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.

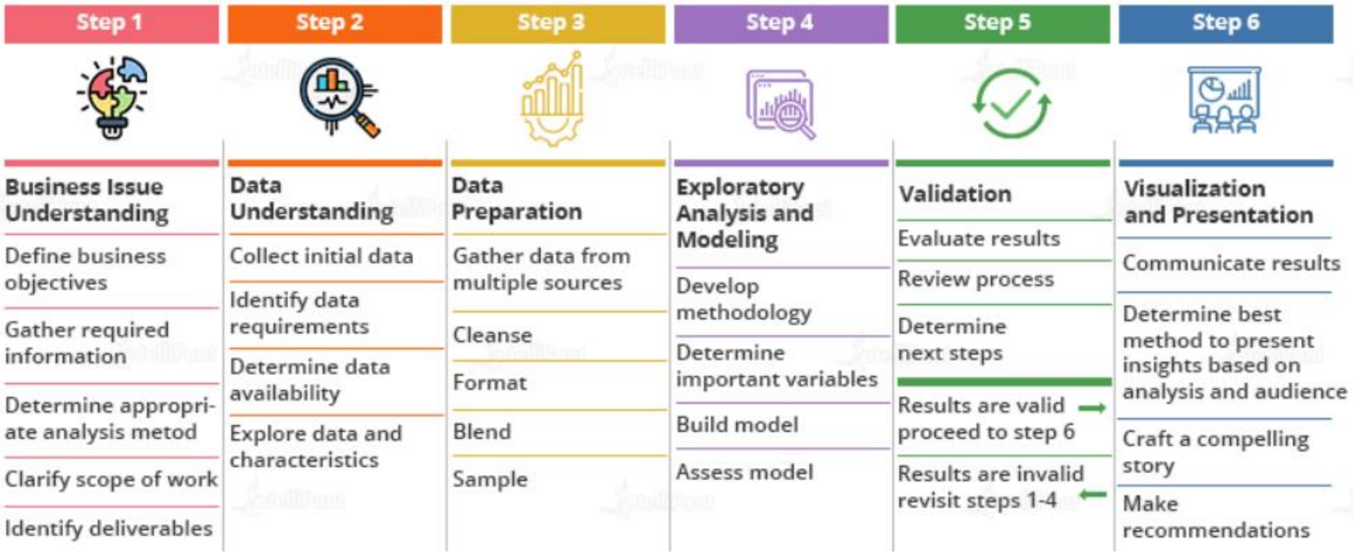
Monitoring and Maintenance:

- 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

Business Issue Understanding

- Define business objectives
- Gather required information
- Determine appropriate analysis metod
- Clarify scope of work
- Identify deliverables

Step 2

Data Understanding

- Collect initial data
- Identify data requirements
- Determine data availability
- Explore data and characteristics

Step 3

Data Preparation

- Gather data from multiple sources
- Cleanse
- Format
- Blend
- Sample

Step 4

Exploratory Analysis and Modeling

- Develop methodology
- Determine important variables
- Build model
- Assess model

Step 5

Validation

- Evaluate results
- Review process
- Determine next steps
- Results are valid proceed to step 6 →
- Results are invalid revisit steps 1-4 ←

Step 6

Visualization and Presentation

- Communicate results
- Determine best method to present insights based on analysis and audience
- Craft a compelling story
- Make recommendations

Session 3 & 4

Prerequisites:

Factorial Notation:

We define, $0! = 1$

For any positive integer n,

$n! = n\times(n-1)\times(n-2)\times\ldots\times1$

e.g.

$1! = 1$

$2! = 2\times1$

$3! = 3\times2\times1$

$4! = 4\times3\times2\times1$ and so on.

Consider $6! = 6\times5\times4\times3\times2\times1$, which we can write as

$6! = 6\times\mathbf{(5\times4\times3\times2\times1)} = 6\times\mathbf{5!}$ $\therefore$ **$\mathbf{n! = n\times[(n-1)!]}$**

$6! = 6\times5\times\mathbf{(4\times3\times2\times1)} = 6\times5\times\mathbf{4!}$

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

Permutation:

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

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

$$\mathrm{nPr} = \frac{n!}{(n-r)!}$$

e.g.

$$\mathrm{nP0} = \frac{n!}{(n-0)!} = 1$$

$$\mathrm{nP1} = \frac{n!}{(n-1)!} = \frac{n(n-1)!}{(n-1)!} = n$$

$$\mathrm{nPn} = \frac{n!}{(n-n)!} = \mathrm{n!}$$

Combination:

Consider selection of ‘r’ objects out of n (r ≤ 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

$$\text{nCr} = \frac{n!}{r!\,(n-r)!}$$

e.g.

$$\text{nC0} = \frac{n!}{0!\,(n-0)!} = \frac{n!}{n!} = 1$$

$$\text{nC1} = \frac{n!}{1!\,(n-1)!} = \frac{n(n-1)!}{(n-1)!} = n$$

$$\text{nCn} = \frac{n!}{1!\,(n-n)!} = \frac{n!}{n!\,0!} = 1$$

$$10\text{C}3 = \frac{10!}{3!\,(10-3)!} = \frac{10!}{3!\,7!}$$

$$10\text{C}7 = \frac{10!}{7!\,(10-7)!} = \frac{10!}{7!\,3!}$$

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

$$\boxed{\textbf{In general, } {}^{n}\mathbf{C}_{r} = {}^{n}\mathbf{C}_{n-r}}$$

$^{10}C_3 = \frac{10!}{3!7!} = \frac{10*9*8*7!}{3!7!} = \frac{10*9*8}{3!}$; $^{12}C_4 =$ ____________

$^{100}C_{97} = {}^{100}C_3 =$ ____________

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 (Ω). 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.

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.

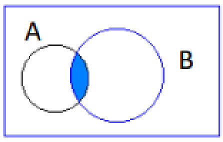
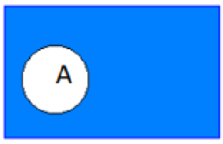
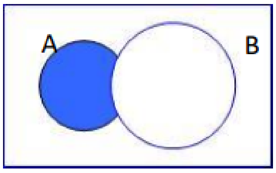
iii. It is concerned with the planning and controlling and with the occurrence of accidents of all kinds.

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 ….. 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 ….. occurring.
7	Complementation	c or $-$	A^c or $\overline{A}$	The event of A not occurring. 
8	Subtraction	-	A-B	The event of A occurring and B not occurring. 

Operation	Interpretation
A', A 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 φ, $P(\varphi) = 0$

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

5) If A1 and A2 two mutually exclusive events then $P(A1 \cup A2\ ) = P(A1\ ) + P(A2\ )$

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 , ..., A_m are mutually exclusive events in S, then

$$P(A_1 \cup A_2 \cup, ... \cup A_m = P(A_1\ ) + P(A_2\ ) + ... + P(A_m)$$

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

Note:
$P(A) + P(\bar{A}) = 1$
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) ≠0 is called conditional probability of event A given B and is denoted by P(A/B).

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

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∩B)

and is given by P(A∩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) and P(B/A) = P(B/A') = P(B)

If A and B are independent events then P(A∩B) = P(A).P(B)

(P(A∩B) = P(A).(B/A) = P(A).P(B) ∴ P(A∩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 1/3 and 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

$$\overset{\text{posterior}}{P(A|B)} = \frac{\overset{\text{likelihood}}{P(B|A)} * \overset{\text{prior}}{P(A)}}{\underset{\text{evidence}}{P(B)}}$$

<u>Ex-22:</u> 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.

<u>Ex-23:</u> 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'.

$\therefore P(A) = \frac{m}{n}$ 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 $\frac{n- m}{n}$ i.e. P(A'):P(A

Session 5 & 6

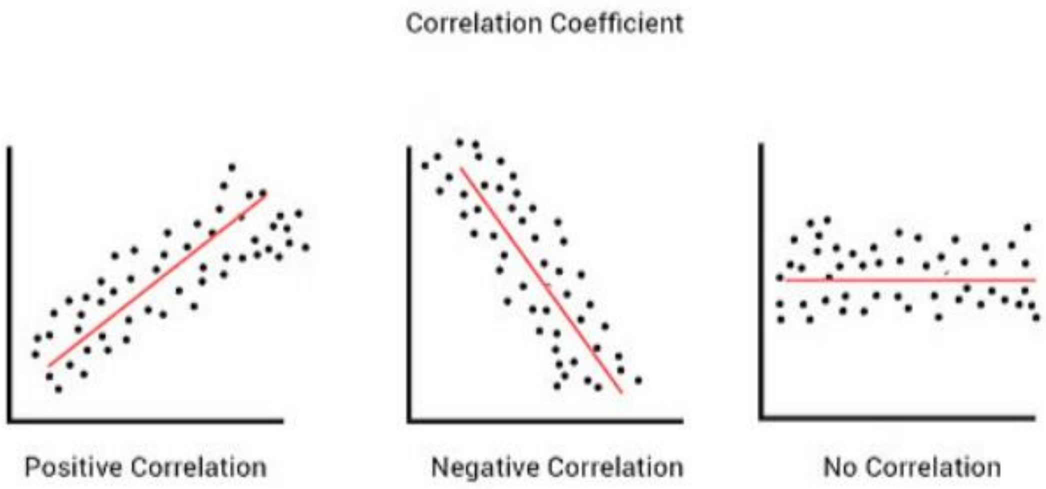
What Is Correlation?

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:

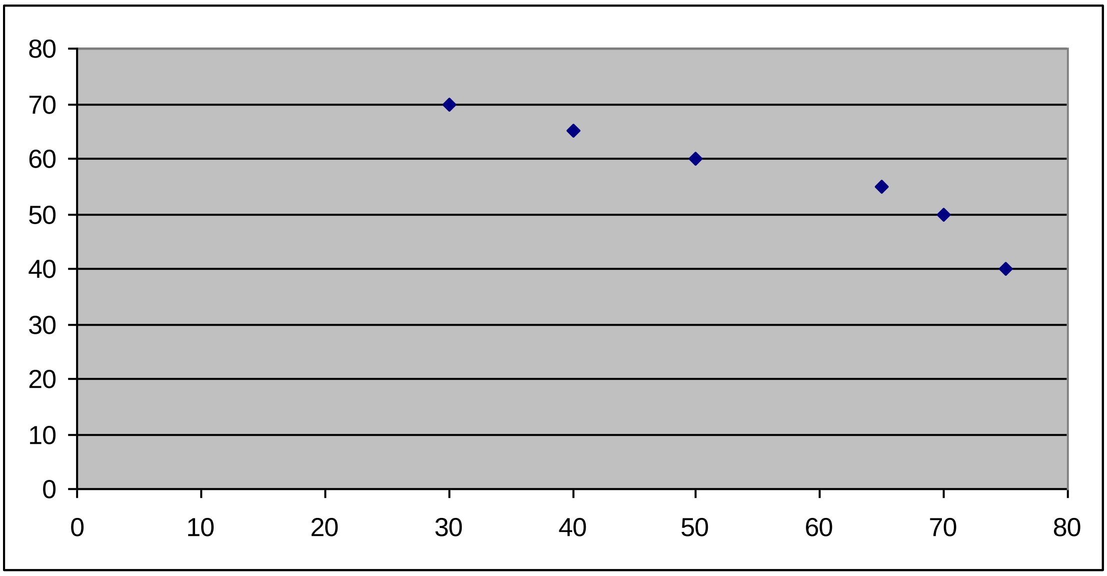
We collect a data of pairs of values of the two variables. Generally these variables are denoted by x & y. These values are considered as x & y co-ordinates respectively and plotted as points on a graph. Such diagrammatic representation of a bivariate data is called as a scatter diagram. From the scatter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

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

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 Cov(X,Y). The formula is given below for both population covariance and sample covariance.

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

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

Session 7 & 8

Session 9 & 10

Session 11 & 12

Session 13 & 14