

# **Advanced Analytics using Statistics PG-DBDA March 24**

## **Session 1 & 2**

**Analytics** is the systematic exploration and analysis of data to uncover meaningful patterns, insights, and trends that can inform decision-making and drive improvements in various aspects of business, science, and other domains. It involves the use of statistical, mathematical, and computational techniques to extract actionable insights from data.

**Understanding Data:** Analytics begins with understanding the data available. This includes both structured data (such as databases and spreadsheets) and unstructured data (such as text documents, images, and videos). The data can come from various sources, including business transactions, customer interactions, sensors, social media, and more.

**Data Preparation:** Before analysis can begin, the raw data often needs to be cleaned, transformed, and formatted to make it suitable for analysis. This process involves tasks such as removing duplicates, handling missing values, standardizing formats, and integrating data from different sources.

**Descriptive Analytics:** Descriptive analytics focuses on summarizing and describing the characteristics of the data. This may involve generating summary statistics, visualizing data through charts and graphs, and exploring relationships between variables. Descriptive analytics helps stakeholders understand what has happened in the past and provides context for further analysis.

**Diagnostic Analytics:** Diagnostic analytics aims to understand why certain events occurred by identifying patterns and correlations in the data. It involves digging deeper into the data to uncover the root causes of observed phenomena or trends. Diagnostic analytics often involves hypothesis testing, correlation analysis, and causal inference techniques.

**Predictive Analytics:** Predictive analytics leverages historical data to forecast future outcomes or trends. This involves building statistical or machine learning models that can make predictions based on patterns observed in the data. Predictive analytics can be used for various purposes, such as sales forecasting, customer churn prediction, risk assessment, and demand forecasting.

**Prescriptive Analytics:** Prescriptive analytics goes beyond prediction to recommend actions or decisions that can optimize outcomes. This involves using optimization and simulation techniques to explore different scenarios and identify the best course of action given specific constraints and objectives. Prescriptive analytics helps organizations make data-driven decisions to improve efficiency, minimize risks, and maximize outcomes.

**Continuous Improvement:** Analytics is an iterative process that requires continuous improvement and refinement. As new data becomes available and business conditions change, analytics models and strategies need to be updated and adapted accordingly. Organizations should establish feedback loops to incorporate insights gained from analytics into decision-making processes and drive ongoing improvement.

Overall, analytics enables organizations to leverage data as a strategic asset to gain competitive advantage, improve operational efficiency, enhance customer experiences, and

drive innovation. By harnessing the power of analytics, businesses and other entities can make smarter decisions and achieve better outcomes in an increasingly data-driven world.

### **Data analytics Life Cycle**

The data analytics lifecycle is a structured approach to extracting insights and value from data. It typically consists of several interconnected stages that guide the process from defining the problem to implementing solutions. Here's a breakdown of the data analytics lifecycle:

#### **Problem Definition:**

- Identify the business problem or opportunity that analytics can address.
- Define clear objectives and key performance indicators (KPIs) to measure success.
- Ensure alignment with organizational goals and stakeholder needs.

#### **Data Collection:**

- Identify relevant data sources both internal and external to the organization.
- Gather data from databases, spreadsheets, files, APIs, sensors, social media, etc.
- Ensure data quality, completeness, and relevance for analysis.

#### **Data Preparation:**

- Cleanse the data by removing duplicates, correcting errors, and handling missing or inconsistent values.
- Transform the data into a suitable format for analysis (e.g., normalization, aggregation, or feature engineering).
- Integrate data from multiple sources if necessary.

#### **Exploratory Data Analysis (EDA):**

- Explore the dataset to understand its structure, distribution, and relationships.
- Visualize data using charts, graphs, and statistical summaries.
- Identify patterns, trends, outliers, and potential insights.

#### **Feature Engineering:**

- Select, create, or transform features that are relevant and predictive for the analysis.
- Apply techniques such as dimensionality reduction, encoding categorical variables, or deriving new features.

#### **Modeling:**

- Select appropriate analytical techniques or algorithms based on the problem and data characteristics.
- Split the data into training, validation, and testing sets.
- Train machine learning or statistical models using the training data.
- Tune hyperparameters and evaluate model performance using validation data.
- Validate the model's performance on unseen data using the testing set

**Interpretation and Evaluation:**

- Interpret the model results in the context of the problem and business objectives.
- Evaluate the model's performance using relevant metrics (e.g., accuracy, precision, recall, or AUC).
- Assess the impact of the analytics solution on the business problem and its alignment with KPIs.

**Deployment:**

- Deploy the analytics solution into production or operational systems.
- Integrate the model into decision-making processes or business workflows.
- Monitor the model's performance in real-world scenarios and collect feedback for continuous improvement.







**Monitoring and Maintenance:**

- Establish monitoring mechanisms to track the performance and behavior of the deployed model.
- Monitor data quality, model drift, and other relevant metrics over time.
- Retrain or update the model periodically with new data to ensure relevance and accuracy.

**Iterative Improvement:**

- Continuously refine and improve the analytics solution based on feedback, changing business requirements, and new data.
- Iterate through the lifecycle stages as needed to address evolving challenges and opportunities.

By following the data analytics lifecycle, organizations can systematically leverage data to derive actionable insights, make informed decisions, and drive business value.

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
					
<b>Business Issue Understanding</b>	<b>Data Understanding</b>	<b>Data Preparation</b>	<b>Exploratory Analysis and Modeling</b>	<b>Validation</b>	<b>Visualization and Presentation</b>
Define business objectives	Collect initial data	Gather data from multiple sources	Develop methodology	Evaluate results	Communicate results
Gather required information	Identify data requirements	Cleanse	Determine important variables	Review process	Determine best method to present insights based on analysis and audience
Determine appropriate analysis method	Determine data availability	Format	Build model	Determine next steps	Results are valid → proceed to step 6
Clarify scope of work	Explore data and characteristics	Blend	Assess model	Results are invalid ← revisit steps 1-4	Craft a compelling story
Identify deliverables		Sample			Make recommendations

## Session 3 & 4

### **Prerequisites:**

#### **Factorial Notation:**

We define,  $0! = 1$

For any positive integer  $n$ ,

$$n! = n \times (n - 1) \times (n - 2) \times \dots \times 1$$

e.g.

$$1! = 1$$

$$2! = 2 \times 1$$

$$3! = 3 \times 2 \times 1$$

$$4! = 4 \times 3 \times 2 \times 1 \text{ and so on.}$$

Consider  $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$ , which we can write as

$$6! = 6 \times (5 \times 4 \times 3 \times 2 \times 1) = 6 \times 5! \therefore n! = n \times [(n - 1)!]$$

$$6! = 6 \times 5 \times (4 \times 3 \times 2 \times 1) = 6 \times 5 \times 4!$$

$$\therefore n! = n \times (n - 1) \times [(n - 2)!] \text{ and so on.}$$

#### **Permutation:**

Consider selection of 'r' objects out of  $n$  ( $r \leq n$ ). If the **order** in which the objects are selected **is important** then such a selection is called as a **Permutation**.

The number of such permutations is denoted by  ${}^n P_r$  and

$${}^n P_r = \frac{n!}{(n - r)!}$$

e.g.

$${}^n P_0 = \frac{n!}{(n - 0)!} = 1$$

$${}^n P_1 = \frac{n!}{(n - 1)!} = \frac{n(n - 1)!}{(n - 1)!} = n$$

$${}^n P_n = \frac{n!}{(n - n)!} = n!$$

**Combination:**

Consider selection of 'r' objects out of n ( $r \leq n$ ). If the **order** in which the objects are selected **is not important** then such a selection is called as a **Combination**.

The number of such combinations is denoted by  ${}^nC_r$  and

$${}^nC_r = \frac{n!}{r!(n-r)!}$$

e.g.

$${}^nC_0 = \frac{n!}{0!(n-0)!} = \frac{n!}{n!} = 1$$

$${}^nC_1 = \frac{n!}{1!(n-1)!} = \frac{n(n-1)!}{(n-1)!} = n$$

$${}^nC_n = \frac{n!}{1!(n-n)!} = \frac{n!}{n!0!} = 1$$

$${}^{10}C_3 = \frac{10!}{3!(10-3)!} = \frac{10!}{3!7!}$$

$${}^{10}C_7 = \frac{10!}{7!(10-7)!} = \frac{10!}{7!3!}$$

$$\therefore {}^{10}C_3 = {}^{10}C_7 \text{ i.e. } {}^{10}C_3 = {}^{10}C_{10-3}$$

<b>In general, <math>{}^nC_r = {}^nC_{n-r}</math></b>
---

$${}^{10}C_3 = \frac{10!}{3!7!} = \frac{10*9*8*7!}{3!7!} = \frac{10*9*8}{3!}; \quad {}^{12}C_4 = \underline{\hspace{2cm}}$$

$${}^{100}C_{97} = {}^{100}C_3 = \underline{\hspace{2cm}}$$

## **Basic Terms:**

### **Random Experiment:**

Consider an action which is repeated under essentially identical conditions. If it results in any one of the several possible outcomes, but it is not possible to predict which outcome will appear, then such an action is called as a Random Experiment

A random experiment is defined as an experiment whose outcome cannot be predicted with certainty

An activity that produces a result or an outcome is called an experiment. It is an element of uncertainty as to which one of these occurs when we perform an activity or experiment. Usually, we may get a different number of outcomes from an experiment. However, when an experiment satisfies the following two conditions, it is called a random experiment.

- (i) It has more than one possible outcome.
- (ii) It is not possible to predict the exact outcome in advance.

### **Outcome**

A possible result of random experiment is called a possible outcome of the experiment.

### **Sample Space:**

The set of all possible outcomes of a random experiment is called the sample space. The sample space is denoted by  $S$  or Greek letter omega ( $\Omega$ ). The number of elements in  $S$  is denoted by  $n(S)$ . A possible outcome is also called a sample point since it is an element in the sample space.

All the elements of the sample space together are called as 'exhaustive cases'.

### **Event:**

Any subset of the sample space is called as an 'Event' and is denoted by any capital letter like  $A$ ,  $B$ ,  $C$  or  $A_1$ ,  $A_2$ ,  $A_3$ ,..

### **Favourable cases:**

The cases which ensure the happening of an event  $A$ , are called as the cases favourable to the event  $A$ . The number of cases favourable to event  $A$  is denoted by  $n(A)$ .

## **Types of Events**

**Elementary Event:** An event consisting of a single outcome is called an elementary event.

**Certain Event:** The sample space is called the certain event if all possible outcomes are favourable outcomes. i.e. the event consists of the whole sample space.

**Impossible Event:** The empty set is called impossible event as no possible outcome is favorable

### **Union of Two Events**

Let A and B be two events in the sample space S. The union of A and B is denoted by  $A \cup B$  and is the set of all possible outcomes that belong to at least one of A and B.

Let S = Set of all positive integers not exceeding 50;

Event A = Set of elements of S that are divisible by 6;

Event B = Set of elements of S that are divisible by 9.

$$A = \{6, 12, 18, 24, 30, 36, 42, 48\}$$

$$B = \{9, 18, 27, 36, 45\}$$

$\therefore A \cup B = \{6, 9, 12, 18, 24, 27, 30, 36, 42, 45, 48\}$  is the set of elements of S that are divisible by 6 or 9.

### **Exhaustive Events**

Two events A and B in the sample space S are said to be exhaustive if  $A \cup B = S$

### **Intersection of Two Events**

Let A and B be two events in the sample space S.

The intersection of A and B is the event consisting of outcomes that belong to both the events A and B.

Let S = Set of all positive integers not exceeding 50,

Event A = Set of elements of S that are divisible by 3,

Event B = Set of elements of S that are divisible by 5.

$$\text{Then } A = \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45, 48\},$$

$$B = \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$$

$\therefore A \cap B = \{15, 30, 45\}$  is the set of elements of S that are divisible by both 3 and 5.

### **Mutually Exclusive Events**

Event A and B in the sample space S are said to be mutually exclusive if they have no outcomes in common. ( $A \cap B = \phi$ ). In other words, the intersection of mutually exclusive events is empty. Mutually exclusive events are also called disjoint events.

If two events A and B are mutually exclusive and exhaustive, then they are called **Complementary events**.

### **Equally Likely Cases:**

Cases are said to be equally likely if they all have the same chance of occurrence i.e. no case is preferred to any other case.



## **Probability Introduction**

Chance is the occurrence of events in the absence of any obvious intention or cause. It is, simply, the possibility of something happening. When the chance is defined in Mathematics, it is called probability.

Probability is the extent to which an event is likely to occur, measured by the ratio of the favourable cases to the whole number of cases possible.

Mathematically, the probability of an event occurring is equal to the ratio of a number of cases favourable to a particular event to the number of all possible cases.

The theoretical probability of an event is denoted as  $P(E)$ .

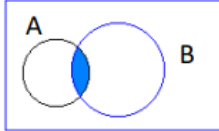

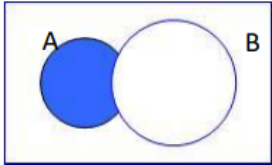
$$P(E) = \frac{\text{Number of Outcomes Favourable to E}}{\text{Number of all Possible Outcomes of the Experiment}}$$

## **Importance of Probability**

The concept of probability is of great importance in everyday life. Statistical analysis is based on this valuable concept. Infact the role played by probability in modern science is that of a substitute for certainty.

The following discussion explains it further:

- i. The probability theory is very much helpful for making prediction. Estimates and predictions form an important part of research investigation. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.
- ii. It has also immense importance in decision making.
- iii. It is concerned with the planning and controlling and with the occurrence of accidents of all kinds.
- iv. It is one of the inseparable tools for all types of formal studies that involve uncertainty.
- v. The concept of probability is not only applied in business and commercial lines, rather than it is also applied to all scientific investigation and everyday life.
- vi. Before knowing statistical decision procedures one must have to know about the theory of probability.
- vii. The characteristics of the Normal Probability. Curve is based upon the theory of probability.

S.NO	Operator	Symbol	Example	Meaning
1	Union	$\cup$	$A \cup B$	The event of either A or B occurring.
2	Finite union	$\bigcup_{i=1}^n A_i$	$\bigcup_{i=1}^3 A_i$	The event of any one of the events $A_1, A_2$ and $A_3$ occurring.
3	Countable union	$\bigcup_{i=1}^{\infty} A_i$	$\bigcup_{i=1}^{\infty} A_i$	The event of any one of the events $A_1, A_2, \dots$ occurring.
4	Intersection	$\cap$	$A \cap B$	The event of both A and B occurring. 
5	Finite intersection	$\bigcap_{i=1}^n A_i$	$\bigcap_{i=1}^3 A_i$	The event of all the events $A_1, A_2$ and $A_3$ occurring.
6	Countable intersection	$\bigcap_{i=1}^{\infty} A_i$	$\bigcap_{i=1}^{\infty} A_i$	The event of all the events $A_1, A_2, \dots$ occurring.
7	Complementation	c or $-$	$A^c$ or $\bar{A}$	The event of A not occurring. 
8	Subtraction	$-$	$A - B$	The event of A occurring and B not occurring. 

Operation	Interpretation
$A', A \text{ or } A^c$	Not A.
$A \cup B$	At least, one of A and B
$A \cap B$	Both A and B
$(A' \cap B) \cup (A \cap B')$	Exactly one of A and B
$(A' \cap B') = (A \cup B)'$	Neither A nor B

**Elementary Properties of Probability:**

1)  $A'$  is complement of  $A$  and therefore  $P(A') = 1 - P(A)$

2) For any event  $A$  in  $S$ ,  $0 \leq P(A) \leq 1$

3) For the impossible event  $\phi$ ,  $P(\phi) = 0$

4) For the certain event  $S$ ,  $P(S) = 1$

5) If  $A_1$  and  $A_2$  two mutually exclusive events then  $P(A_1 \cup A_2) = P(A_1) + P(A_2)$

6) If  $A \subseteq B$ , then  $P(A) \leq P(B)$  and  $P(A' \cap B) = P(B) - P(A)$

7) Addition theorem: For any two events  $A$  and  $B$  of a sample space  $S$ ,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

8) For any two events  $A$  and  $B$ ,  $P(A \cap B') = P(A) - P(A \cap B)$

9) For any three events  $A$ ,  $B$  and  $C$  of a sample space  $S$ ,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - (P(A \cap C) + P(A \cap B \cap C))$$

10) If  $A_1, A_2, \dots, A_m$  are mutually exclusive events in  $S$ , then

$$P(A_1 \cup A_2 \cup \dots \cup A_m) = P(A_1) + P(A_2) + \dots + P(A_m)$$

**Ex-1:** If a die is rolled find the probability that number on the uppermost face of the die is

- a) an odd no.
- b) a prime no.
- c) greater than 2.

**Ex-2:** If three coins are tossed simultaneously, find the probability of getting

- a) exactly one head
- b) at least one head
- c) no head.

**Ex-3:** Find the probability that a leap year selected at random contains 53 Sundays.

**Ex-4:** In a housing society, half of the families have a single child per family, while the remaining half have two children per family. If a child is picked at random, find the probability that the child has a sibling.

**Ex-5:** A box contains 6 white and 4 black balls. 2 balls are selected at random and the colour is noted. Find the probability that

- a) Both balls are white
- b) Both balls are black.
- c) Balls are of different colours.

**Ex-6:** If all the letters of the word EAR are arranged at random. Find the probability that the word begins and ends with a vowel.

**Ex-7:** If all the letters of the word EYE are arranged at random find the probability that the word begins and ends with vowels.

**Ex-8:** If all the letters of word EQUATION are arranged at random, find the probability that the word begins and ends with a vowel.

**Ex-9:** 7 boys and 3 girls are arranged in a row. Find the probability that there is at least one boy between 2 girls.

**Ex-10:** 6 books on accountancy, 5 books of economics and 4 books of mathematics are to be arranged in the shelf find the probability that all the books of one subject are together.

**Complement of an event:** The complement of an event A is denoted by  $\bar{A}$  or  $A'$  or  $A^C$  and it contains all the elements of the sample space which do not belong to A.

e.g.random experiment: an unbiased die is rolled.

$$S = \{1, 2, 3, 4, 5, 6\}$$

(i) Let A: number on the die is a perfect square

$$\therefore A = \{1, 4\} \therefore \bar{A} = \{2, 3, 5, 6\}$$

(ii) Let B: number on the die is a prime number

$$\therefore B = \{2, 3, 5\} \therefore \bar{B} = \{1, 4, 6\}$$

**Note:**

$$P(A) + P(\bar{A}) = 1$$

$$\text{i.e. } P(A) = 1 - P(\bar{A})$$

**Ex-11:** If 3 dice are tossed simultaneously, find the probability that the sum of the 3 numbers is less than 17.

**Addition Theorem of Probability:**

**Result:** If A and B are any two events then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Note:**

1)  $A \cup B$  : either A or B or both;

$A \cup B$  : at least one of A & B

2) If A & B are mutually exclusive,  $P(A \cap B) = 0$

$$\therefore P(A \cup B) = P(A) + P(B)$$

$$3) P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

**Ex-13:** The probability that a particular film gets award for best direction is 0.7. The probability that it gets award for best acting is 0.4. The probability that the film gets award for both is 0.2. Find the probability that the film gets

- a) at least one award
- b) no award

**Ex-14:** Two cards are selected at random from a pack of 52 cards, find the probability that two cards are

- a) Red or face cards
- b) Aces or jacks.

### **Conditional Probability**

Let S be a sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S.

Then the probability of occurrence of event A under the condition that event B has already occurred and  $P(B) \neq 0$  is called conditional probability of event A given B and is denoted by  $P(A/B)$ .

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0$$

### **Multiplication theorem**

Let S be sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S.

Then the probability of occurrence of both the events is denoted by  $P(A \cap B)$

and is given by  $P(A \cap B) = P(A).P(B/A)$

### **Independent Events**

Let S be sample space associated with the given random experiment.

Let A and B be any two events defined on the sample space S. If the occurrence of either event, does not affect the probability of the occurrence of the other event, then the two events A and B are said to be independent.

Thus, if A and B are independent events then,

$$P(A/B) = P(A/B') = P(A) \text{ and } P(B/A) = P(B/A') = P(B)$$

If A and B are independent events then  $P(A \cap B) = P(A).P(B)$

$$(P(A \cap B) = P(A).(B/A) = P(A).P(B) \therefore P(A \cap B) = P(A).P(B))$$

### **If A and B are independent events then**

a) A and B' are also independent event

b) A' and B' are also independent event

**Ex-15:** 2 shooters are firing at target. The probability that they hit the target are  $\frac{1}{3}$  and  $\frac{1}{2}$  respectively. If they fire independently find the probability that

- a) both hit the target.
- b) Nobody hits the target.
- c) At least one hits the target.
- d) Exactly one hits the target.

**Ex-17:** Suppose A & B are two independent events such that  $P(A) = 0.4$ ,  $P(A \cup B') = 0.7$ . Find  $P(A \cup B)$ .

**Ex-18:** Three vendors were asked to supply a component. The respective probabilities that the component supplied by them is 'good' are 0.8, 0.7 and 0.5. Each vendor supplies only one component. Find the probability that at least one component is 'good'.

**Ex-19:** The chance of a student passing a test is 20%. The chance of student passing the test and getting above 90% marks is 5%. Given that a student passes the test, find the probability that the student gets above 90% marks.

**Ex-20:** A box contains 6 white and 4 black balls. One ball is selected at random and its colour is noted. The ball is replaced and two balls of the opposite colour are added and then second ball is selected at random find the probability that both balls are white.

**Ex-21:** A shop has equal number of LED bulbs of two different types. The probability that the life of an LED bulb is more than 100 hours given that it is of type-1 is 0.7 and given that it is of type-2 is 0.4. If an LED bulb is selected at random, find the probability that the life of the bulb is more than 100 hours.

## Bayes Theorem

Bayes' Theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence

- Bayes' Theorem allows you to update the predicted probabilities of an event by incorporating new information.
- Bayes' Theorem was named after 18th-century mathematician Thomas Bayes.
- It is often employed in finance in calculating or updating risk evaluation.
- The theorem has become a useful element in the implementation of machine learning.
- The theorem was unused for two centuries because of the high volume of calculation capacity required to execute its transactions.

Applications of Bayes' Theorem are widespread and not limited to the financial realm. For example, Bayes' theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test. Bayes' theorem relies on incorporating prior probability distributions in order to generate posterior probabilities.

Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed.

Posterior probability is the revised probability of an event occurring after taking into consideration the new information. Posterior probability is calculated by updating the prior probability using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Bayes' Theorem formula

posterior

↓

likelihood

↓

prior

↙

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

↑

evidence

**Ex-22:** In a bolt factory, three machine P, Q and R produce 25%, 35% and 40% of the total output respectively. It is found that in their production, respectively 5%, 4% and 2% are defective bolts. If a bolt is selected at random and found defective, find the probability that it is produced by machine Q.

**Ex-23:** A certain test for a particular cancer is known to be 95% accurate. A person submits to the test and the results are positive. Suppose that the person comes from a population of 1,00,000 where 2,000 people suffer from that disease. What can we conclude about the probability that the person under test has that particular disease?



**ODDS (Ratio of two complementary probabilities):**

Let  $n$  be number of distinct sample points in the sample space  $S$ . Out of  $n$  sample points,  $m$  sample points are favourable for the occurrence of event  $A$ . Therefore remaining  $(n-m)$  sample points are favourable for the occurrence of its complementary event  $A'$ .

$$\therefore P(A) = \frac{m}{n} \text{ and } P(A') = \frac{n-m}{n}$$

Ratio of number of favourable cases to number of unfavourable cases is called as odds in favour of event  $A$  which is given by  $\frac{m}{n-m}$  i.e.  $P(A):P(A')$ .

Ratio of number of unfavourable cases to number of favourable cases is called as odds against event  $A$  which is given by  $\frac{n-m}{m}$  i.e.  $P(A'):P(A)$

## Session 5 & 6

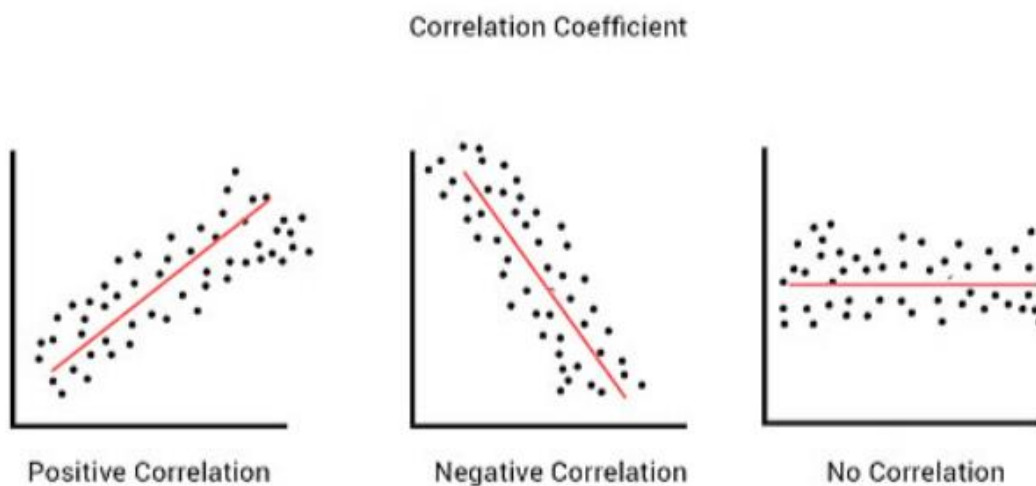
### What Is Correlation?

Correlation refers to the statistical relationship between the two entities. It measures the extent to which two variables are linearly related. For example, the height and weight of a person are related, and taller people tend to be heavier than shorter people.

You can apply correlation to a variety of data sets. In some cases, you may be able to predict how things will relate, while in others, the relation will come as a complete surprise. It's important to remember that just because something is correlated doesn't mean it's causal.

There are three types of correlation:

- **Positive Correlation:** A positive correlation means that this linear relationship is positive, and the two variables increase or decrease in the same direction.
- **Negative Correlation:** A negative correlation is just the opposite. The relationship line has a negative slope, and the variables change in opposite directions, i.e., one variable decreases while the other increases.
- **No Correlation:** No correlation simply means that the variables behave very differently and thus, have no linear relationship



### **Scatter diagram:**

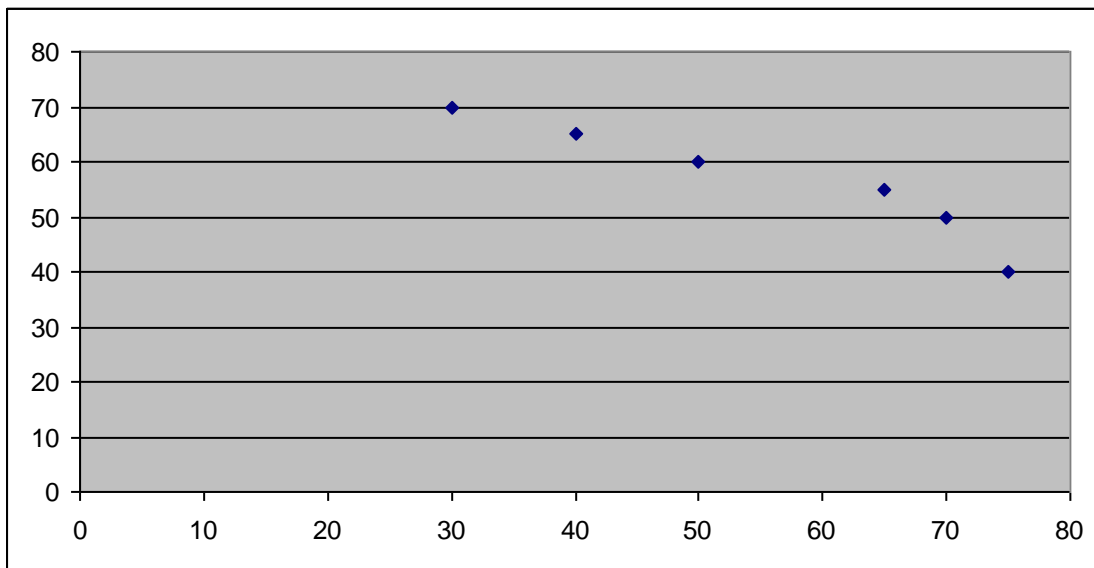
We collect a data of pairs of values of the two variables. Generally these variables are denoted by  $x$  &  $y$ . These values are considered as  $x$  &  $y$  co-ordinates respectively and plotted as points on a graph. Such diagrammatic representation of a bivariate data is called as a scatter diagram. From the scatter diagram a rough idea of the nature of relationship between the two variables can be drawn as follows.

**Ex:** Draw a scatter diagram for the following data and give your comments.

x	30	40	50	65	70	75
y	70	65	60	55	50	40

**Answer:** Scatter Diagram:

Comment: There is negative correlation between  $x$  &  $y$



**Ex:** Draw a scatter diagram for the following data and comment.

Demand	15	20	18	22	25	30
Price	32	19	25	15	12	10

## Correlation Coefficient

The correlation coefficient,  $r$ , is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When  $r$  is close to 0 this means that there is little relationship between the variables and the farther away from 0  $r$  is, in either the positive or negative direction, the greater the relationship between the two variables.

$$r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

### **Note:**

- 1)  $r$  lies between -1 & 1 i.e.  $-1 \leq r \leq 1$
- 2) If  $r = 1$ , there is perfect positive correlation
- 3) If  $0 < r < 1$ , there is positive correlation
- 4) If  $r = -1$ , there is perfect negative correlation
- 5) If  $-1 < r < 0$ , there is negative correlation
- 6) If  $r = 0$ , there is no correlation
- 7) Correlation Coefficient is independent of change of origin & change of scale.

**Ex:** Calculate correlation coefficient for the following data. Comment on your findings.

Marks in Statistics	53	59	72	43	93	35	55	70
Marks in Economics	35	49	63	36	75	28	38	76

**Ex:** Calculate Karl Pearson's Coefficient of correlation for the following data.

X	17	8	12	13	10	12
Y	13	7	10	11	8	11

**Ex:** Find the Karl Pearson's correlation coefficient for the following data.

x	10	14	12	18	20	16
y	20	30	20	35	25	20

**Spearman's Rank Correlation Coefficient:**

In this method, ranks are assigned to the data. The ranks are given to the x-series & y-series separately. The highest observation is given rank '1', the next highest observation is given rank '2' and so on. Suppose,  $R_1$  &  $R_2$  are the ranks of the x & y respectively and  $d = R_1 - R_2$  then

$$r = 1 - \left\{ \frac{6 \sum d^2}{n(n^2 - 1)} \right\}$$

where  $n$  = number of pairs of observations

**Ex:** Calculate the Spearman's rank correlation coefficient for the following data.

x	15	12	16	13	17	14	18	11
y	17	14	20	25	23	24	22	21

**Ex:** Calculate the Spearman's rank correlation coefficient for the following data.

x	50	63	40	70	45	65	38	53	52
y	48	30	35	60	55	33	25	54	50

**"Causation is not correlation"** is a fundamental concept in statistics and scientific research. It essentially means that just because two variables are correlated (meaning they tend to vary together), it doesn't necessarily mean that one causes the other to happen.

Here's an example to illustrate this:

Let's say we observe a strong positive correlation between ice cream sales and the number of drownings at the beach. During the summer months, both ice cream sales and drownings tend to increase. However, it would be incorrect to conclude that eating ice cream causes people to drown or vice versa.

There could be a third variable at play here, such as temperature. Warmer temperatures in the summer lead to increased ice cream consumption as well as more people going to the beach and swimming, which in turn increases the risk of drownings. So, in this example, temperature is the common cause behind both variables—ice cream sales and drownings—rather than one causing the other directly.

To establish causation, researchers often conduct controlled experiments or use sophisticated statistical methods to account for potential confounding variables. These methods help them determine if changes in one variable directly lead to changes in another variable, thus establishing a cause-and-effect relationship.

## Covariance Meaning

**Covariance** is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

## Types of Covariance

Covariance can have both positive and negative values. Based on this, it has two types:

1. Positive Covariance
2. Negative Covariance

### Positive Covariance

If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

### Negative Covariance

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

## Covariance Formula

Covariance formula is a statistical formula, used to evaluate the relationship between two variables. It is one of the statistical measurements to know the relationship between the variance between the two variables. Let us say X and Y are any two variables, whose relationship has to be calculated. Thus the covariance of these two variables is denoted by  $\text{Cov}(X,Y)$ . The formula is given below for both population covariance and sample covariance.

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

---

If  $\text{cov}(X, Y)$  is greater than zero, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.

If  $\text{cov}(X, Y)$  is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.

If  $\text{cov}(X, Y)$  is zero, then we can say that there is no relation between two variables.

## **Regression**

What is Linear Regression?

“Linear regression predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values.

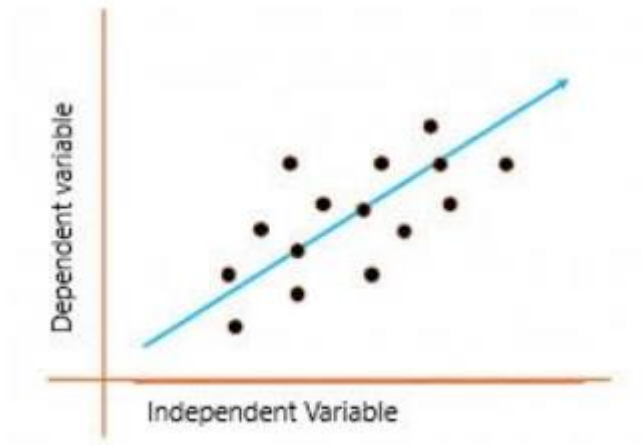
Applied in various domains like economics and finance, this method analyzes and forecasts data trends. It can extend to multiple linear regression involving several independent variables and logistic regression, suitable for binary classification problems

### Simple Linear Regression

**In a simple linear regression, there is one independent variable and one dependent variable. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables. The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.**

Linear regression is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression. If there is a single input variable **X**(independent variable), such linear regression is **simple linear regression**.





The graph above presents the linear relationship between the output(y) and predictor(X) variables. The blue line is referred to as the *best-fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.

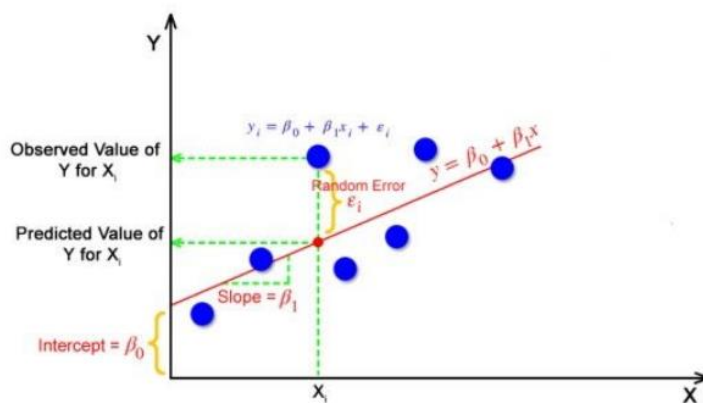
### Simple Regression Calculation

To calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

$$Y_i = \beta_0 + \beta_1 X_i$$

where  $Y_i$  = Dependent variable,  $\beta_0$  = constant/Intercept,  $\beta_1$  = Slope/Intercept,  $x_i$  = Independent variable.

This algorithm explains the linear relationship between the dependent(output) variable y and the independent(predictor) variable X using a straight line  $Y = B_0 + B_1 X$ .



But how the linear regression finds out which is the best fit line?

The goal of the linear regression algorithm is to get the **best values for  $B_0$  and  $B_1$**  to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

## Random Error(Residuals)

In regression, the difference between the observed value of the dependent variable( $y_i$ ) and the predicted value(**predicted**) is called the residuals.

$$\epsilon_i = y_{\text{predicted}} - y_i$$

$$\text{where } y_{\text{predicted}} = B_0 + B_1 X_i$$

## Outliers

What is an outlier?

An outlier is an observation “that appears to deviate markedly from other members of the sample in which it occurs”

What causes outliers?

- Human errors, e.g. data entry errors
- Instrument errors, e.g. measurement errors
- Data processing errors, e.g. data manipulation
- Sampling errors, e.g. extracting data from wrong sources
- Not an error, the value is extreme, just a ‘novelty’ in the data

A dilemma

- Outliers can be genuine values
- The trade-off is between the loss of accuracy if we throw away “good” observations, and the bias of our estimates if we keep “bad” ones
- The challenge is twofold:
  1. to figure out whether an extreme value is good (genuine) or bad (error)
  2. to assess its impact on the statistics of interest

Outlier treatment is the process of identifying and handling outliers in a dataset. Outliers are defined as observations that fall outside of the general pattern of the data and can have a significant impact on the analysis and modeling of the data.

There are several methods for identifying and treating outliers, including:

**Z-score method:** This method calculates the standard deviation and mean of the data, and any observation that falls more than 3 standard deviations away from the mean is considered an outlier.

**Interquartile range method:** This method calculates the interquartile range (IQR) of the data, and any observation that falls outside of the lower and upper limits of the box plot, which are defined as  $Q1 - 1.5 * IQR$  and  $Q3 + 1.5 * IQR$ , respectively, is considered an outlier.

**Clustering methods:** Clustering methods such as DBScan and KMeans can also be used to identify outliers by grouping similar data points together and identifying any data points that do not belong to any cluster.

**Visualization techniques:** Visualization techniques such as box plots and scatter plots can also be used to identify outliers by visually identifying any points that fall outside of the general pattern of the data.

There are many other advanced methods that we will read about in the following section of the article.

Once outliers have been identified, there are several methods for handling them. The appropriate method for handling outliers will depend on the specific dataset and the goals of the analysis. It is important to carefully consider the impact of outliers on the data and the appropriate method for handling them before proceeding with any analysis or modeling. Some common techniques include:

**Deleting the outlier observations:** This is a simple method, but it can lead to a loss of information if the outliers are actually meaningful observations.

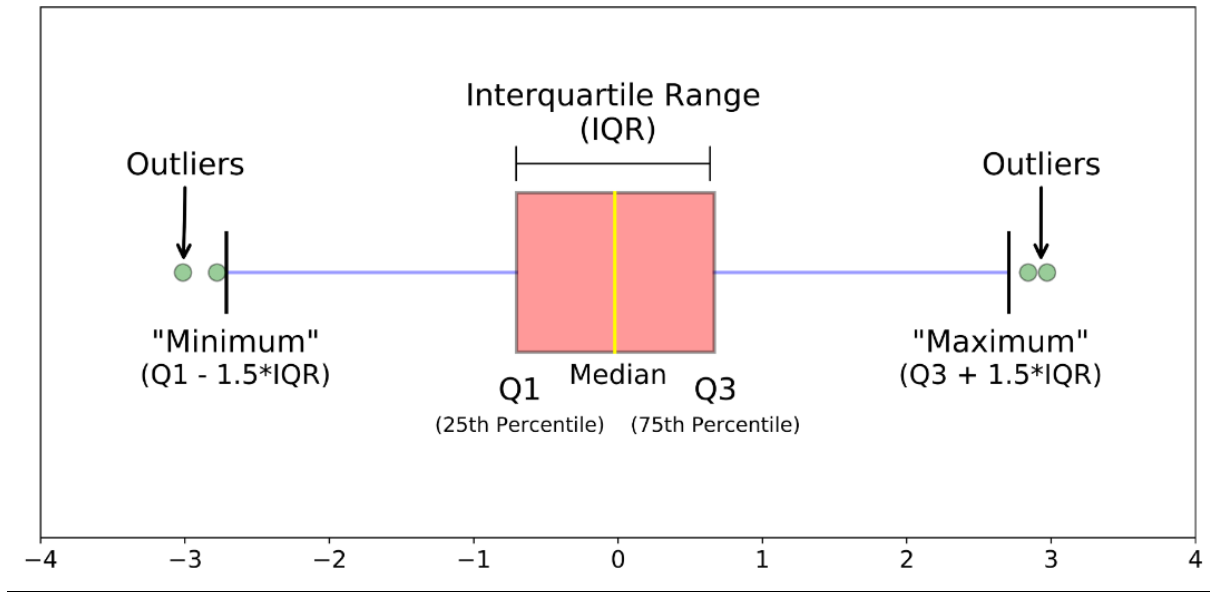
**Trimming the data:** This method involves removing a certain percentage of the largest and smallest observations.

**Winsorizing:** This method replaces the outliers with a value that is closer to the center of the data.

**Log transformation:** This method can be used when the data is positively skewed and the outliers are on the high end of the distribution.

**Z-score standardization:** This method replaces each observation with its z-score, which is the number of standard deviations away from the mean.

**Cap and floor:** This method replaces the outlier with a maximum and minimum value respectively.



## **Session 7 & 8**

## **Session 9 & 10**

### **Descriptive Statistical measures**

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics. The mean, or the average, is calculated by adding all the figures within the data set and then dividing by the number of figures within the set.

For example, the sum of the following data set is 20: (2, 3, 4, 5, 6). The mean is 4 ( $20/5$ ). The mode of a data set is the value appearing most often, and the median is the figure situated in the middle of the data set. It is the figure separating the higher figures from the lower figures within a data set.

### **Central Tendency**

Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables and general discussions to help people understand the meaning of the analysed data.

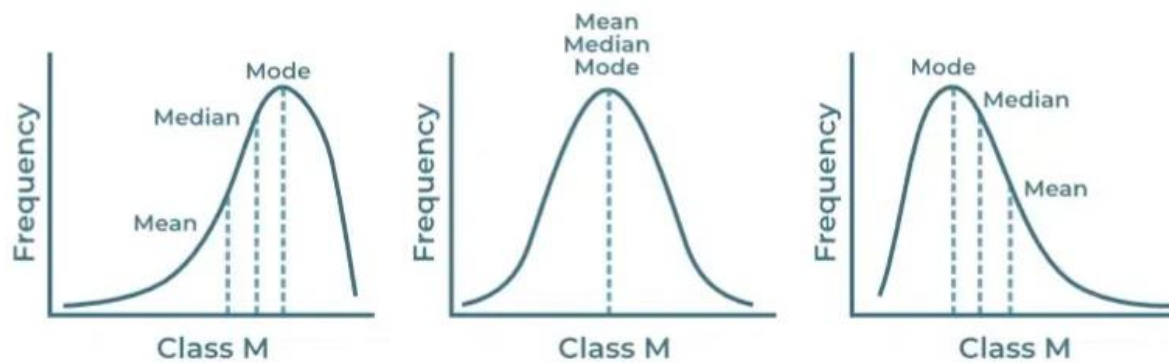
Measures of central tendency describe the centre position of a distribution for a data set. A person analyses the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analysed data set.

### **Measures of Variability**

Measures of variability (or the measures of spread) aid in analysing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.

So, while the average of the data maybe 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).



### **Measures of Central Tendency (Averages)**

One of the most important objectives of statistical analysis is to get one single value that describes the characteristic of the entire data. An average is the value of the variable which is representative of the entire data. It gives us idea about the concentration of the values in the central part of the distribution.

### **Types of Averages:**

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

### **Requisites of a good average:**

- 1) Easy to understand
- 2) Simple to compute
- 3) Based on all the items.
- 4) Not unduly affected by extreme observations.
- 5) Rigidly defined.
- 6) Capable for further algebraic treatment.
- 7) Easy to interpret

### **Arithmetic Mean(A.M.):**

#### **Arithmetic mean for Raw Data:**

The A.M. of  $N$  observations  $X_1, X_2, \dots, X_N$  is denoted by  $\bar{X}$  and is defined as follows.

$$\text{A.M.} = \bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{\sum X}{N}$$

**Ex:** The grades of a student on six examinations were 84, 91, 72, 68, 87 and 78. Find the arithmetic mean of the grades.

**Ex:** The mean of 10 observations was found to be 20. Later on it was discovered that the observations 24 and 34 were wrongly noted as 42 and 54. Find the corrected mean.

#### **Arithmetic Mean for a frequency distribution:**

Consider a data of  $n$  observations  $X_1, X_2, \dots, X_n$  occurring with respective frequencies  $f_1, f_2, \dots, f_n$ . Then the A.M. is denoted by  $\bar{X}$  and is defined as follows.

$$\text{A.M.} = \bar{X} = \frac{f_1 X_1 + f_2 X_2 + \dots + f_n X_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum fX}{\sum f}$$

We denote  $\sum f$  by  $N$ , called as total frequency

$$\therefore \text{A.M.} = \bar{X} = \frac{\sum fX}{N}$$

**Ex:** The following table gives the monthly income of 20 families in a city. Calculate the arithmetic mean.

Income (in '00 Rs.)	16	20	30	35	45	50
No. of families	2	5	4	6	2	1



**Ex:** Use the following frequency distribution of heights to find the arithmetic mean of height of 100 students at XYZ university.

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

**Ex:** Use the following frequency distribution of weekly wages to find the arithmetic mean of wage of employees at P & R company.

Weekly Wage (\$)	Number of employees
250.00 – 259.99	8
260.00 – 269.99	10
270.00 – 279.99	16
280.00 – 289.99	14
290.00 – 299.99	10
300.00 – 309.99	5
310.00 – 319.99	2

**Ex:** Find the missing frequency if the mean is 21.9

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	2	5	--	13	21	16	8	3

### **Combined Mean:**

**Ex:** The average marks of a group of 100 students in Statistics are 60 and for other group of 50 students, the average marks are 90. Find the average marks of the combined group of 150 students.

**Answer:** Given:  $N_1 = 100$ ,  $\bar{X}_1 = 60$ ,  $N_2 = 50$ ,  $\bar{X}_2 = 90$

$$\begin{aligned}
 \text{Combined Mean} = \bar{X} &= \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} \\
 &= \frac{(100 \times 60) + (50 \times 90)}{100 + 50} = \underline{70}
 \end{aligned}$$

**Weighted Arithmetic Mean:**

Sometimes we associate certain weighing factor (or weights) with the numbers  $X_1, X_2, \dots, X_n$ . Suppose the weights are  $w_1, w_2, \dots, w_n$ . The weighted arithmetic mean is denoted by  $\bar{X}$  and is defined as follows.

$$\begin{aligned}\text{Weighted A.M.} = \bar{X} &= \frac{w_1 X_1 + w_2 X_2 + \dots + w_n X_n}{w_1 + w_2 + \dots + w_n} \\ &= \frac{\sum wX}{\sum w}\end{aligned}$$

**Ex:** A student's grades in laboratory, lecture and recitation parts of a Physics course were 71, 78 and 89 respectively. If the weights of these grades are 2, 4 and 5 respectively, what is the appropriate average grade?

**Answer:**

X	w	wX
71	2	142
78	4	312
89	5	445
	$\Sigma w = 11$	$\Sigma wX = 899$

$$\text{Weighted A.M.} = \bar{X} = \frac{\sum wX}{\sum w} = \frac{899}{11} = 81.7273$$

## **MERITS AND DEMERITS OF MEAN**

### **Merits:**

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. It is based on all the observations.
4. It is capable of further algebraic treatment.
5. Of all the averages, A.M. is least affected by sampling fluctuations i.e. it is a stable average.

### **Demerits:**

1. It cannot be obtained by mere inspection nor can it be located graphically.
2. It cannot be obtained even if a single observation is missing. It is affected by extreme values.
3. It is affected by extreme values
4. It cannot be calculated for frequency distribution having open end class- intervals e.g. class-intervals like below 10 or above 50 etc.
5. It may be a value which may not be present in the data.
6. Sometimes, it gives absurd results. e.g. Average number of children per family is 1.28.
7. It cannot be used for the study of qualitative data such as intelligence, honesty, beauty etc.

Even though A.M. has various demerits, it is considered to be the best all averages as it satisfies most of the requisites of a good average. A.M. is called the Ideal Average.

**Median:**

The **Median** of a set of numbers arranged in order of magnitude is either the middle value or arithmetic mean of the two middle values.

**Median for raw data:**

Suppose there are N observations.

If N is an odd number,

$$\text{median} = \left( \frac{N + 1}{2} \right)^{\text{th}} \text{ observation in order sample}$$

If N is an even number,

$$\text{median} = \text{average of } \left( \frac{N}{2} \right)^{\text{th}} \text{ \& } \left( \frac{N}{2} + 1 \right)^{\text{th}} \text{ observation in order sample}$$

**Ex:** The number of ATM transactions per day was recorded at 15 locations in a large city. The data were as follows.

35, 49, 225, 50, 30, 65, 40, 55, 52, 76, 48, 325, 47, 32, 60.

Find the median number of transactions.

**Answer:** Given observations in ascending order:

30, 32, 35, 40, 47, 48, 49, **50**, 52, 55, 60, 65, 76, 225, 325

N = 15, which is an 'odd' number

$$\left( \frac{N + 1}{2} \right) = 8$$

$$\begin{aligned} \text{median} &= \left( \frac{N + 1}{2} \right)^{\text{th}} \text{ observation in order sample} \\ &= 8^{\text{th}} \text{ observation in order sample} = \mathbf{\underline{50}} \end{aligned}$$

**Ex:** The following table gives the data of weight of 20 students at a University. Find the median weight.

138, 146, 168, 146, 161, 164, 158, 126, 173, 145,  
150, 140, 138, 142, 135, 132, 147, 176, 147, 142

**Ex:** Find the median of the following data.

X	10	15	25	40	60	75
f	3	4	6	17	12	7

**Answer:**  $\Sigma f = N = 49$ , N is an odd number

x	10	15	25	40	60	75
f	3	4	6	17	12	7
less than cumulative freq	3	7	13	30	42	49

$$\text{median} = \left( \frac{N + 1}{2} \right)^{\text{th}} \text{ observation in order sample}$$

$$= 25^{\text{th}} \text{ observation in order sample} = \underline{40}$$

### **Median for grouped data:**

Suppose the total number of observations is N. The “median class” is defined as the class interval for which the less than cumulative frequency is just greater than  $N/2$ . Then,

$$\text{median} = \ell_1 + \left\{ \frac{(\ell_2 - \ell_1)}{f} \left( \frac{N}{2} - \text{c.f.} \right) \right\}$$

$\ell_1$  = lower class boundary of median class

$\ell_2$  = upper class boundary of median class

f = frequency of median class

c.f. = less than cumulative frequency of class preceding median class

**Ex:** Find the median of the following data.

Weight (pounds)	Frequency
118 – 126	3
127 – 135	5
136 – 144	9
145 – 153	12
154 – 162	5
163 – 171	4
172 – 180	2

**Answer:** Given:  $N = \Sigma f = 40$  ;  $N/2 = 20$

Weight (pounds)	Frequency	Less Than Cumulative Frequency
118 – 126	3	3
127 – 135	5	8
136 – 144	9	<u>17</u>
<u>145 – 153</u>	<u>12</u>	29
154 – 162	5	34
163 – 171	4	38
172 – 180	2	40

The median class is 145 – 153

$$\begin{aligned}
 \text{median} &= \ell_1 + \left\{ \frac{(\ell_2 - \ell_1)}{f} \left( \frac{N}{2} - \text{c.f.} \right) \right\} \\
 &= 144.5 + \left[ \left( \frac{153.5 - 144.5}{12} \right) \times (20 - 17) \right] = \underline{146.75}
 \end{aligned}$$

## **MERITS AND DEMERITS OF MEDIAN**

### **Merits:**

- 1.It is easy to understand and easy to calculate.
- 2.It is quite rigidly defined.
- 3.It can be computed for a distribution with open-end classes.
- 4.In majority of the cases, it is one of the values in the data.
- 5.It can be determined graphically.
- 6.Since median is a positional average, it can be computed even if the observations at the extremes are unknown.
- 7.It is not highly affected by fluctuations in sampling.
- 8.It can be calculated even for qualitative data.

### **Demerits:**

- 1.When the number of observations is large, the pre-requisite of arranging observations in ascending/descending order of magnitude is a difficult process.
2. It is not based on all observations and hence, may not be a proper representative.
3. It is not capable of further mathematical treatment.
4. Since it does not require information about all the observation, it is insensitive to some changes

**Mode:**

The **Mode** of a set of numbers is that value which occurs with the greatest frequency, that is, it is the most common value.

**Note:**

- 1) Sometimes the **Mode** may not exist.
- 2) Even if the **Mode** exists, sometimes it may not be unique.

**Ex:** The reaction times of an individual to a certain stimulus were measured as 0.53, 0.46, 0.50, 0.49, 0.52, 0.44, 0.55, 0.53, 0.40 and 0.56. Find the mode.

**Ex:** Three teachers of a subject reported examination grade of 79, 74 and 82 in their classes, which consisted of 32, 25 and 17 students respectively. Determine the mode.

**Mode for a grouped data:**

“Modal class” is the class interval with maximum frequency.

$$\text{mode} = \ell_1 + \left\{ (\ell_2 - \ell_1) \left( \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \right\}$$

$\ell_1$  = lower class boundary of modal class

$\ell_2$  = upper class boundary of modal class

$f_0$  = frequency of the class preceding modal class

$f_1$  = frequency of modal class

$f_2$  = frequency of the class next to modal class



## **MERITS AND DEMERITS OF MODE**

### **Merits:**

1. It is easy to understand and simple to calculate.
2. It is not affected by extreme values or sampling fluctuations.
3. It can be calculated for distribution with open-end classes.
4. It can be determined graphically.
5. It is always present within the data and is the most typical value of the given set of data.
6. It is applicable to both, qualitative and quantitative data.

### **Demerits**

1. It is not rigidly defined.
2. It is not based on all observations.
3. It is not capable of further mathematical treatment.
4. It is indeterminate if the modal class is at the extreme of the distribution.
5. If the sample of data for which mode is obtained is small, then such mode has no significance.

**Geometric Mean (G.M.):****Geometric Mean for Raw Data:**

The G.M. of N observations  $X_1, X_2, \dots, X_N$  is denoted by G and is defined as follows.

$$\text{G.M.} = G = \sqrt[N]{X_1 X_2 \dots X_N}$$

**Ex:** Find the Geometric Mean of the following data.

28.5, 73.6, 47.2, 31.5 and 64.8.

**Answer:**  $N = 5$

$$\text{G.M.} = G = \sqrt[N]{X_1 X_2 \dots X_N}$$

$$= \sqrt[5]{28.5 \times 73.6 \times 47.2 \times 31.5 \times 64.8} = \underline{\underline{45.8258}}$$

**Geometric Mean for a frequency distribution:**

Consider a data of n observations  $X_1, X_2, \dots, X_n$  occurring with respective frequencies  $f_1, f_2, \dots, f_n$ . Then the G.M. is denoted by G and is defined as follows.

$$\text{G.M.} = G = \sqrt[N]{(X_1)^{f_1} (X_2)^{f_2} \dots (X_n)^{f_n}} \quad \text{where } N = \Sigma f$$

**Harmonic Mean(H.M.):****Harmonic mean for Raw Data:**

The H.M. of N observations  $X_1, X_2, \dots, X_N$  is denoted by H and is defined as follows.

$$H = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_N}}$$
$$\therefore \frac{1}{H} = \frac{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_N}}{N}$$
$$\therefore \frac{1}{H} = \frac{1}{N} \left( \frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_N} \right)$$
$$\therefore \frac{1}{H} = \frac{1}{N} \sum \left( \frac{1}{X} \right)$$

**Ex:** Cities A, B and C are equidistant from each other. A motorist travels from A to B at 30 mph, from B to C at 40 mph and from C to A at 50 mph. Find his average speed.

**Answer:**

$$\frac{1}{H} = \frac{1}{N} \sum \left( \frac{1}{X} \right) = \frac{1}{3} \left( \frac{1}{30} + \frac{1}{40} + \frac{1}{50} \right) = \frac{47}{1800}$$
$$\therefore H = \text{Harmonic Mean} = \frac{1800}{47} = \underline{\underline{38.2979}}$$

**Harmonic Mean for a frequency distribution:**

Consider a data of n observations  $X_1, X_2, \dots, X_n$  occurring with respective frequencies  $f_1, f_2, \dots, f_n$ . Then the H.M. is denoted by H and is defined as follows.

$$H = \frac{N}{\frac{f_1}{X_1} + \frac{f_2}{X_2} + \dots + \frac{f_n}{X_n}} \quad \text{where } N = \Sigma f$$

**Ex:** An airplane travels distances of 2500, 1200 and 500 miles at speeds 500, 400 and 250 mph respectively. Find the Harmonic mean.

**Answer:**

$$\begin{aligned} \text{Harmonic Mean} &= \frac{\Sigma w}{\Sigma \left( \frac{w}{X} \right)} \\ &= \frac{2500 + 1200 + 500}{\left( \frac{2500}{500} \right) + \left( \frac{1200}{400} \right) + \left( \frac{500}{250} \right)} = \underline{\underline{420}} \end{aligned}$$

### **Quartiles, Deciles and Percentiles:**

Suppose we arrange a set of data in order of magnitude. The values which divide the set into four equal parts are denoted by  $Q_1, Q_2, Q_3$  and are called as the 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> Quartiles respectively. Similarly, the values which divide the set into ten equal parts are denoted by  $D_1, D_2, \dots, D_9$  and are called as the 1<sup>st</sup>, 2<sup>nd</sup>, ..., 9<sup>th</sup> Deciles respectively. And, the values which divide the set into hundred equal parts are denoted by  $P_1, P_2, \dots, P_{99}$  and are called as the 1<sup>st</sup>, 2<sup>nd</sup>, ..., 99<sup>th</sup> Percentiles respectively. Collectively, Quartiles, Deciles and Percentiles are called as **Quantiles**.

$$Q_i = i^{\text{th}} \text{ Quartile} = L_1 + \left\{ \frac{c}{f} \left( \frac{iN}{4} - \text{c.f.} \right) \right\} \text{ where } i = 1, 2, 3$$

$L_1$  = lower class boundary of  $i^{\text{th}}$  quartile class

$f$  = frequency of  $i^{\text{th}}$  quartile class

c.f. = less than cumulative freq of class preceding  $i^{\text{th}}$  quartile class

$c$  = class width

$$D_i = i^{\text{th}} \text{ Decile} = L_1 + \left\{ \frac{c}{f} \left( \frac{iN}{10} - \text{c.f.} \right) \right\} \text{ where } i = 1, 2, \dots, 9$$

$L_1$  = lower class boundary of  $i^{\text{th}}$  decile class

$f$  = frequency of  $i^{\text{th}}$  decile class

c.f. = less than cumulative freq of class preceding  $i^{\text{th}}$  decile class

$c$  = class width

$$P_i = i^{\text{th}} \text{ Percentile} = L_1 + \left\{ \frac{c}{f} \left( \frac{iN}{100} - \text{c.f.} \right) \right\} \text{ where } i = 1, 2, \dots, 99$$

$L_1$  = lower class boundary of  $i^{\text{th}}$  percentile class

$f$  = frequency of  $i^{\text{th}}$  percentile class

c.f. = less than cumulative freq of class preceding  $i^{\text{th}}$  percentile class

$c$  = class width

**Ex:** For the following data of age, calculate the 1<sup>st</sup> Quartile, 5<sup>th</sup> Decile and 54<sup>th</sup> Percentile.

Age in years	Number of persons
0 – 10	6
10 – 20	8
20 – 30	13
30 – 40	18
40 – 50	16
50 – 60	13
60 – 70	12
70 – 80	9
80 – 90	4
90 – 100	1

**Answer:**  $N = \Sigma f = 100$

Class Interval	Frequency	Less Than Cumulative Frequency
0 – 10	6	6
10 – 20	8	14
20 – 30	13	27
30 – 40	18	45
40 – 50	16	61
50 – 60	13	74
60 – 70	12	86
70 – 80	9	95
80 – 90	4	99
90 – 100	1	100

1) **To find Q<sub>1</sub>:**

Consider  $\frac{iN}{4}$ , put  $i = 1$

$\frac{1 \times N}{4} = 25$ , class containing  $Q_1$  is 20-30

$$Q_i = i^{\text{th}} \text{ Quartile} = L_1 + \left\{ \frac{c}{f} \left( \frac{iN}{4} - c.f. \right) \right\}, \text{ put } i = 1$$

$$Q_1 = L_1 + \left\{ \frac{c}{f} \left( \frac{1 \times N}{4} - c.f. \right) \right\} = 20 +$$

$$\left\{ \frac{10}{13} (25 - 14.) \right\} = \underline{\underline{28.4615}}$$

2) **To find D<sub>5</sub>:**

Consider  $\frac{iN}{10}$ , put  $i = 5$

$\frac{5 \times N}{10} = 50$ , class containing  $Q_1$  is 40-50

$$D_i = i^{\text{th}} \text{ Decile} = L_1 + \left\{ \frac{c}{f} \left( \frac{iN}{10} - c.f. \right) \right\}, \text{ put } i = 5$$

$$D_5 = L_1 + \left\{ \frac{c}{f} \left( \frac{5N}{10} - c.f. \right) \right\} = 40 + \left\{ \frac{10}{16} (50 - 45.) \right\} =$$

**43.125**

3) **To find P<sub>54</sub>:**

$$\text{Consider } \frac{iN}{100}, \text{ put } i = 54$$

$$\frac{54 \times N}{100} = 54, \text{ class containing } Q_1 \text{ is } 40-50$$

$$P_i = i^{\text{th}} \text{ Percentile} = L_1 + \left\{ \frac{c}{f} \left( \frac{iN}{100} - \text{c.f.} \right) \right\}, \text{ put } i = 54$$

$$P_{54} = L_1 + \left\{ \frac{c}{f} \left( \frac{54N}{100} - \text{c.f.} \right) \right\} = 40 +$$

$$\left\{ \frac{10}{16} (54 - 45.) \right\} = \underline{\underline{45.625}}$$



## Comparison between Central Tendencies

We have studied five different measures of central tendency. It is obvious that no single measure can be the best for all situations. The most commonly used measures are mean, median and mode. It is not desirable to consider any one of them to be superior or inferior in all situations. The selection of appropriate measure of central tendency would largely depend upon the nature of the data; more specifically, on the scale of measurement used for representing the data and the purpose on hand.

The data obtained on nominal scale, we can count the number of cases in each category and obtain the frequencies. We may then be interested in knowing the class which is most popular or the most typical value in the data. In such cases, mode can be used as the appropriate measure of central tendency.

e.g. Suppose in a genetical study, for a group of 50 family members, we want to know most common colour of eyes. Then we count the number of persons for each different colour of eye. Suppose 3 persons have light eyes, 6 persons have brown eyes, 12 with dark grey eyes and 29 persons are with black eyes. Then the most common colour of eyes (i.e. mode) for this group of people is 'black'.

When the data is available on ordinal scale of measurement i.e. the data is provided in rank order, use of median as a measure of central tendency is appropriate. Suppose in a group of 75 students, 10 students have failed, 15 get pass class, 20 secure second class and 30 are in first class. The average performance of the students will be the performance of the middlemost student (arranged as per rank) i.e. the performance of 38th student i.e. second class; which is the median of the data. Median is only a point on the scale of measurement, below and above which lie exactly 50% of the data. Median can also be used (i) for truncated (incomplete) data, provided we know the total number of cases and their positions on the scale and (ii) when the distribution is markedly skewed.

Arithmetic Mean is the most commonly used measure of central tendency. It can be calculated when the data is complete and is represented on interval or ratio scale. It represents the centre of gravity of the data i.e. the measurements in any sample are perfectly balanced about the mean. In computation of simple A.M., equal importance is given to all observations in the data. It is preferred because of its high reliability and its applicability to inferential statistics. Thus, A.M. is more precise, reliable and stable measure of central tendency.

## What are the objectives of computing dispersion?

### (1) Comparative study

- Measures of dispersion give a single value indicating the degree of consistency or uniformity of distribution. This single value helps us in making comparisons of various distributions.
- The smaller the magnitude (value) of dispersion, higher is the consistency or uniformity and vice-versa.

### (2) Reliability of an average

- A small value of dispersion means low variation between observations and average. It means that the average is a good representative of observation and very reliable.
- A higher value of dispersion means greater deviation among the observations. In this case, the average is not a good representative and it cannot be considered reliable.

### (3) Control the variability **Measure**

- Different measures of dispersion provide us data of variability from different angles, and this knowledge can prove helpful in controlling the variation.
- Especially in the financial analysis of business and medicine, these measures of dispersion can prove very useful.

### (4) Basis for further statistical analysis

- Measures of dispersion provide the basis for further statistical analysis like computing correlation, regression, test of hypothesis, etc.

Measure of dispersion:

- 1) Absolute unit of measurement is same
- 2) Relative when the units are different

## Range

Definition: If L is the largest observation in the data and S is the smallest observation, then range is the difference between L and S. Thus,

$$\text{Range} = L - S$$

For a frequency distribution, range may be considered as the difference between the largest and the smallest class-boundaries.

Range is a crude and simplest measure of dispersion. It measures the scatter of observations among themselves and not about any average.

The corresponding relative measure is

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

Note:

1. Range is a suitable measure of dispersion in case of small groups. In the branch of statistics known as Statistical Quality Control, range is widely used. It is also used to measure the changes in the prices of shares. Variation in daily temperatures at a certain place are measured by recording maximum temperature and minimum temperature. Range is also used in medical sciences to check whether blood pressure, haemoglobin count etc. are normal.
2. The main drawback of this measure is that it is based on only two extreme values, the maximum and the minimum, and completely ignores all the remaining observations.

## Quartile Deviation

We have seen earlier that range, as a measure of dispersion, is based only on two extreme values and fails to take into account the scatter of remaining observations within the range. To overcome this drawback to an extent, we use another measure of dispersion called Inter-Quartile Range. It represents the range which includes middle 50% of the distribution. Hence,

$$\text{Inter-Quartile Range} = Q_3 - Q_1$$

where,  $Q_3$  and  $Q_1$  represent upper and lower quartiles respectively.

Half of Inter-Quartile-Range i.e. Semi-Inter-Quartile Range =  $\frac{Q_3 - Q_1}{2}$  is also

used as absolute measure of dispersion. The semi-inter-quartile range is popularly known as Quartile Deviation (Q.D.)

$$QD = \frac{Q_3 - Q_1}{2}$$

The corresponding relative measure of dispersion is called coefficient of quartile deviation and is defined as

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

Note:

1. Q.D. is independent of extreme values. It is a better representative and more reliable than range.
2. Q.D. gives an idea about the distribution of middle half of the observations around the median.
3. Whenever median is preferred as a measure of central tendency, quartile deviation is preferred as a measure of dispersion. However, like median, quartile deviation is also not capable of further algebraic treatment, as it does not take into consideration all the values of the distribution.
4. For a symmetric distribution,

$$Q_1 = \text{Median} - Q.D. \text{ and } Q_3 = \text{Median} + Q.D.$$

## STANDARD DEVIATION (S.D.)

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical techniques.

Definition:

Standard Deviation (S.D.) is defined as the positive square root of the arithmetic mean of the squares of the deviations of the observations from their arithmetic mean.

The arithmetic mean of the squares of the deviations of the observations from their A.M. is called variance

Thus  $SD = +\sqrt{\text{variance}}$

Note: The coefficient of variation is considered to be the most appropriate measure for comparing variability of two or more distributions. The importance of C.V. can be explained with the following example: Suppose milk bags are filled with automatic machine, the amount of milk being 1 litre per bag. Setting the machine for C.V. = 0 (i.e. zero variability) is impossible, because of chance causes which exist in any process and are beyond human control. Let us assume that the machine is set for C.V. less than or equal to 1. Then, using statistical law, we can expect approximately 99.73% of the bags to contain milk quantity ranging between atleast 970 mls. and at the most 1030 mls. Usually, this variation is not noticable and hence acceptable to customers But, if the machine is set for C.V. say equal to 5, we can see that about 16% of the bags will contain 900 mls. or less of milk, which is definitely not acceptable. Thus, one has to take utmost care to reduce C.V.

In manufacturing process, with reference to quality control section and in pharmaceutical industries, C.V. plays a very important role. In quality control section, efforts are made to improve the quality by producing items as per given specifications. The extent of deviation from given specifications can be measured using C. V. The lower is the value of C.V., better is the quality of the items produced. Due to competition, almost all industries have reduced the C.V. of their goods to a considerable extent in last few years.

In pharmaceutical industries, C.V. is as low as 1 or less than 1. The variation in the weights of tablets is almost negligible.

In industrial production, C.V. depends upon raw material used. A good quality of raw material will result in homogeneous end product. In chemical and pharmaceutical industries, C.V. can be reduced by thorough mixing and pounding of the raw material.

## **MERITS AND DEMERITS OF STANDARD DEVIATION**

### **Merits:**

- 1.It is rigidly defined.
- 2.It is based on all observations.
- 3.It is capable of further algebraic treatment.
- 4.It is least affected by sampling fluctuations.

### **Demerits:**

- 1.As compared to other measures it is difficult to calculate.
- 2.It cannot be calculated for distribution with open-end class-intervals.
- 3.It gives more importance (weightage) to extreme values and less importance to the values close to A.M. that is unduly affected due to extreme observations
- 4.It cannot be calculated for qualitative data.

## Skewness and Kurtosis:

### Introduction

*“Skewness essentially is a commonly used measure in descriptive statistics that characterizes the asymmetry of a data distribution, while kurtosis determines the heaviness of the distribution tails.”*

Understanding the shape of data is crucial while practicing data science. It helps to understand where the most information lies and analyze the outliers in a given data. In this article, we'll learn about the shape of data, the importance of skewness, and kurtosis in statistics. The types of **skewness and kurtosis** and Analyze the shape of data in the given dataset. Let's first understand what skewness and kurtosis is.

### What Is Skewness?

Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side.

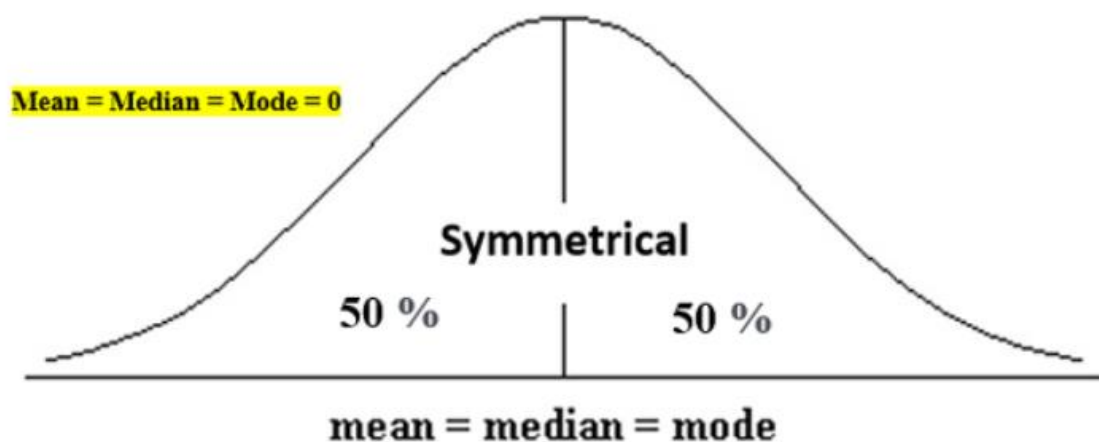
Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side. Skewness helps in understanding the shape and outliers in a dataset.

Depending on the model, skewness in the values of a specific independent variable (feature) may violate model assumptions or diminish the interpretation of feature importance.

A probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data exhibits skewness, which is a measure of asymmetry in statistics.

A skewed data set, typical values fall between the first quartile (Q1) and the third quartile (Q3).

The normal distribution helps to know a skewness. When we talk about normal distribution, data symmetrically distributed. The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.



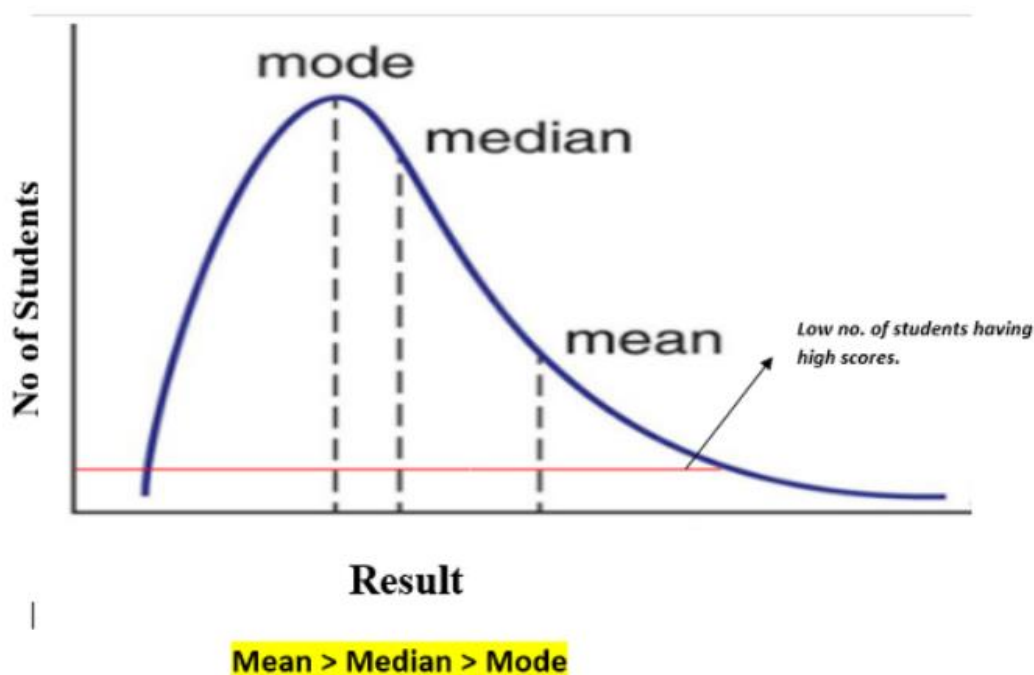
In a symmetrically distributed dataset, both the left-hand side and the right-hand side have an equal number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness comes into the picture.



## Types of Skewness

### Positive Skewed or Right-Skewed (Positive Skewness)

In statistics, a positively skewed or right-skewed distribution has a long right tail. It is a sort of distribution where the measures are dispersing, unlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other. This makes Positively Skewed Distribution a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero



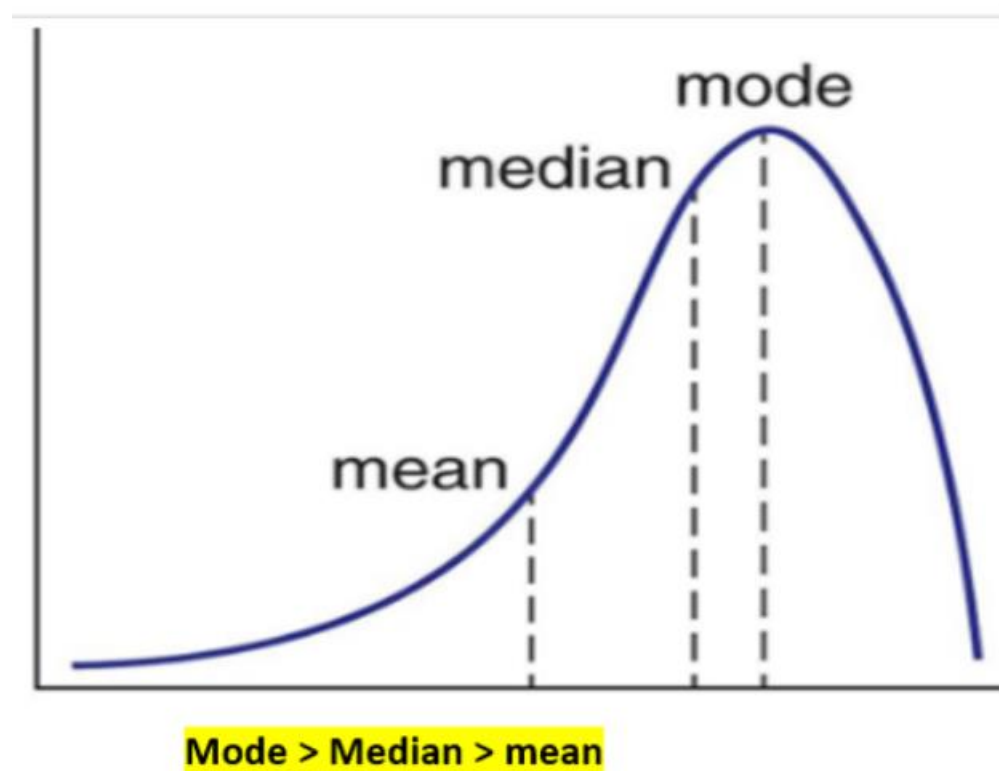
In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the most frequent value.

Extreme positive skewness is not desirable for a distribution, as a high level of skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.

## Negative Skewed or Left-Skewed (Negative Skewness)

A distribution with a long left tail, known as negatively skewed or left-skewed, stands in complete contrast to a positively skewed distribution. In statistics, negatively skewed distribution refers to the distribution model where more values are plotted on the right side of the graph, and the tail of the distribution is spreading on the left side.

In negatively skewed, the mean of the data is less than the median (a large number of data-points pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.



*Median is the middle value, and mode is the most frequent value. Due to an unbalanced distribution, the median will be higher than the mean.*

## How to Calculate the Skewness Coefficient?

Various methods can calculate skewness, with Pearson's coefficient being the most commonly used method.

### **Pearson's first coefficient of skewness**

*To calculate skewness values, subtract the mode from the mean, and then divide the difference by standard deviation.*

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

As Pearson's correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of **-1 to +1**. That accurately shows the range of the correlation values.

Pearson's first coefficient of skewness is helping if the data present high mode.

However, if the data exhibits low mode or multiple modes, it is preferable not to use Pearson's first coefficient, and instead, Pearson's second coefficient may be superior, as it does not depend on the mode.

### Pearson's second coefficient of skewness

subtract the median from the *mean*, *multiply the difference by 3*, and *divide the product by the standard deviation*.

$$\text{Pearson's second coefficient} = \frac{3 (\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\text{Mean} - \text{Mode} \approx 3 (\text{Mean} - \text{Median})$$

*Rule of thumb :*

- For skewness values between -0.5 and 0.5, the data exhibit approximate symmetry.
- Skewness values within the range of -1 and -0.5 (negative skewed) or 0.5 and 1 (positive skewed) indicate slightly skewed data distributions.
- Data with skewness values less than -1 (negative skewed) or greater than 1 (positive skewed) are considered highly skewed.

## What Is Kurtosis?

Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution.

Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution. Kurtosis helps in analyzing the characteristics and outliers of a dataset.

The measure of Kurtosis refers to the tailedness of a distribution. Tailedness refers to how often the outliers occur.

Peakedness in a data distribution is **the degree to which data values are concentrated around the mean**. Datasets with high kurtosis tend to have a distinct peak near the mean, decline rapidly, and have heavy tails. Datasets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

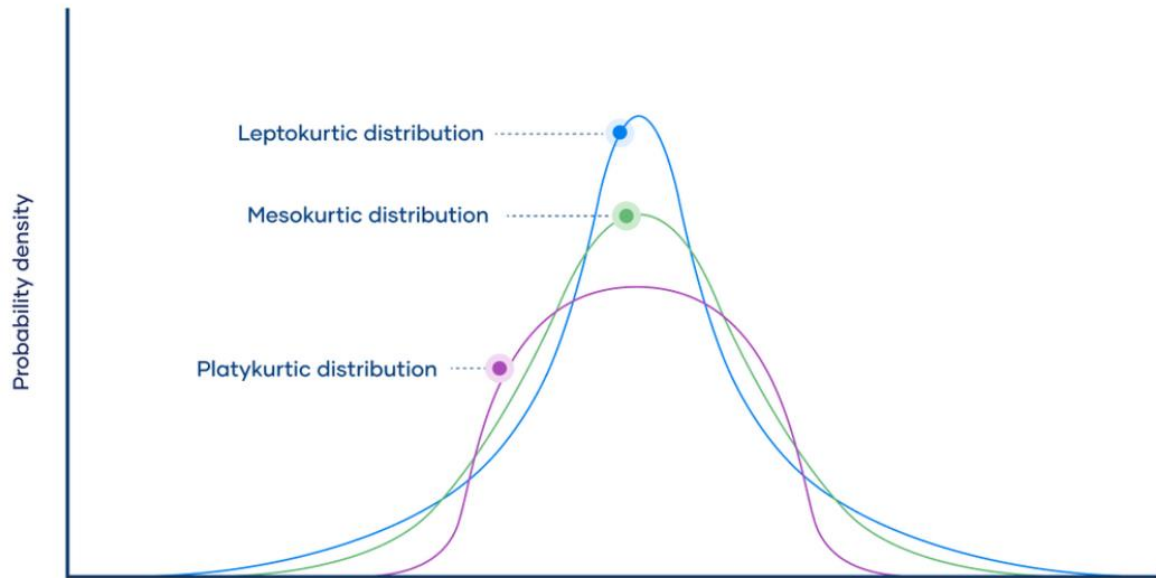
## Types of Kurtosis

Kurtosis is a statistical measure that describes the shape of a probability distribution's tails relative to its peak. There are three main types of kurtosis:

1. **Mesokurtic:** A distribution with mesokurtic kurtosis has a similar peak and tail shape as the normal distribution. It has a kurtosis value of around 0, indicating that its tails are neither too heavy nor too light compared to a normal distribution.
2. **Leptokurtic:** A distribution with leptokurtic kurtosis has heavier tails and a sharper peak than the normal distribution. It has a positive kurtosis value, indicating that it has more extreme outliers than a normal distribution. This type of distribution is often associated with higher peakedness and a greater probability of extreme values.
3. **Platykurtic:** A distribution with platykurtic kurtosis has lighter tails and a flatter peak than the normal distribution. It has a negative kurtosis value, indicating that it has fewer extreme outliers than a normal distribution. This type of distribution is often associated with less peakedness and a lower probability of extreme values.

## Types of Excess Kurtosis

1. *Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).*
2. *Mesokurtic (kurtosis same as the normal distribution).*
3. *Platykurtic or short-tailed distribution (kurtosis less than normal distribution).*



### ***Leptokurtic ( $Kurtosis > 3$ )***

Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. Extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.

### ***Platykurtic ( $Kurtosis < 3$ )***

Platykurtic having a thin tail and stretched around the center means most data points are present in high proximity to the mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

### ***Mesokurtic ( $Kurtosis = 3$ )***

Mesokurtic is the same as the normal distribution, which means kurtosis is near 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

## Difference Between Skewness and Kurtosis

### *Key Differences of Skewness and Kurtosis*

1. Skewness evaluates how much a distribution deviates from symmetry, while Kurtosis gauges the degree of its peakiness or flatness.
2. Skewness is a measure derived from the third moment, whereas Kurtosis stems from the fourth moment.
3. The range of values for both Skewness and Kurtosis spans from negative infinity to positive infinity.
4. Perfect symmetry and normality are indicated by both zero skewness and zero kurtosis.
5. Skewness can impact the central tendency of a distribution, whereas kurtosis can influence its tail behavior.
6. Both Skewness and Kurtosis provide insights into the shape characteristics of distributions.



Session 11 & 12

Session 13 & 14