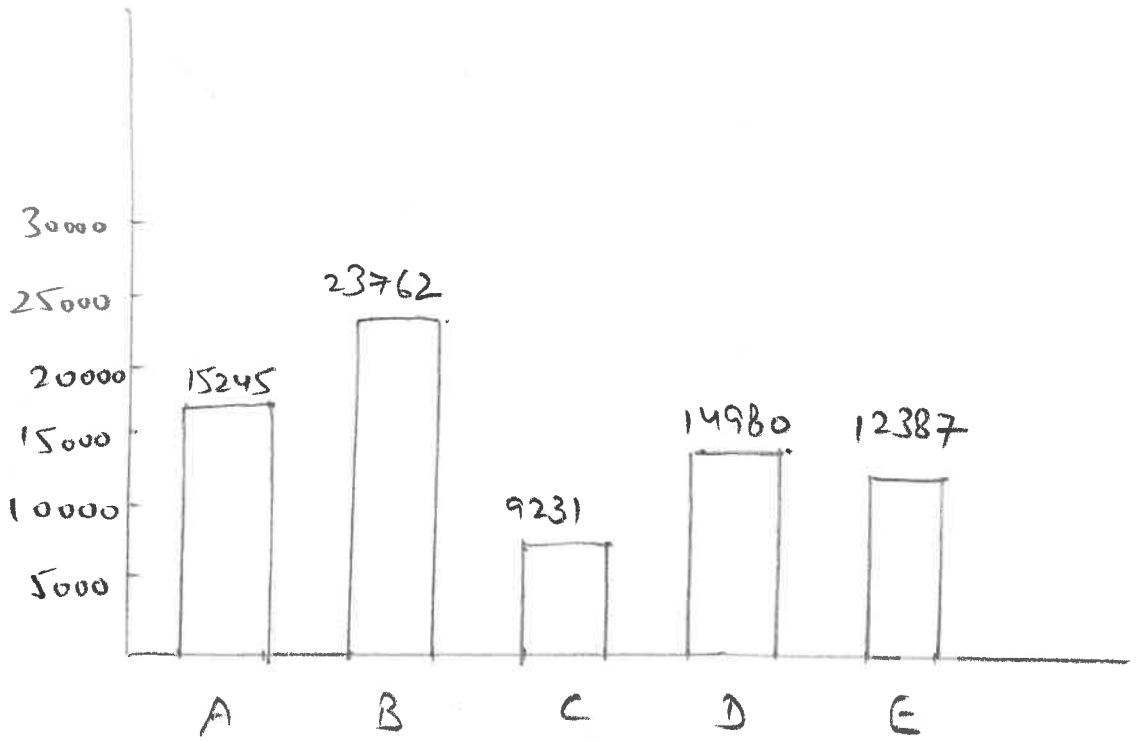
Number of Accidents (Class interval)	No. of weeks (Frequency)
0-4	5
5-9	22
10-14	13
15-19	8
20-24	2

4.3.3 Graphical Representation of data :-

Graphs help to understand the data easily. All statistical packages, MS Excel and open-office.org offer a wide range of graphs. In case of qualitative data most common graphs are bar charts and pie charts.

4.3.4 Bar Chart :- A bar chart consist of a series of rectangles (or bars). The height of each rectangle is determined by the frequency of that category.

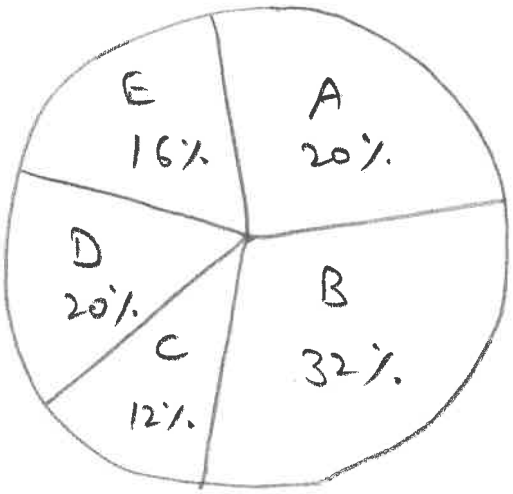
Example :- The sales of a popular soft drink in the year 2010-11, in five geographical regions, denoted as A, B, C, D and E are 15245, 23762, 9231, 14980 and 12387 respectively measured in 10000 USD. A bar chart of this date is as below :-



4.3.5 Pie Chart :- A pie chart is used to emphasize relative proportion or shares of each category. It's a circular chart divided into sectors, illustrating relative frequency. The relative frequency in each category or sector is proportional to the arc length of that sector or the area of that sector or the central angle of that sector.

Example :- In the previous example, if the soft drink has its market only in five geographical regions, denoted as A, B, C, D, E.

A total sale of the soft drink is 75605 times 10000 USD. A pie chart can be plotted to have the idea of the shares of different markets.



4.4.1 REQUISITES OF A MEASURE OF CENTRAL TENDENCY

The following are few requirements to be satisfied by an average or a measure of central tendency:

1. It should be rigidly defined.
2. It should be based on all the observations.
3. It should be easy to understand and calculate.
4. It should have sampling stability.
5. It should be capable of further algebraic treatment.
6. It should not be unduly affected by extreme observations.

According to Professor Bowley, "Averages are statistical constants which enable us to comprehend in a single effort the significance of the whole."

There are five types of averages in common use :-
(i) Arithmetic mean (ii) Median (iii) Mode
(iv) Geometric mean (v) Harmonic mean.

4.4.2 ARITHMETIC MEAN

In case of individual observations i.e. where frequency is not given,

$$A.M.\ \bar{x} = \frac{1}{n}\Sigma x$$

If the frequency distribution is given,

x : $x_1, x_2, x_3, \ldots, x_n$

f : $f_1, f_2, f_3, \ldots, f_n$, then

$$A.M = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n} = \frac{\Sigma fx}{\Sigma f} \quad ; N = \Sigma f$$

In the case of continuous series having equal class intervals, say of width h, we use a different formula (Step deviation method)

Let $u = \frac{x-a}{h}$ then $x = a + hu$

$\therefore \Sigma fx = \Sigma f(a+hu) = a\Sigma f + h\Sigma fu$

Dividing both sides by $N = \Sigma f$; we get

$$\frac{\Sigma fx}{N} = a + \frac{h\Sigma fu}{N}$$

$$\Rightarrow \boxed{\bar{x} = a + h\frac{\Sigma fu}{N}} \quad ; \text{ where } u = \frac{x-a}{h}$$

Example 1. Find the mean from the following data:

Marks	No. of Students
Below 10	5
Below 20	9
Below 30	17
Below 40	29
Below 50	45
Below 60	60
Below 70	70
Below 80	78
Below 90	83
Below 100	85

Solution: The freq. distribution table can be written as:

Marks	Mid Value (x)	f	$x-55$	$u = \frac{x-55}{10}$	fu
0-10	5	5	-50	-5	-25
10-20	15	4	-40	-4	-16
20-30	25	8	-30	-3	-24
30-40	35	12	-20	-2	-24
40-50	45	16	-10	-1	-16
50-60	55	15	0	0	0
60-70	65	10	10	1	10
70-80	75	8	20	2	16
80-90	85	5	30	3	15
90-100	95	2	40	4	8
		$N = \Sigma f = 85$			$\Sigma fu = -56$

$$\text{Mean } \bar{x} = a + h\frac{\Sigma fu}{N} = 55 + 10\left(\frac{-56}{85}\right)$$

$$\bar{x} = 55 - \frac{112}{17} = 48.41 \text{ marks}$$

Example 2. The mean of 200 items was 50. Later on it was discovered that two items were misread as 92 and 8 instead of 192 and 88. Find out the correct mean.

Solution :- Here incorrect values are 192 & 88.

so incorrect value of $\bar{x} = 50$, $n = 200$

$$\because \bar{x} = \frac{\Sigma x}{n} \Rightarrow \Sigma x = n\bar{x}$$

using incorrect value of $\bar{x}$,

Incorrect $\Sigma x = 200 \times 50 = 10000$

$\therefore$ Corrected value of $\Sigma x = 10000 - (92+8) + (192+88)$

$= 10180$

$$\text{Correct mean} = \frac{\text{Corrected } \Sigma x}{n} = \frac{10180}{200} = 50.9$$

4.4.3 MEDIAN

Median is the central value of the variable when the values are arranged in ascending or descending order of magnitude.

For an ungrouped frequency distribution, if n values of the variate are arranged in ascending or descending order of magnitude.

(a) When n is odd, the middle value i.e $\left(\frac{n+1}{2}\right)^{th}$ value gives the median.

(b) When n is even, there are two middle values $\left(\frac{n}{2}\right)^{th}$ & $\left(\frac{n}{2}+1\right)^{th}$ the arithmetic mean of these two values gives the median.

For a grouped frequency distribution, the median is given by the formula,

$$Md = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

where, l ≡ lower limit of median class, where median class is the class corresponding to cummulative frequency just $\geq \frac{N}{2}$

h ≡ width of the median class;

f ≡ frequency of median class; $N = \Sigma f$;

C ≡ Cummulative frequency of the class preceding the median class

Example:- Obtain the median for following frequency distribution

x	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6

Solution:- The cummulative frequency distribution table is given below:

x	f	C.f
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120

Here $N = 120 \Rightarrow \frac{N+1}{2} = 60.5$; Since C.f is greater than $\frac{N+1}{2}$ is 65 and the value of x corresponding to C.f. 65 is 5, hence median is 5.

Example:- Find median from the following data:

Marks	No. of students
Below 10	15
Below 20	35
Below 30	60
Below 40	84
Below 50	94
Below 60	127
Below 70	198
Below 80	249

Solution: Construct the cummulative frequency table:

Marks	No. of Students (f)	C.f.
0-10	15	15
10-20	20	35
20-30	25	60
30-40	24	84
40-50	10	94
50-60	33	127
60-70	71	198
70-80	51	249

Here $N = 249$

$\frac{N}{2} = 124.5$, $\therefore$ median class is $50-60$; $l = 50$

$h = 10$, $f = 33$, $C = 94$

$$\therefore \text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - C\right) = 50 + \frac{10}{33}(124.5 - 94)$$

$$= 59.24 \text{ marks}$$

4.4.4 **MODE**: It is the point of maximum frequency or the point of greatest density. In other words, the mode or modal value of the distribution is that value of the variate for which frequency is maximum.

In case of continuous frequency distribution, mode is given by the formula : $\left[\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h\right]$

where, $l \equiv$ lower limit, $h \equiv$ width

$f_m \equiv$ frequency of modal class.

f_1 & f_2 are the frequencies of the classes preceding and succeeding the modal class respectively.

✱ For a symmetrical distribution, mean, median & mode coincide.

✱ Empirical formula $[\text{Mode} = 3\text{Median} - 2\text{Mean}]$

Example :- Calculate the mode from the following frequency distribution:

Size (x) :	4	5	6	7	8	9	10	11	12	13
frequency(f) :	2	5	8	9	12	14	14	15	11	13

Solution :- Method of Grouping :

Size x	Frequency					
	I	II	III	IV	V	VI
4	2	7	13			
5	5					
6	8	17		15		
7	9		21		22	
8	12	26				29
9	14		[28]	35		
10	14	[29]			[40]	
11	15			[40]		[43]
12	11	24	26		39	
13	13					

Explanation :

In column I : original frequencies are written.

In column II : frequencies of column I are combined two by two.

In column III : leave the first frequency of column I and combine the others two by two.

In column IV : frequencies of column I are combined three by three.

In column V : leave the first frequency of column I & combine the others three by three.

In column VI : leave the first two frequencies in column I & combine the others three by three.

* (a) Maximum frequencies are written in boxes.
  (b) All operations are done in column I.

Now we frame another table in which against every maximum item of columns I to VI, we write down the corresponding size or sizes (x). This size (x) which occurs maximum number of times is "MODE."

Columns	Size of items having max. frequency
I	11
II	10, 11
III	9, 10
IV	10, 11, 12
V	8, 9, 10
VI	9, 10, 11

Since the item 10 occurs maximum number times (i.e. 5 times), hence mode is 10.

Example:- Find the mode of the following:

Marks:	1-5	6-10	11-15	16-20	21-25
No. of candidates:	7	10	16	32	24

Marks:	26-30	31-35	36-40	41-45
No. of candidates:	18	10	5	1

Solution: Here, the greatest frequency 32 lies in the class 16-20, hence the modal class is 16-20. But the actual limits of this class are 15.5 - 20.5.

$\Rightarrow l = 15.5,\ f_m = 32,\ f_1 = 16,\ f_2 = 24,\ h = 5$

$\therefore$ Mode $= l + \dfrac{f_m - f_1}{2f_m - f_1 - f_2} \times h = 15.5 + \dfrac{32-16}{64-16-24} \times 5$

$= 18.83$ marks

4.4.5 GEOMETRIC MEAN

(a) Geometric mean (G.M) of n individual observations $x_1, x_2, \ldots, x_n$ $(x_i \neq 0)$ is the n^{th} root of their product.

$$\therefore G = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{1/n}$$

$$\log_{10} G = \frac{1}{n}(\log x_1 + \log x_2 + \log x_3 + \ldots + \log x_n)$$

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

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

(b) If $x_1, x_2, \ldots, x_n$ occur $f_1, f_2, \ldots, f_n$ times respectively & $N = \sum_{i=1}^{n} f_i$, then G.M is given by,

$$G = (x_1^{f_1} x_2^{f_2} \ldots\ldots x_n^{f_n})^{1/N}$$

$$\log_{10} G = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n)$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i$$

$$G = \text{antilog}\left[\frac{1}{N}\sum_{i=1}^{n} f_i \cdot \log x_i\right]$$

(C) In the case of continuous frequency distribution, x is taken to be the value corresponding to the mid points of the class-interval.

Example: Compute the geometric mean from the following data:

Marks:	0-10	10-20	20-30	30-40	40-50
No. of Students:	10	5	8	7	20

Solution:

x :	5	15	25	35	45
f :	10	5	8	7	20
$\log_{10} x$:	0.6990	1.1761	1.3979	1.5441	1.6532
$f \cdot \log x$:	6.9900	5.8805	11.1832	10.8087	33.0640

here $\Sigma f \log x = 67.9264$

$\Rightarrow \log_{10} G = \frac{1}{N}\Sigma f \log x = \frac{67.9264}{50} = 1.3585$

$\Rightarrow G = \text{antilog}_{10} 1.3585 = 22.83$

4.4.6 HARMONIC MEAN

H.M. of a number of observations is the reciprocal of arithmetic mean of the reciprocals of the given values. Thus, the harmonic mean H of n observations $x_1, x_2, \ldots, x_n$ is

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$$

If $x_1, x_2, \ldots, x_n$ have the frequencies $f_1, f_2, \ldots, f_n$ resp. then h.m is given by,

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{x_i}} = \frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\cdots+\frac{f_n}{x_n}} \quad ; N = \Sigma f_i$$

Example: Find out the harmonic mean of the following data-

Marks (out of 150) :	10	20	40	60	120
No. of Students :	2	3	6	5	4

Solution:

x :	10	20	40	60	120	$\Sigma f = N$
f :	2	3	6	5	4	
$1/x$:	0.1	0.05	0.025	0.017	0.008	
f/x :	0.2	0.150	0.150	0.085	0.032	

$$H.M = \frac{N}{\Sigma f/x} = \frac{20}{0.617} = 32.4$$

Example: An aeroplane flies along the four sides of a square at speeds of 100, 200, 300 & 400 km/hr respectively. What is the average speed of the aeroplane in its flight around the square?

Solution: When equal distances are covered with unequal speeds, the harmonic mean is the proper average,

$$\therefore \text{Average Speed} = \frac{4}{\frac{1}{100}+\frac{1}{200}+\frac{1}{300}+\frac{1}{400}} = 192 \text{ km/hr.}$$

4.5.1 RANGE

Range is the most simple measure of dispersion and is based on the location of the largest and the smallest values in the data.

"Thus, the range is defined to be the difference between the largest and the lowest observed values in a data set."

Range (R) = Highest value of an observation − lowest value of an observation

= H − L

For grouped frequency distributions of values in the data set, the range is the difference b/w the upper class limit of the last class and the lower class limit of first class.

COEFFICIENT OF RANGE The relative measure of range, called the coefficient of range is

obtained by applying the following formula,

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

Example:- The following are the sales figures of a firm for the last 12 months.

Months :	1	2	3	4	5	6	7	8	9	10	11	12
Sales : (Rs '000)	80	82	82	84	84	86	86	88	88	90	90	92

Calculate the range and coefficient of range for sales.

Solution: Given that $H = 92$ & $L = 80$, therefore

$$\text{Range} = H - L = 12$$

$$\& \text{ Coefficient of range} = \frac{H-L}{H+L} = \frac{12}{172} = 0.069$$

Example: The following data shows the waiting time (to the nearest 100th of a minute) of telephone calls to be matured:

Waiting Time :	0.10–0.35	0.36–0.61	0.62–0.87	0.88–1.13	1.14–1.39
frequency :	6	10	8	8	4

Calculate the range and coefficient of range.

Solution: Given that $H = 1.39$ & $L = 0.10$, therefore

$$\text{Range} = H - L = 1.39 - 0.10 = 1.29 \text{ minutes}$$

$$\& \text{ Coefficient of range} = \frac{H-L}{H+L} = \frac{1.39-0.10}{1.39+0.10} = 0.865$$

4.5.2 QUARTILE DEVIATION

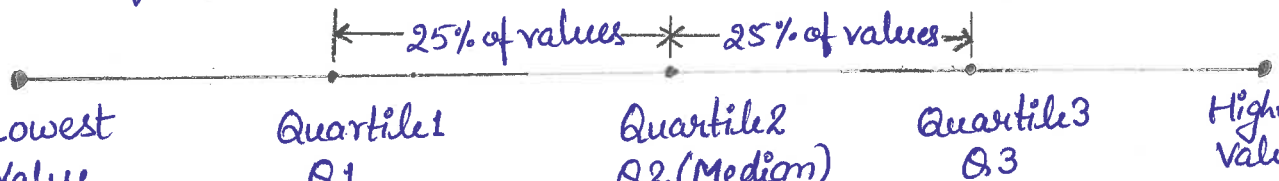
Quartile: In statistics, quartiles are three points Q_1, Q_2, Q_3, that devide a dataset into four equal parts, with each part containing 25% of the data.

Quartile deviation is the dispersion in the middle of the data where it defines the spread of the data.

$$\left[\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}\right]$$

For a symmetrical distribution, median lies midway b/w Q_1 & Q_3.

The quartile deviation measures the average range of 25% of the values in the data set



In a non-symmetrical distribution, the two quartiles Q_1 & Q_3 are at equal distance from median, i.e.

$$\text{Median} - Q_1 = Q_3 - \text{Median}.$$ Thus median ± quartile deviation covers exactly 50 percent of the observed values in the data set.

4.5.3 Coefficient of Quartile deviation: Since quartile deviation is an absolute measure of variation, therefore its value gets affected by the size & number of observed values in the data set. The Q.D. of two or more than two sets of data may differ. Due to this reason, to compare

the degree of variation in the different sets of data, we compute the relative measure corresponding to Q.D., called the coefficient of Q.D;

$$\text{Coeff. of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example: Use an appropriate measure to evaluate the variation in the following data:

Farm size (acre)	below 40	41-80	81-120	121-160	161-200	201-240	241 & above
No. of farms	394	461	391	334	169	113	148

Solution: Since the frequency distribution has open-end class-intervals on the two extreme sides, therefore Q.D. would be an appropriate measure of variation. The computation is as follows:

Farm size (acre)	No. of Farms (f)	Cumulative frequency C.f (less than)
below 40	394	394
41-80	461	855 ← Q_1 class
81-120	391	1246
121-160	334	1580 ← Q_3 class
161-200	169	1749
201-240	113	1862
241 and above	148	2010
	$\Sigma f = 2010$	

Q_1 = Value of $\left(\frac{n}{4}\right)$th observation = $2010 \div 4$ or 502.5th observation. This observation lies in the class 41-80.

Therefore,

$$Q_1 = l + \frac{(n/4) - cf}{f} \times h$$

$$Q_1 = 41 + \frac{502.5 - 394}{461} \times 40 = 41 + 9.41 = 50.41 \text{ acres}$$

Q_3 = Value of $(3n/4)$th observation

$= (3 \times 2010) \div 4$ or 1507.5th observation

This observation lies in the class 121-160. Therefore

$$Q_3 = l + \frac{(3n/4) - cf}{f} \times h$$

$$Q_3 = 121 + \frac{1507.5 - 1246}{334} \times 40$$

$$= 121 + 31.31$$

$$= 152.31 \text{ acres}$$

Thus, the quartile deviation is given by,

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{152.31 - 50.41}{2} = 50.95 \text{ acres}$$

& coefficient of $Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{50.95}{202.72} = 0.251$

4.5.4 "Mean Deviation"

Introduction:

* Mean deviation is a measure of dispersion that calculates the average of the absolute differences between each data point and the mean (or median).

- It provides a clear, easily interpretable value for the average spread of data.
- In engineering, where consistency and precision are critical, mean deviation is a valuable tool for quality control and process analysis.

Objectives of Calculation of Mean Deviation.

1. Measure of Consistency/Uniformity
   - To find how consistent or uniform a set of data values is.

   Example: In engineering, check if machine parts produced have nearly the same dimensions.
2. Understanding Variability:
   - To quantify how much the data values deviate from the central tendency.
   - Helps in comparing the spread of two or more data sets.
3. Comparative Studies:
   - To compare variations of two different groups. (i.e performance of two batches, two machines etc)
   - The smaller the MD, the more consistent the data.
4. Decision-Making in Engineering & Industry
   - To decide whether a process, product, or system is working within acceptable tolerance limits.
   - Computer Network Performance.
     Small MD → Network is stable and reliable.
   - Software Quality Testing
     Small MD → Software runs consistently
     → Reliable for deployment.

- Data science & Manchine learning:
  - MD of errors (absolute deviations from predicted values) can be used as a simple error metric.
    i.e Predicting house prices → Smaller MD means the model is giving consistent predictions close to actual values.
- Concrete cube test strength
  If MD is small → construction quality is reliable.

Definition: The mean deviation (MD) is a average of absolute deviations (ignoring the sign) of all values from a central value.

$$MD = \frac{\Sigma |x_i - A|}{N}$$

Where: x_i : data values

A : central value

N : Total number of observations

Step-by-Step Method (Ungrouped data).

1. Find the central value (A)
2. compute absolute deviation $|x_i - A|$
4. Take the sum of deviations
5. Divide by total number of observations (N)
6. Calculate the MD

Ex 1: Find the mean deviation about the mean for data : 4, 8, 6, 10, 12.

Sol^n :- central value : Mean $(\bar{x} = A) = \frac{4+8+6+10+12}{5}$

$$= \frac{40}{5} = 8$$

Deviations from mean $|x_i - 8|$ are

$= |4-8|, |8-8|, |6-8|, |10-8|, |12-8|$

$= 4, 0, 2, 2, 4$

$$MD = \frac{\Sigma |x_i - A|}{N} = \frac{4+0+2+2+4}{5} = \frac{12}{5} = 2.4$$ A.

Exe2: Voltages recorded from a transformer during Testing are: 215, 220, 218, 222, 225 (in volts) Find the MD about the mean. Also, interpret the result.

Soln: Mean $(\bar{x}) = \frac{215+220+218+222+225}{5} = 220V$

Deviations: $|215-220|, |220-220|, |218-220|, |222-220|, |225-220| = 5, 0, 2, 2, 5$

Sum of deviations $= 5+0+2+2+5 = 14$.

$M.D. = \frac{\Sigma|x_i - \bar{x}|}{N} = \frac{14}{5} = 2.8\ V$

Voltage variation is small → transformer performance is stable.

Exe3: A software engineer recorded the execution times (in milliseconds) of a program across 6 test runs: 120, 125, 118, 130, 122, 127

Find the Mean Deviation about the mean to check the consistency of program execution.

Sol: Mean $(\bar{x}) = \frac{120+125+118+130+122+127}{6} = \frac{742}{6} \approx 123.67$

Mean deviations: $|120-123.67|, |125-123.67|, |118-123.67|, |130-123.67|, |122-123.67|, |127-123.67|$

$= 3.67, 1.33, 5.67, 6.33, 1.67, 3.33$ $\quad +3.33$

Mean Deviation (MD) $= \frac{3.67+1.33+5.67+6.33+1.67}{6}$

$= \frac{21}{6} = 3.5$

The mean deviation (MD) of 3.5 ms indicates that the program execution time are fairly consistent, The software performance is stable.

Exe4: The ping times of 5 packets over a network are recorded as: 32, 35, 30, 34, 33 milisecond. Calculate the mean deviation to check network stability.

Soln: Mean $(\bar{x}) = \frac{32+35+30+34+33}{5} = \frac{164}{5}$

$= 32.8$ ms

$$\text{Mean deviation} = \frac{\Sigma |x_i - \bar{x}|}{N}$$

$$= \frac{|32-32.8| + |35-32.8| + |30-32.8| + |34-32.8| + |33-32.8|}{5}$$

$$= \frac{0.8 + 2.2 + 2.8 + 1.2 + 0.2}{5} = \frac{7.2}{5} = 1.44$$

Small mean deviation → Network latency is stable.

Exe5: The number of patients seen in the emergency ward of a hospital for a sample of 5 days in the last month was 153, 147, 151, 156 and 153. Determine the mean absolute deviation and interpret.

Sol: Mean of the patients $(\bar{x}) = \frac{153+147+151+156+153}{5}$

$= 152$

The calculation Table:

Number of Patients (x_i)	Deviation ($x_i - \bar{x}$)	Absolute Deviation $\lvert x_i - \bar{x} \rvert$
153	$153 - 152 = 1$	1
147	$147 - 152 = -5$	5
151	$151 - 152 = -1$	1
156	$156 - 152 = 4$	4
153	$153 - 152 = 1$	1
		$\Sigma \lvert x_i - \bar{x} \rvert = 12$

$$\text{Mean Deviation} = \frac{\Sigma |x_i - \bar{x}|}{N} = \frac{12}{5} = 2.4$$

$\approx$ 3 patients (approx)

The mean absolute deviation is 3 patients per day. The deviation in the number of patients' falls in the interval (152 ± 3) patients per day.

"Mean Deviation for Grouped data"

Formula $MD. = \frac{1}{N} \Sigma f_i |(x_i - \bar{x})|$

Working Steps: Let class mid points be x_i, frequencies f_i, and $N = \Sigma f_i$

- calculate Mean $\bar{x}$
- caculate $|x_i - \bar{x}|$
- calculate $f_i |x_i - \bar{x}|$ & $\Sigma f_i |x_i - \bar{x}|$
- Find $MD = \frac{1}{N} \Sigma f_i |x_i - \bar{x}|$.

Exe: Find the mean deviation from mean for the following frequency distribution of sales (₹ in thousands) in a co-operative store:

Sales	50-100	100-150	150-200	200-250	250-300	300-350
No. of days:	11	23	44	19	8	7

Solution: Calculation for MD

Sales (Rs)	Mid value (x_i)	Frequency (f_i)	$d_i=(x_i-A)/h$ ($h=50$)	$f_i d_i$	$\lvert x_i-\bar{x}\rvert$	$f\lvert(x_i-\bar{x})\rvert$
50-100	75	11	-2	-22	104.91	1154.01
100-150	125	23	-1	-23	54.91	1262.93
150-200	(175) ← A	44	0	0	4.91	216.04
200-250	225	19	1	19	45.09	856.71
250-300	275	8	2	16	95.09	760.72
300-350	325	7	3	21	145.09	1015.63
		$N=112$		$\Sigma f_i d_i = 11$		$\Sigma f_i\lvert x_i-\bar{x}\rvert = 5266.04$

Now $$\bar{x} = A + \left(\frac{\Sigma f_i d_i}{N}\right) \times h = 175 + \frac{11}{112} \times 50$$

$$= 175 + 0.0982 \times 50 = 175 + 4.910 = ₹\ 179.91 \text{ per day}$$

$$MD = \frac{\Sigma f_i |x_i - \bar{x}|}{N} = \frac{5266.04}{112} = ₹\ 47.01$$

∴ The average sales are ₹ 179.91 per day and the mean deviation of sales is ₹ 47.01 per day.

Exe Performance of electronic components

An electronics company tested a batch of 500 resistors for their resistance (in ohms). The results are shown in the table. Calculate the mean deviation to evaluate the component performance.

Resistance (ohms):	10-20	20-30	30-40	40-50	50-60
No. of Resistors:	50	120	180	100	50

Sol: Calculation Table

Resistance (in ohms)	Mid values (x_i)	No. of Resistors (f_i)	$d_i = \frac{(x_i - A)}{h}$ ($h = 10$)	$f_i d_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
10-20	15	50	-2	-100	19.60	980
20-30	25	120	-1	-120	9.60	1152
30-40	35 A	180	0	0	0.40	72
40-50	45	100	1	100	10.40	1040
50-60	55	50	2	100	20.40	1020

$N = \Sigma f_i = 500$ $\quad \Sigma f_i d_i = -20$ $\quad \Sigma f_i \lvert x_i - \bar{x} \rvert = 4264$

Now $\bar{x} = A + \frac{\Sigma f_i d_i}{N} \times h = 35 + \frac{(-20)}{500} \times 10$

$= 35 + \left(-\frac{2}{5}\right) = 35 + (-0.40) = 34.06$

$MD = \frac{\Sigma f_i \lvert x_i - \bar{x} \rvert}{N} = \frac{4264}{500} = 8.528.$

The mean deviation of 8.528 Ω provides an average measure of the variability of the resistance values in the batch of resistors.

Advantages of Mean Deviations

- Simple and easy to calculate
- Considers all data values
- Better than range as it is not based only on extremes
- Less affected by squaring of extreme values (unlike variance/SD)
- Useful in economics, social sciences & engineering
- Applicable to ungrouped discrete and grouped data.

Disadvantages of Mean Deviations

- Not widely used in advanced statistics.
- Cannot be applied in many statistical methods (correlations, regression etc.)
- Difficult to use in algebraic/statistical proofs (due to absolute values)
- Less reliable for skewed distributions.

4.5.5

Coefficient of Mean Deviation

The coefficient of mean deviation is a relative measure of dispersion which makes comparison between different data sets.

* The coefficient of mean deviation is obtained by dividing the mean deviation by a mean.

$$\text{Coefficient of MD} = \frac{\text{Mean Deviation}}{\text{Average Value (Mean)}}$$

Exe: Vibration sensor readings (mm): 2.0, 2.1, 1.9, 2.2, 2.0. Compute the coefficient of MD and interpret the result.

Sol: We have $n = 5$,

$$\text{Mean } \bar{x} = \frac{2.0 + 2.1 + 1.9 + 2.2 + 2.0}{5} = \frac{10.2}{5} = 2.04 \text{ mm}$$

$$\text{Mean deviation} = \frac{|2.0-2.04| + |2.1-2.04| + |1.9-2.04| + |2.2-2.04| + |2.0-2.04|}{5}$$

$$= \frac{0.04 + 0.06 + 0.14 + 0.16 + 0.04}{5} = \frac{0.44}{5} = 0.088 \text{ mm}$$

$$\text{CMD: Coefficient of Mean Deviation} = \frac{MD}{\bar{x}} = \frac{0.088}{2.04}$$

$$\approx 0.04314 = (4.314\%)$$

Interpretation

- Average (Mean) Vibration deviation ≈ 0.088 mm
- CMD = 4.314%
- If assembly demands < 2% variation for precision, then CMD is too high → Investigate mounting, balance or damping.
- Perform corrective maintenance.

Exe: Find out the coefficient of mean deviation in the following series:

Age:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Person:	20	25	32	40	42	35	10	8

Sol: Calculation Table

Class	x_i	f_i	$d_i = \frac{(x_i - A)}{h}$	$f_i d_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
0-10	5	20	-3	-60	31.5	630
10-20	15	25	-2	-50	21.5	537.5
20-30	25	32	-1	-32	11.5	368
30-40	35	A 40	0	0	1.5	60
40-50	45	42	1	42	8.5	357
50-60	55	35	2	70	18.5	647.5
60-70	65	10	3	30	28.5	285
70-80	75	8	4	32	38.5	308

$N = \Sigma f_i = 212$ $\quad \Sigma f_i d_i = 320$ $\quad \Sigma f_i \lvert x_i - \bar{x} \rvert = 3193$

$$\bar{x} = A + \frac{\Sigma f_i d_i}{N} \times h = 35 + \frac{320}{212} \times 10 \quad (h = UL - LL = 10 - 0 = 10)$$

$$= 35 + \frac{320}{212} = 36.5$$

$$\text{Mean deviation} = \frac{\Sigma f_i \lvert x_i - \bar{x} \rvert}{N} = \frac{3193}{212} = 15.1$$

$$\text{coefficient of Mean deviation (CMD)} = \frac{MD}{\bar{x}}$$

$$= \frac{15.1}{36.5} = 0.41$$

4.6.1 STANDARD DEVIATION:-

The standard deviation is a statistical measure that shows how much "spread" or "variability" is present in the sample.

Standard deviation is also known as root mean square deviation and is defined as

$$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$$

, $\bar{X}$ = Mean.

N = Total no of data points.

X = each data point.

$$\text{Variance} = \sigma^2$$

$$\text{or, } \sigma = \sqrt{\text{Variance}}$$

* A small S.D, high degree of uniformity of the observations as well as homogenity of the series.

* A high S.D, the greater will be the magnitude of deviation of the values from their mean, more variability or inconsistency.

Calculations of Standard Deviation

(I) Ungrouped data:-

(i) By taking deviations from actual mean.

$$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$$

$$= \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

(ii) Deviations taken from assumed mean:

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

$d = X - A$, $A \rightarrow$ Assumed mean.

Example:- Find the standard deviation from the monthly income of ten employees working in a company:

Employees	Monthly income	Employees	Monthly income
A	11,700	F	15,500
B	10,500	G	14,320
C	11,200	H	16,800
D	12,150	I	15,400
E	13,200	J	17,210

Solution:- Calculation of S.D (Actual mean)

Employees	Monthly Income (X)	$X - \bar{X}$ $(X - 13,798)$	$(X-\bar{X})^2$
A	11,700	− 2098	44,01,604
B	10,500	− 3298	1,08,76,804
C	11,200	−2598	67,49,609
D	12,150	−1648	27,15,904
E	13,200	−598	3,57,604
F	15,500	1702	28,96,804
G	14,320	522	2,72,484
H	16,800	3002	90,12,004
I	15,400	1602	25,66,404
J.	17,210	3412	1,16,41,744
$N = 10$	$\Sigma X = 1,37,980$	$\Sigma(X-\bar{X}) = 0$	$\Sigma(X-\bar{X})^2 = 5,14,90,960$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{1,37,980}{10} = 13,798$$

$$\sigma = \sqrt{\frac{\Sigma (X-\bar{X})^2}{N}} = \sqrt{\frac{5,14,90,960}{10}} = 2269.16$$

Calculation of S.D (Assumed Mean)

Employees	Monthly Income X	X − A = d. A = 13,700	d^2
A	11,700	−2000	40,00,000
B	10,500	−3,200	1,02,40,000
C	11,200	−2,500	62,50,000
D	12,150	−1550	24,02,500
E	13,200	−500	2,50,000
F	15,500	1800	32,40,000
G	14,320	620	3,84,400
H	16,800	3100	96,10,000
I	15,400	1700	28,90,000
J	17,210	3510	1,23,20,100
		$\Sigma d = 980$	$\Sigma d^2 = 5,15,87,000$

$N = 10$

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2} = \sqrt{\frac{5,15,87,000}{10} - \left(\frac{980}{10}\right)^2}$$

$$= \sqrt{51,58,700 - 9,604} = \sqrt{51,49,096}$$

$$= 2,269.16$$

* Answer on both the methods are same. If actual mean is not a whole number, assumed mean should be preferred.

Calculation of S.D — Grouped Data:

(i) By taking deviations from actual mean :—

$$\sigma = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}}$$

Or, $$\sigma = \sqrt{\frac{\Sigma f X^2}{N} - \left(\frac{\Sigma f X}{N}\right)^2}$$

(ii) Deviations taken from Assumed Mean

$$\sigma = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2} \times i$$

Example:-1:- A study of 100 engineering companies gives the following information.

Profit (₹ in crore)	0-10	10-20	20-30	30-40	40-50	50-60
No. of companies	8	12	20	30	20	10

Calculate the standard deviation of the profit earned.

Solution:-

Profit (in crore)	0-10	10-20	20-30	30-40	40-50	50-60
mid value (m)	5	15	25	(35) → A	45	55
$d = (m - A/h)$	-3	-2	-1	0	1	2

$= \frac{m-35}{10}$

No. of companies (f)	8	12	20	30	20	10
					$\Sigma f = 100$	
fd	-24	-24	-20	0	20	20
fd^2	72	48	20	0	20	40
						$(\Sigma fd^2 = 200)$

$\Sigma fd = -28$

Standard deviation $= \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$

$= \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 10$ $\quad [N = \Sigma f, \; i = 10]$

$= \sqrt{2 - 0.078} \times 10 = 13.863$

4.6.2 COMBINED STANDARD DEVIATION (Two groups):-

Combined standard deviation of two groups is denoted by σ_{12} and is computed as follows:

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

σ_{12} = combined standard deviation

σ_1 = standard deviation of group 1

σ_2 = standard deviation of group 2

$d_1 = |\bar{X}_1 - \bar{X}_{12}|$ $\quad d_2 = |\bar{X}_2 - \bar{X}_{12}|$ $\quad \bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$

Example:- For a group of 50 male workers, the mean and standard deviation of their monthly wages ₹6300 and ₹900, respectively. For a group of 40 female workers, these are ₹5400 and ₹600, respectively. Find the standard deviation of monthly wages for the combined group of workers.

Solution:-

Given that, Male workers: $N_1 = 50$

$\bar{X}_1 = 6300$, $\sigma_1 = 900$

Female workers: $N_2 = 40$, $\bar{X}_2 = 5400$

$\sigma_2 = 600$

Combined Mean $= \bar{X}_{12} = \dfrac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$

$= \dfrac{50 \times 6300 + 40 \times 5400}{50 + 40} = 5,900$

and combined standard deviation:

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2)}{N_1 + N_2}}$$

$$= \sqrt{\frac{50(8,10,000 + 1,60,000) + 40(3,60,000 + 2,50,000)}{50+40}}$$

$$= ₹900$$

where $d_1 = \overline{X}_{12} - \overline{X}_1 = 5,900 - 6300 = -400$

$d_2 = \overline{X}_{12} - \overline{X}_2 = 5,900 - 5400 = 500$

4.6.3 Coefficient of Variation :-

A relative measure is called the coefficient of variation (CV). (developed by Karl Pearson)

Uses :-

* Comparing two or more datasets expressed in different units of measurement

* Comparing datasets that are in same unit of measurement but the mean value of datasets are not same.

coefficient of variation (CV) =

$$\frac{\text{Standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

($\sigma \rightarrow$ S.D, $\bar{x}$ = Mean)

Example:- The weekly sales of two products A and B were recorded as given below:

Product A : 59 75 27 63 27 28 56

Product B : 150 200 125 310 330 250 225

Find out which of the two shows greater fluctuation in sales.

Solution:- Calculating coefficient of variation for both the products to compare fluctuation in their sales.

Product A: Let A = 56 be the assumed mean of sales for product A.

calculation of mean and standard deviation of product A:-

Sales(x)	frequency(f)	$d = x - A$ $= x - 56$	fd	fd^2
27	2	-29	-58	1682
28	1	-28	-28	784
56 → A	1	0	0	0
59	1	3	3	9
63	1	7	7	49
75	1	19	19	361
	N = 7		-57	2885

$$\bar{x} = A + \frac{\Sigma fd}{N} = 56 - \frac{57}{7} = 47.86$$

$$\sigma_A^2 = \frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2 = \frac{2885}{7} - \left(\frac{-57}{7}\right)^2$$

$$= 412.14 - 66.30 = 354.84$$

$$\sigma_A = \sqrt{354.84} = 18.59$$

$$CV(A) = \frac{\sigma_A}{\bar{x}} \times 100 = \frac{18.59}{47.86} \times 100 = 38.84$$

$\therefore CV(A) = 38.84$

(c)

Calculation of Mean and Standard deviation for product B:-

Let A = 225

Sales (x)	frequency (f)	d = x − A = x − 225	fd	fd^2
125	1	−100	−100	10,000
150	1	−75	−75	5625
200	1	−25	−25	625
225	1	0	0	0
250	1	25	25	625
310	1	85	85	7225
330	1	105	105	11,025
	N = 7		15	35,125

$$\bar{x} = A + \frac{\Sigma fd}{N} = 225 + \frac{15}{7} = 227.14$$

$$\sigma_B^2 = \frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2 = \frac{35,125}{N} - \left(\frac{15}{7}\right)^2$$

$$= 5017.85 - 4.59 = 5013.26$$

$$\sigma_B = \sqrt{5013.26} = 70.80$$

$$CV(B) = \frac{\sigma_B}{\bar{x}} \times 100 = \frac{70.80}{227.14} \times 100 = 31.17$$

Since the coefficient of variation for product A is more than that of product B, therefore the sales fluctuation on case of product is A is higher.

Example:- Suppose that the sample of polythene bags from two manufacturers A and B are tested by a prospective buyer for bursting pressure, with the following results:

Bursting pressure (lbs)	Number of bags A	B
5.0 - 9.9	2	9
10.0 - 14.9	9	11
15.0 - 19.9	29	18
20.0 - 24.9	54	32
25.0 - 29.9	11	27
30.0 - 34.9	5	13
	110	110

which sets of bags has the highest bursting pressure? which has more uniform pressure? If prices are the same, which manufacturer's bag would be preferred by the buyer? why?

Solution:-

For determining which sets of bags has the highest average bursting pressure, we need to calculate arithmetic mean and for finding out which bags should be preferred, need to calculate C.V.

Calculation of Mean and S.D for ⓐ.

Bursting Pressure	m.p X	f	$d = \frac{X - A}{i} = \left(\frac{X - 17.45}{5}\right)$	fd	fd^2
4.95 - 9.95	7.45	2	-2	-4	8
9.95 - 14.95	12.45	9	-1	-9	9
14.95 - 19.95	(17.45) A	29	0	0	0
19.95 - 24.95	22.45	54	1	54	54
24.95 - 29.95	27.45	11	2	22	44
29.95 - 34.95	32.45	5	3	15	45
		N = 110		Σfd = 78	Σfd² = 160.

$$\bar{x} = A + \frac{\Sigma fd}{N} \times i = 17.45 + \frac{78}{110} \times 5$$

$$= 17.45 + 3.55 = 21$$

$$\sigma^{ⓐ} = \sqrt{\left[\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2\right]} \times i \quad \{ i = \text{interval length} \}$$

$$= \sqrt{\frac{160}{110} - \left(\frac{78}{110}\right)^2} \times 5$$

$$= 0.976 \times 5 = 4.879$$

$$C.V(A) = \frac{\sigma}{\bar{x}} \times 100 = \frac{4.879}{21} \times 100$$

$$= 23.23\%$$

Calculation of Mean and S.D for B

Bursting Pressure (lbs.)	m.P X	f	$d=\frac{(X-17.45)}{5}$	fd	fd^2
4-95 - 9-95	7-45	9	-2	-18	36
9-95 - 14-95	12-45	11	-1	-11	11
14-95 - 19-95	17-45	18	0	0	0
19-95 - 24-95	22-45	32	1	32	32
24-95 - 29-95	27-45	27	2	54	108
29-95 - 34-95	32-45	13	3	39	117
		$N=110$		$\Sigma fd=96$	$\Sigma fd^2=304$

$$\bar{x} = A + \frac{\Sigma fd}{N} \times i = 17.45 + \frac{96}{110} \times 5$$

$$= 21.81.$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{304}{110} - \left(\frac{96}{110}\right)^2} \times 5$$

$$= 1.4149 \times 5 = 7.0745$$

$$C.V(B) = \frac{\sigma}{\bar{x}} \times 100 = \frac{7.0745}{21.81} \times 100 = 32.44\%$$

Since the average bursting pressure is higher for manufacture B, Hence the bags of manufacturer B have a higher bursting pressure. The bags of manufacturer A have more uniform pressure since the coefficient of variation is less for manufacturer A.

If the price are same, the bags of manufacturer A should be preferred by the buyer because they have more uniform pressure.

4.6.4 Applications:-

(I) Mini Project:- Application of coefficient of variation in Engineering Quality Control.-

Title:- Analysis of shaft diameter using coefficient of variation (CV).

Objective:-

* To measure the variability in manufactured shaft diameters.
* To apply statistical tool (Mean, S.D, C.V) in evaluating machine precision.
* To interpret CV as an indicator of quality control and process stability.

Problem Statement:-

In manufacturing industries, ensuring dimensional accuracy is essential for product performance and assembly. Even small variation in shaft diameters can cause issues in fitting, wear or mechanical failure. The coefficient of variation (CV) provides a normalized measure of variability that helps engineers assess the consistency of production.

Data collection

Measured diameters of 10 machined shafts (mm)

10.02, 9.98, 10.01, 9.99, 10.03, 9.97, 10.01, 10.00, 9.98, 10.00

Methodology

* Calculate the mean ($\bar{x}$) of the data
* Compute standard deviation (σ)
* Determine the coefficient of variation

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

* Interpret results in terms of process precision.

Calculations:-

* Mean = $\bar{x}$ = 9.999 mm.
* S.D = σ = 0.01814 mm.
* $CV = \frac{0.01814}{9.999} \times 100 = 0.181\%$

Results and Discussion:-

* The CV of 0.181% indicates extremely low relative variability.
* This shows the machining process is highly precise and well-controlled.
* A CV below 1% in dimensional measurements is generally considered very good in production.

Conclusions:

The study demonstrates that the coefficient of variation is a powerful tool for assessing manufacturing quality control. In this case, the low CV proves that the machining process ensures high production reliability.

Applications:

* Can be extended to evaluate material property variability (i.e., tensile strength, hardness)
* Useful in supply chain evaluation, comparing consistency of suppliers
* Applicable in civil, electrical and thermal systems where uniformity is critical.

4.7 Moments :-

Moments are statistical tools, used in statistical investigations. Moments are said to be a set of statistical parameters to measure a distribution.

4.7.1 Moments about Mean (Central Moments)

If $x_1, x_2, \ldots x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, \ldots f_n$ respectively then r^{th} moment μ_r about mean $\bar{x}$ is defined as

$$\mu_r = \frac{1}{N} \sum_{i=1}^{n} f_i (x_i - \bar{x})^r, \quad r = 0, 1, 2 \ldots$$

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

[Why :- If actual Mean $\bar{x} \in I$]

and $\bar{x} = \frac{1}{N} \Sigma f_i x_i$

If $r = 0$, $\mu_0 = \frac{1}{N} \Sigma f_i (x_i - \bar{x})^0 = 1$

If $r = 1$, $\mu_1 = \frac{1}{N} \Sigma f_i (x_i - \bar{x})^1 = 0$

If $r = 2$, $\mu_2 = \frac{1}{N} \Sigma f_i (x_i - \bar{x})^2 =$ Variance

If $r = 3$, $\mu_3 = \frac{1}{N} \Sigma f_i (x_i - \bar{x})^3$

If $r = 4$, $\mu_4 = \frac{1}{N} \Sigma f_i (x_i - \bar{x})^4$ and so on.

Note - In case of a frequency distribution with class intervals, the values of x are the midpoints of the interval.

Example Calculate first four moments for the following frequency distribution:-

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	1	6	10	15	11	7

Solution

Marks	No. of students f	Mid value x	fx	$f(x-\bar{x})$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
0-10	1	5	5	-30	900	-27000	810000
10-20	6	15	90	-120	2400	-48000	960000
20-30	10	25	250	-100	1000	-1000	100000
30-40	15	35	525	0	0	0	0
40-50	11	45	495	110	1100	11000	11000
50-60	7	55	385	140	2800	56000	112000
	$\Sigma f = 50$		$\Sigma fx = 1750$	0	8200	-18000	3100000

Mean $\bar{x} = \frac{1}{N}\Sigma fx = \frac{1}{50} \times 1750 = 35$

$$\mu_1 = \frac{1}{N}\Sigma f(x-\bar{x}) = \frac{1}{50} \times 0 = 0$$

Variance $\mu_2 = \frac{1}{N}\Sigma f(x-\bar{x})^2 = \frac{1}{50} \times 8200 = 164$

Third Central Moment

$$\mu_3 = \frac{1}{N}\Sigma f(x-\bar{x})^3 = \frac{1}{50} \times (-18000) = -360$$

Fourth Central Moment

$$\mu_4 = \frac{1}{N}\Sigma f(x-\bar{x})^4 = \frac{1}{50} \times 3100000 = 62000$$ Ans.

Why:- If $\bar{x} \notin I$ then let assumed mean $= a$

4.7.2 Moments about any number (Raw Moments):-

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - a)^r, \quad r = 0, 1, 2 \text{ ——}$$

where "a" is any number and $N = \sum_{i=1}^{n} f_i$ = Sum of frequencies

for $r = 0$, $\mu_0' = \frac{1}{N}\Sigma f_i (x_i - a)^0 = 1$

for $r = 1$, $\mu_1' = \frac{1}{N}\Sigma f_i (x_i - a)^1 = \bar{x} - a$

for $r = 2$, $\mu_2' = \frac{1}{N}\Sigma f_i (x_i - a)^2$

for $r = 3$, $\mu_3' = \frac{1}{N}\Sigma f_i (x_i - a)^3$

for $r = 4$, $\mu_4' = \frac{1}{N}\Sigma f_i (x_i - a)^4$ and so on

Note- To ease our calculation work by defining $u = \frac{x - a}{h}$, we have

$$\mu_r' = \frac{1}{N}\left[\Sigma f_i u_i^r\right] h^r, \quad r = 0, 1, 2 \text{ ——}$$

Example Calculate the variance and third central moment from the following data:-

x	0	1	2	3	4	5	6	7	8
f	1	9	26	59	72	52	29	7	1

Solution Mean $\bar{x} = \frac{1}{N}\Sigma f_i x_i = 3.97$

Let assumed mean $x = a = 4$

x_i	f_i	$x_i - 4$	$f_i(x_i-4)$	$f_i(x_i-4)^2$	$f_i(x_i-4)^3$
0	1	-4	-4	16	-64
1	9	-3	-27	81	-243
2	26	-2	-52	104	-208
3	59	-1	-59	59	-59
4	72	0	0	0	0
5	52	1	52	52	52
6	29	2	58	116	232
7	7	3	21	63	189
8	1	4	4	16	64
	256		-7	507	-37

$$\mu_1' = \frac{1}{\Sigma f_i} \Sigma f_i (x_i - 4) = \frac{1}{256}(-7) = -0.02734$$

$$\mu_2' = \frac{1}{\Sigma f_i} \Sigma f_i (x_i - 4)^2 = \frac{1}{256}(507) = 1.9805$$

$$\mu_3' = \frac{1}{\Sigma f_i} \Sigma f_i (x_i - 4)^3 = \frac{1}{256}(-37) = -0.1445$$

$\mu_1 = 0$

Variance $\mu_2 = \mu_2' - (\mu_1')^2 = 1.97975$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 0.0178997$$

4.7.3

Relation between μ_r and μ_r' :-

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$\mu_4 = \mu_4' - 4\mu_3'(\mu_1') + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

Why:- If assumed mean $a = 0$

4.7.4 Moment about the origin :-

$$\nu_r = \frac{1}{N} \sum_{i=1}^{n} f_i x_i^r$$

If $r = 0$, $\nu_0 = \frac{1}{N} \Sigma f_i x_i^0 = 1$

If $r = 1$, $\nu_1 = \frac{1}{N} \Sigma f_i x_i = \bar{x}$

If $r = 2$, $\nu_2 = \frac{1}{N} \Sigma f_i x_i^2$ and so on.

4.7.5 Relation between μ_r and ν_r :-

$$\nu_1 = \bar{x}$$

$$\nu_2 = \mu_2 + (\bar{x})^2$$

$$\nu_3 = \mu_3 + 3\mu_2(\bar{x}) + (\bar{x})^3$$

$$\nu_4 = \mu_4 + 4\mu_3(\bar{x}) + 6\mu_2(\bar{x})^2 + (\bar{x})^4$$

4.7.6 Karl Pearson's β and γ coefficients :-

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\gamma_1 = +\sqrt{\beta_1} \quad \text{and} \quad \gamma_2 = \beta_2 - 3$$

Example - In a certain distribution, the first four moments about the point $x = 4$ are -1.5, 17, -30, 308. Find the moments about mean and about origin. Also calculate β_1 and β_2.

Solution :- We have $A = 4$,

$$\mu_1' = -1.5, \quad \mu_2' = 17, \quad \mu_3' = -30, \quad \mu_4' = 308$$

Moments about mean :- $\mu_1 = 0$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 17 - (-1.5)^2 = 14.75$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 39.75$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 = 342.3125$$

Moments about origin :-

$$\nu_1 = \bar{x} = \mu_1' + A = -1.5 + 4 = 2.5$$

$$\nu_2 = \mu_2 + (\bar{x})^2 = 14.75 + (2.5)^2 = 21$$

$$\nu_3 = \mu_3 + 3\mu_2(\bar{x}) + (\bar{x})^3 = 166$$

$$\nu_4 = \mu_4 + 4\mu_3(\bar{x}) + 6\mu_2(\bar{x})^2 + (\bar{x})^4 = 1332$$

Karl Pearson's β coefficients :-

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.492377, \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = 1.573398$$

Example :- The first three moments of a distribution about the value "2" of the variable are 1, 16 and -40. Show that the mean is 3, Variance is 15 and $\mu_3 = -86$.

Solution:- We have

$$a = 2,\ \mu_1' = 1,\ \mu_2' = 16 \text{ and } \mu_3' = -40$$

$$\mu_1' = \bar{x} - a \Longrightarrow \bar{x} = \mu_1' + a = 1 + 2 = 3$$

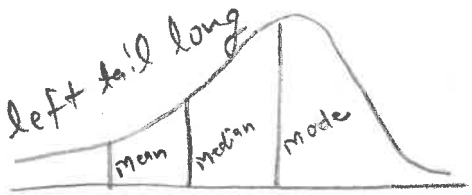
$$\text{Variance } \mu_2 = \mu_2' - (\mu_1')^2 = 16 - (1)^2 = 15$$

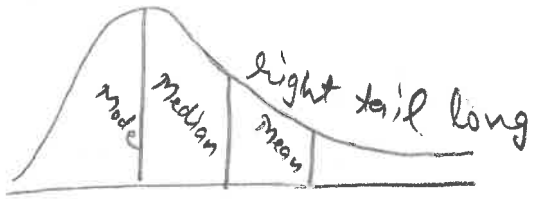
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = -40 - 48 + 2 = -86$$

4.8 Skewness :-

Skewness means lack of Symmetry. It indicates whether the curve is turned more to one side than to other i.e whether the curve has a longer tail on one side. Skewness can be positive as well as negative.

(a) Negatively skewed distribution :- Here left tail is longer than the right tail.





(b) Positively skewed distribution :- The right tail of the curve will be longer than the left.

Note :- In skew distribution mean, median and mode are not equal.

4.8.1 Tests of skewness :-

(1) There is no skewness in the distribution if A.M = mode = Median

(2) There is no skewness in the distribution if

Sum of frequencies which are less than mode

= Sum of freq. which are greater than mode

(3) The distribution is negatively skewed if A.M. is less than Mode.

(4) The curve is not symmetrical if

A.M ≠ Median ≠ Mode

4.8.2

Moments coefficients of skewness :-

$$\text{Moment coeff of skewness} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \pm\sqrt{\beta_1} = \gamma_1$$

For a symmetrical distribution, its value would come out to be zero.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

The sign of $\sqrt{\beta_1}$ is to be taken as that of μ_3.

Example Calculate the moment coefficient of skewness for the following distribution:-

Classes	2.5 – 7.5	7.5 – 12.5	12.5 – 17.5	17.5 – 22.5	22.5 – 27.5	27.5 – 32.5	32.5 – 37.5
frequency	8	15	20	32	23	17	5

Solution:- $\bar{x} = \frac{\Sigma fx}{\Sigma f}$, Let $A = 20$, $h = 5$

Classes	f	Mid x	$u = \frac{x-A}{h}$	fu	fu^2	fu^3
2.5–7.5	8	5	−3	−24	72	−216
7.5–12.5	15	10	−2	−30	60	−120
12.5–17.5	20	15	−1	−20	20	−20
17.5–22.5	32	20	0	0	0	0
22.5–27.5	23	25	1	23	23	23
27.5–32.5	17	30	2	34	68	136
32.5–37.5	5	35	3	15	45	135
	120			−2	288	−62

$$\mu_1' = \left(\frac{\Sigma fu}{N}\right)h = \frac{-2}{120}(5) = -0.083$$

$$\mu_2' = \left(\frac{\Sigma fu^2}{N}\right)h^2 = \frac{288}{120}(5)^2 = 60$$

$$\mu_3' = \left(\frac{\Sigma fu^3}{N}\right)h^3 = \frac{-62}{162}(5)^3 = -64.583$$

Now $\mu_2 = \mu_2' - (\mu_1')^2 = 59.993$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 = -49.644$$

$\therefore$ Moment coeff. of skewness $= \dfrac{\mu_3}{\sqrt{\mu_2^3}} = \dfrac{-49.644}{\sqrt{(59.993)^3}}$

$$= -0.1068$$

Example The first three central moments of a distribution are 0, 15, −31. Find the moment coefficient of skewness.

Solution:- we have

$$\mu_1 = 0, \quad \mu_2 = 15, \quad \mu_3 = -31$$

Moment coeff of skewness $= \dfrac{\mu_3}{\sqrt{\mu_2^3}}$

$$= \frac{-31}{\sqrt{(15)^3}} = -0.53$$

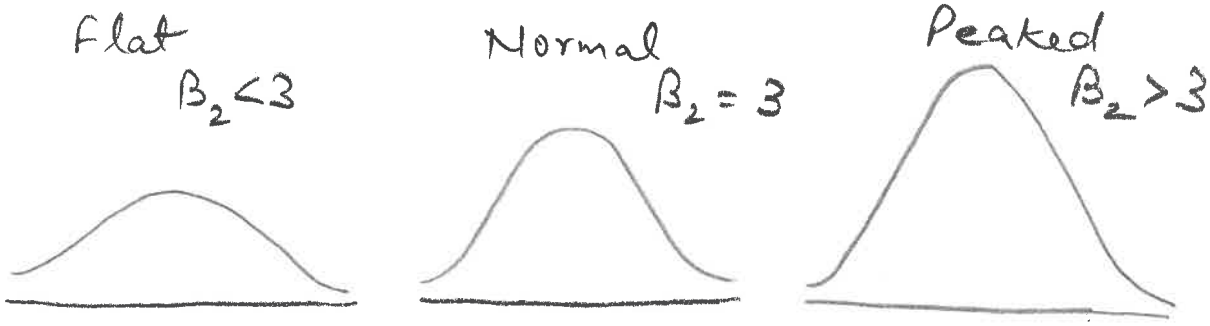
4.9 Kurtosis :-

The relative flatness or peakedness of the top is called kurtosis and is measured by β_2

4.9.1 Measure of Kurtosis :-

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

The kurtosis of a distribution is also measured by using Greek letter "γ_2" which is defined as $\gamma_2 = \beta_2 - 3$



Platykurtic	Mesokurtic	Leptokurtic
$\gamma_2 < 0$	$\gamma_2 = 0$	$\gamma_2 > 0$

Example :- Find out the kurtosis of the date

Class interval	0–10	10–20	20–30	30–40
Frequency	1	3	4	2

Solution :- $\bar{x} = \frac{1}{N} \Sigma f_i x_i$, $N = \Sigma f_i$

Let assumed mean $A = 25$

Class	Frequency f_i	mid value x_i	$f_i(x_i - A)$	$f(x-A)^2$	$f(x-A)^3$	$f(x-A)^4$
0-10	1	5	-20	-8000	-8000	160000
10-20	3	15	-30	-300	-3000	30000
20-30	4	25	0	0	0	0
30-40	2	35	20	200	2000	20000
			-30	900	-9000	210000

$$\mu_1' = \frac{1}{\Sigma f_i} \Sigma f_i (x_i - A) = \frac{-30}{10} = -3$$

$$\mu_2' = \frac{1}{\Sigma f_i} \Sigma f_i (x_i - A)^2 = \frac{900}{10} = 90$$

$$\mu_3' = \frac{1}{\Sigma f} \Sigma f (x - A)^3 = \frac{-9000}{10} = -900$$

$$\mu_4' = \frac{1}{\Sigma f} \Sigma f (x - A)^4 = \frac{210000}{10} = 21000$$

Now $\mu_2 = \mu_2' - (\mu_1')^2 = 90 - 9 = 81$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 14817$$

$$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} = 2.258$$

$$\gamma_2 = \beta_2 - 3 = -0.742$$

Example:- The first four moments of a distribution about $x=4$ are $1, 4, 10, 45$. Obtain the various characteristics of the distribution on the basis of given information. Comment upon the nature of the distribution.

Solution:- $A=4$, $\mu_1' = 1$, $\mu_2' = 4$, $\mu_3' = 10$, $\mu_4' = 45$

$$\mu_1 = 0, \quad \mu_2 = \mu_2' - (\mu_1')^2 = 3$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 = 26$$

Skewness:- $\gamma_1 = \dfrac{\mu_3}{\sqrt{\mu_2^3}} = 0$

$\therefore$ The distribution is symmetrical.

Kurtosis:- $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 2.89 < 3$

$\therefore$ The distribution is platykurtic.

4.10 Real-World Applications of Moments, Skewness & Kurtosis :-

Let's use a concrete real-world example from manufacturing (quality control) – a field where moment analysis is frequently used.

Quality control in Manufacturing Bolts :-

A factory manufactures bolts and checks the length of bolts (in mm) produced daily. They record how many bolts fall into specific length intervals :-

Length (mm) x_i	Frequency f_i
48	5
49	12
50	20
51	10
52	3

How the first ~~m~~ four moments Control the Quality of bolt production.

Solution :- Total frequency $N = \Sigma f_i = 50$

Step-1 $\quad \text{Mean} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{2494}{50} = 49.88 \text{ mm}$

* The average bolt length is 49.88 mm, close to 50 mm (the target)

Step-2 Variance (2^{nd} central Moment) $\rightarrow$

$$\mu_2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{\Sigma f_i} = \frac{53.14}{50} = 1.06$$

$$S.D = \sqrt{\text{var}} \simeq 1.03 \text{ mm}$$

* This measures the spread of bolt lengths.

Step-3 Skewness (using 3^{rd} central Moment) $\rightarrow$

$$\mu_3 = \frac{\Sigma f_i (x_i - \bar{x})^3}{\Sigma f_i} = \frac{1.47}{50} = 0.029$$

$$\text{Skewness} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{0.029}{(1.06)^{3/2}} = 0.027 \simeq 0$$

* The distribution is nearly symmetric.
* No series bias in short/long bolts).

Step-4 Kurtosis (using 4th central Moment)

$$\mu_4 = \frac{\Sigma f_i (x_i - \bar{x})^4}{\Sigma f_i} = \frac{146.0}{50} = 2.92$$

$$\text{Kurtosis} = \frac{\mu_4}{\mu_2^2} = \frac{2.92}{(1.06)^2} = 2.6 < 3$$

* slightly platykurtic (flatter than Normal)
* lower kurtosis shows fewer extreme outliers.

Interpretation for Quality Control:-

(1) Mean (first moment about origin) → Ensures average bolt length ≈ 50 mm (on target)

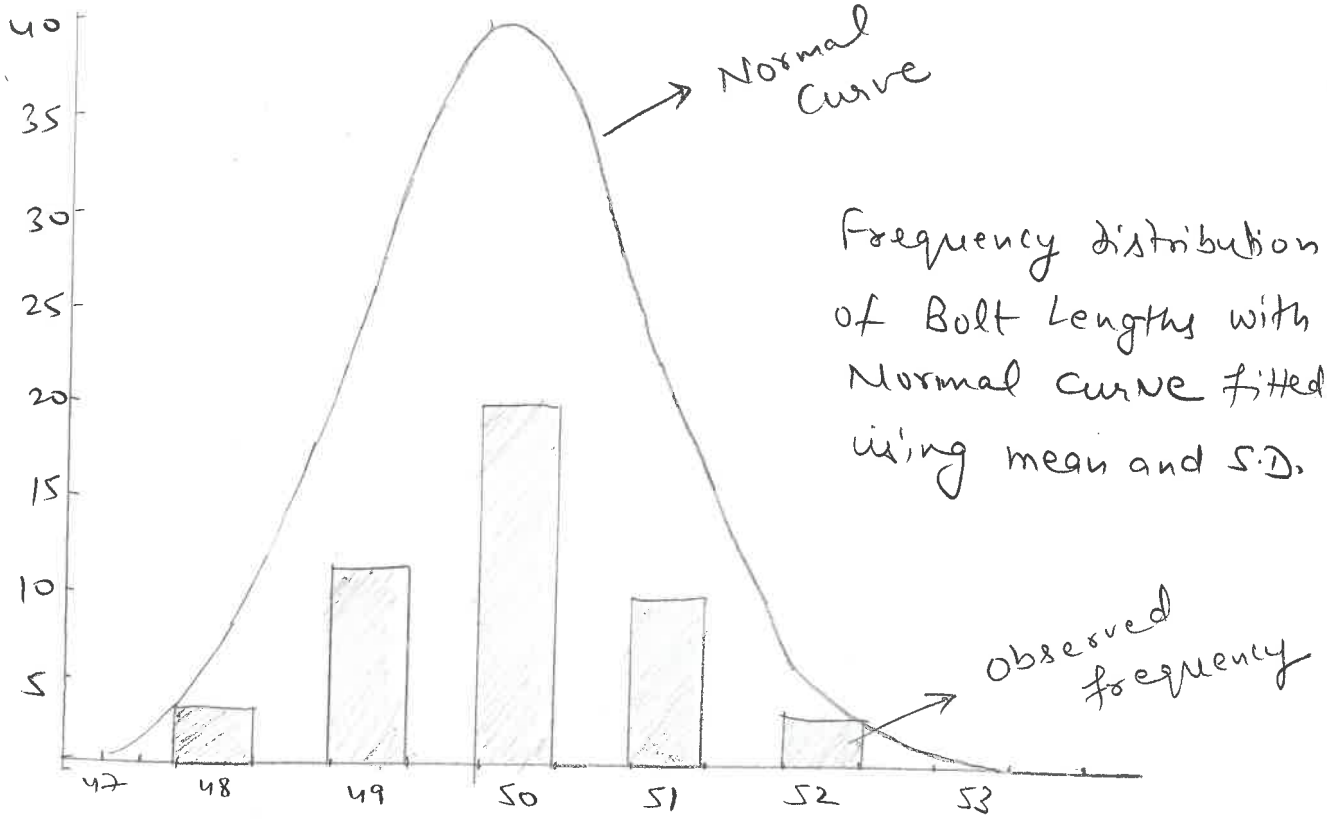
(2) Variance (second moment) → Measures Variability, small variance means most bolts are close to the target.

(3) Skewness (Third Moment) → Indicates Symmetry, near zero skewness means no bias toward short or long bolts.

(4) Kurtosis (Fourth moment) → Indicates shape of distribution, lower kurtosis shows

fewer extreme outliers.

Conclusion → Together, these four moments help monitor and control consistency, accuracy and reliability in bolt manufacturing.



Frequency distribution of Bolt Lengths with Normal curve fitted using mean and S.D.

* Most bolts are around 49-50 mm.
* Very few bolts are at the extremes (48 & 52 mm)

This visualization reinforces what moments analysis showed:-

(1) Average length is on target,
(2) Variability is low
(3) Distribution is symmetric.
(4) Few extreme deviations.

EXERCISE 4.1

Ques1. Statistics plays a vital role in engineering & technology fields, explain

(a) What is statistics and the reasons why engineering students need to learn statistics and develop statistical thinking & analysis?

(b) Discuss the importance and scope of statistics also highlight the major limitations.

Ques2. Explain the need for data in statistical analysis and the types of data.

Ques3. Describe the different methods of collecting the data with suitable examples for engineering applications.

Ques4. What is the case study method of data collection? Illustrate its importance with an example.

Ques 5. Explain inclusive and exclusive methods of classification with suitable examples.

Ques 6. Construct a frequency distribution of the following marks (out of 50) using (a) Inclusive method.
(b) Exclusive method.

Data : 5, 12, 18, 20, 25, 29, 33, 35, 40, 42, 48

Ques 7. The following data relate to area in millions of square kilometer of oceans of the world.

Ocean :	Pacific	Atlantic	Indian	Antarctic	Arctic
Area : (Million sq.km.)	70.8	41.2	28.5	7.6	4.8

Ques 8. The rate of increase in population of a country during the last three decades is 5%, 8%, 12%. Find the average rate of growth during last three decades.

Ans: 108.2

Ques 9. A given machine is assumed to depreciate 40% in value in the 1st year, 25% in 2nd year & 10% per year for the next three years, each percentage being calculated on the diminishing value. What is the average depreciation recorded on the diminishing value for the period of five years?

Ans: 15.85%

Ques 10. Find the harmonic mean of the following data:

Dividend yield (percent) :	2-6	6-10	10-14
No. of companies :	10	12	18

Ques 11. The following are the prices of shares of a company from Monday to Saturday:

Days:	Mon.	Tues.	Wed.	Thur.	Fri.	Sat.
Price (Rs):	200	210	208	160	220	250

Calculate the range & its coefficient.

Ans: Range = 90 rs.
Coeff. of range
= 0.219.

Ques 12. The following sample shows the weekly number of roads accidents in a city during a two year period:

Number of Accidents :	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44
frequency :	5	12	32	27	11	9	4	3	1

Determine coefficient of quartile deviation.

Ans: Q.D = 0.561

Ques 13. Calculate mean deviation for the following data:

Class Interval :	0-10	10-20	20-30	30-40	40-50
freq. :	5	9	12	7	3

Ans. 9.259

Ques 14. The weekly sales of two products A & B were recorded as given below:

Product A :	59	75	27	63	27	28	56
Product B :	150	200	125	310	330	250	225

Find out which of the two shows greater fluctuation in sales.

Ans. C.V of A is more than B.

Ques 15. From the analysis of monthly wages paid to employees in two service organizations X & Y, the following results were obtained:

	Organization X	Organization Y
Number of wage-earners	550	650
Average monthly wages	5000	4500
Variance of distribution of wages	900	1600

In which organization is there greater variability in individual wages of all the wage earners taken together?

Ans: $\sigma_{12} = 251.68$

Ques16: The first four moments of a distribution, about the value '35' are −1.8, 240, −1020 & 144000. Find $\mu_1, \mu_2, \mu_3, \mu_4$.

Ans: $\mu_4 = 0$, $\mu_2 = 234.76$, $\mu_3 = 264.36$
$\mu_4 = 141290.11$.

Ques17: The first three moments of a distribution about value '2' of the variable are 1, 16, −40 respectively. Find the first three moments about origin.

Ans: $\nu_1 = 3$, $\nu_2 = 24$, $\nu_3 = 76$.

Ques18: For a distribution, the mean is 10, variance is 16, γ_1 is 1, β_2 is 4. Find the first four moments about origin.

Ans: $\nu_1 = 10$, $\nu_2 = 116$, $\nu_3 = 1544$, $\nu_4 = 23184$

Ques19: Calculate β_1 & β_2 from the following distribution,

x :	0	1	2	3	4	5	6	7	8
f :	1	8	28	56	70	56	28	8	1

Ans: $\beta_1 = 0$
$\beta_2 = 2.75$

Ques20: Calculate first four moments about mean for the following data :

Class-Interval :	0–10	10–20	20–30	30–40	40–50
f :	10	20	40	20	10

Ans: $\mu_4 = 0$, $\mu_2 = 120$, $\mu_3 = 0$
$\mu_4 = 36000$.

Ques21: The first three central moments of a distribution are 0, 2.5, 0.7. Find the value of moment coefficient of skewness.

Ans: 0.17708

Ques22: The first three moments of a frequency distribution about value '5' are −0.55, 4.46 & −0.43. Find moment coefficient of skewness.

Ans: 0.7781

Ques 23: Calculate the moment coefficient of skewness for the following data:

Marks:	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students:	8	12	20	30	15	10	5

Ans: 0.0726

Ques 24: The first four moments about mean of a frequency distribution are 0, 100, -7 and 35000. Discuss the kurtosis.

Ans: Leptokurtic

Ques 25: The first four moments of a distribution about $x=4$ are 1, 4, 10 and 45. Obtain the various characteristics of the distribution on the basis of the given information. Comment upon the nature of the distribution.

Ans: $\beta_1 = 0$, $\beta_2 < 3$.