Basic Probability

DATAPLAY

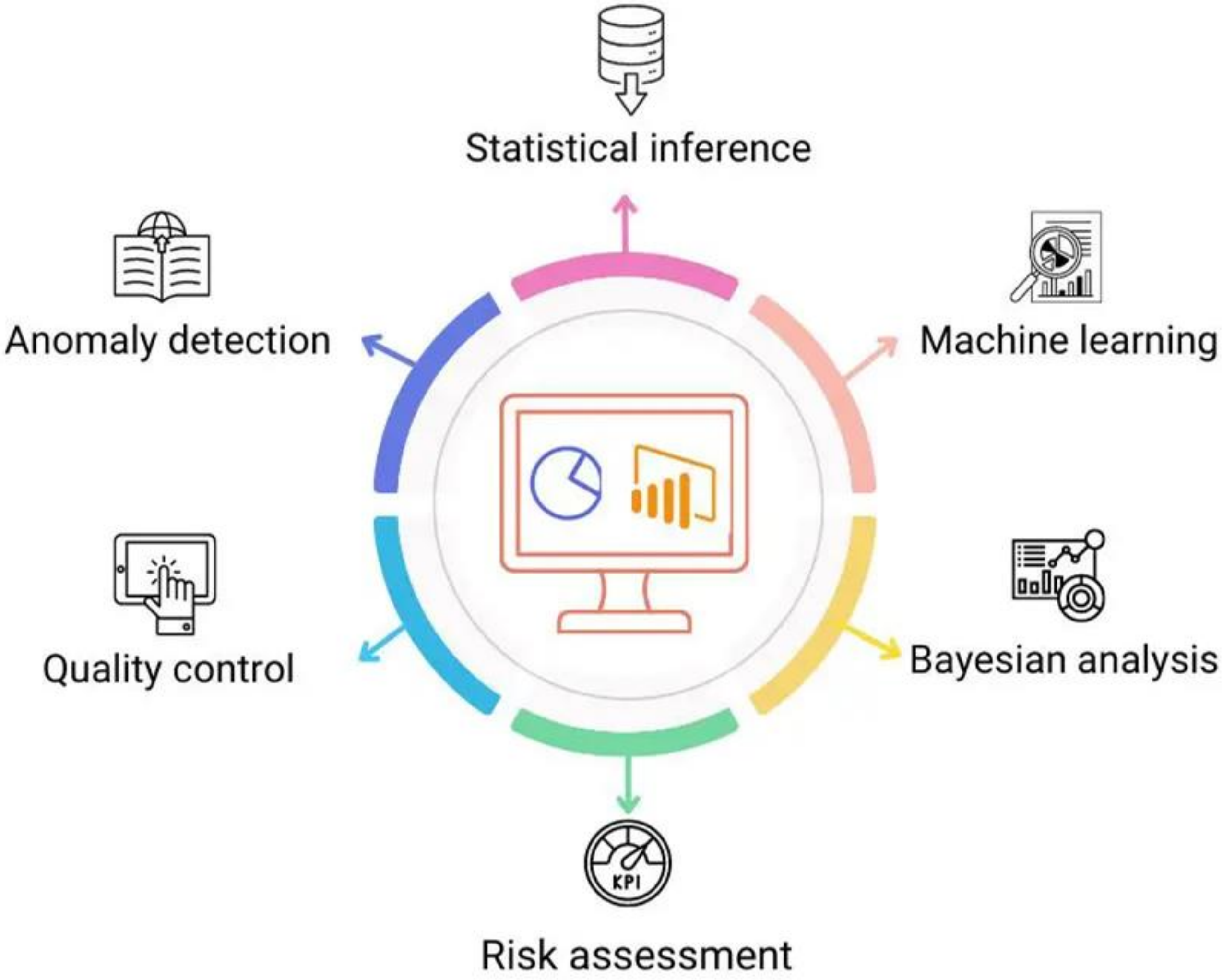
- Applications of Probability in Data Science
- What is Probability?
- Sets and Set Theory (Union and Intersection)
- Random Experiment, Sample Space, Event, Trial
- Types of Events
  - Equally Likely
  - Mutually Exclusive
  - Collectively Exhaustive
  - Independent Event
  - Dependent Event
- Additive And Multiplicative Rule of Probability
- Conditional Probability
- Factorial
- Permutation and Combination
- Gender Prediction using Naïve Bayes Algorithm

Advance Probability

- Random Variable vs Algebraic Variable
  - Discrete Random Variable v/s Continuous Random Variable
- Probability Distribution Function
  - Probability Density Function (pdf)
  - Probability Mass Function (pmf)
- Cumulative Distribution Function
  - Cumulative Density Function (cdf)
  - Cumulative Mass Function (cmf)
- Mean v/s Expected Value
  - Mean and Variance of Random Variable
- Types of Probability Distribution
  - Discrete Probability Distributions
    - Uniform Distribution
    - Bernoulli Distribution
    - Binomial Distribution
    - Geometric Distribution
    - Poisson Distribution

- Types of Probability Distribution
  - Uniform Distribution
  - Gaussian / Normal Distribution
  - Standard Normal / Z Distribution
  - Skewed Normal Distribution
  - Exponential Distribution
- Information Theory and Entropy
- Joint Distribution/ Joint Probability
- Marginal Probability
- Probability v/s Likelihood v/s Log Odds
- Statistical Relationships between two Random Variables
- Causality , Correlation , Covariance and Dependence
- Relationship among statistical Relationships
- Pearson and Spearman rank correlation

Application of Probability in Data Science



What is Probability?

DATAPLAY

Layman Explanation : Probability is a way of dealing with uncertainty or randomness. Probability is a measure of how likely an event is to occur, the chances of its occurrence.

Technical Explanation : Probability is the long-term frequency of an event.

We mean that in practice, the probability of an event occurring can be estimated or understood by observing its frequency over a large number of trials or occurrences.The concept relies on the idea that if an experiment (like flipping a coin) is repeated many times under similar conditions, the proportion of times the event (like heads) occurs will converge towards its probability.

Formulating Probability

DATAPLAY

Calculating Probability : To find the probability of an event, divide the number of ways it can happen by the total possible outcomes.

Range -> [0-1]

0: Impossible (Event won't happen)
1: Certain (Event will definitely happen)

$$\text{Probability} = \frac{\text{Outcome of Interest}}{\text{All possible outcomes}} = \frac{m}{n}$$

Theoretical probability (Classical or a priori probability)
Experimental probability (Empirical or Observed probability)

DATAPLAY

Experimental Probability vs. Theoretical Probability

Experimental Probability is found by repeating an experiment and observing the outcomes.

$$P(\text{event}) = \frac{\text{number of times event occurs}}{\text{total number of trials}}$$

Example:

A coin is tossed 10 times:
A head is recorded 7 times and a tail 3 times.

$$P(\text{head}) = \frac{7}{10}$$

$$P(\text{tail}) = \frac{3}{10}$$

Theoretical Probability is what is expected to happen based on mathematics

$$P(\text{event}) = \frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$$

Example:

A coin is tossed.

$$P(\text{head}) = \frac{1}{2}$$

$$P(\text{tail}) = \frac{1}{2}$$

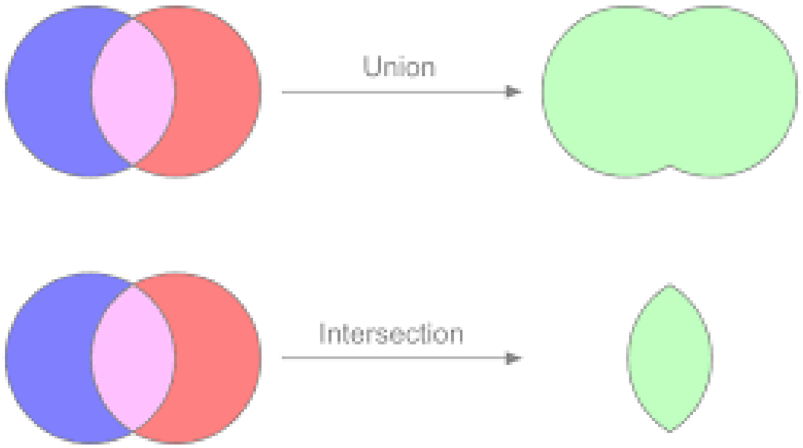
Theoretical probability is expected to match experimental probability when the theoretical model accurately reflects the real-world situation, especially in simple and idealized scenarios.

Sets and Set Theory

Sets: Collections of unique objects, like {2, 4, 6, 8}. No order, no duplicates.

Key Ideas:

Subset: A set within another set (e.g., {2, 4} is a subset of {2, 4, 6, 8}).
Union (OR): Combine elements from two sets (e.g., {2, 4, 6} ∪ {6, 8} = {2, 4, 6, 8}).
Intersection (AND): Elements common to both sets (e.g., {1, 2} ∩ {3, 2} = {2}).



Complement: Elements not in the set (e.g., {2, 4} and {6, 8} are complements in set {2, 4, 6, 8}).
Cardinality: Number of elements in a set (e.g., |{2, 4, 6, 8}| = 4).
Universal Set: All relevant elements in a specific situation

Random Experiment, Sample Space, Event, Trial

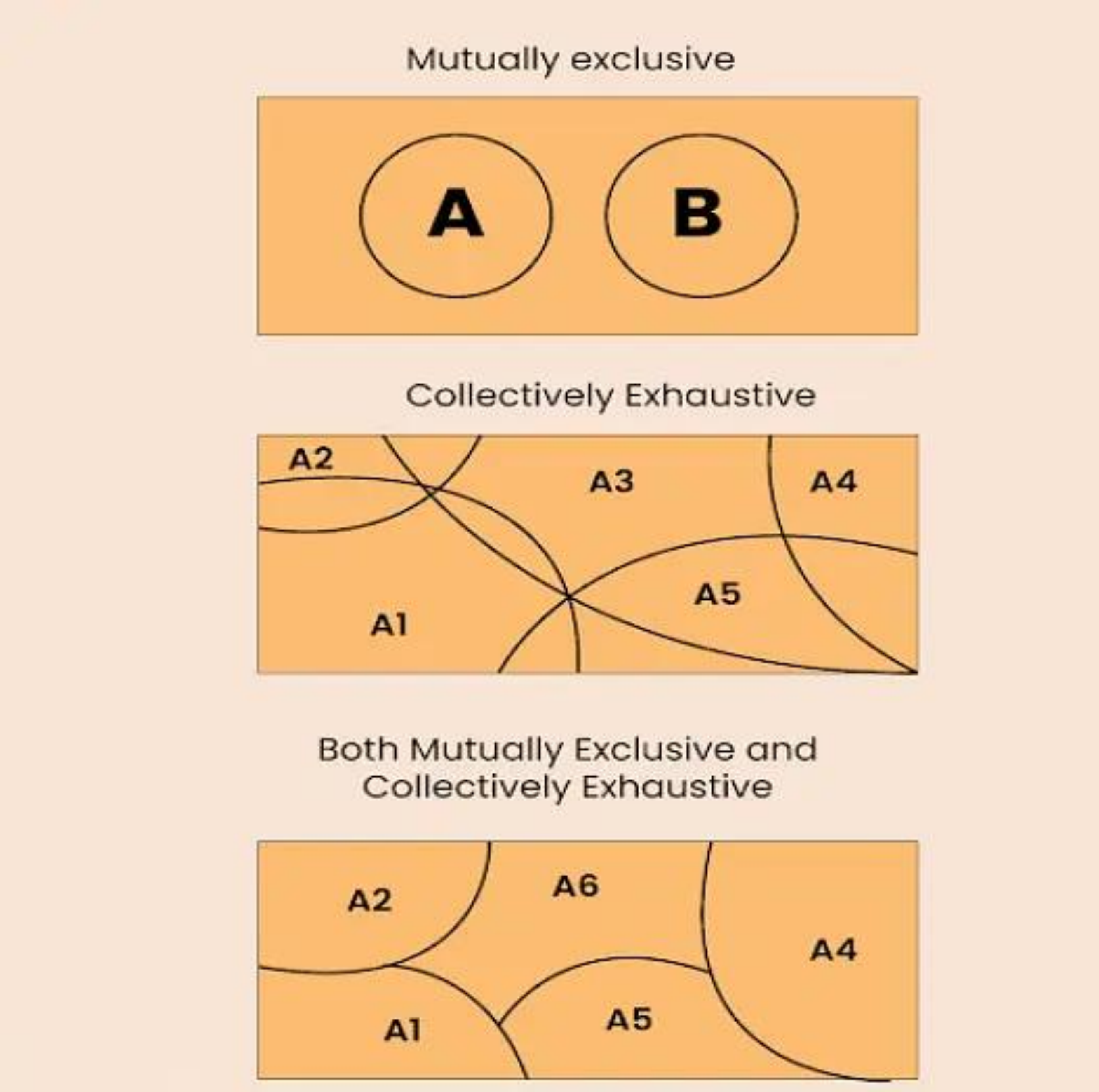
DATAPLAY

Concept	Definition	Example
Random Experiment	A process or procedure with uncertain outcomes. The exact outcome of the process is not known in advance and can differ from one trial to another.	1. Rolling a fair six-sided die 2. Tossing a coin
Sample Space	The set of all possible outcomes of an experiment.	1. Sample space of rolling a die: {1, 2, 3, 4, 5, 6} 2. Sample space of tossing a coin: {H,T}
Event	A subset of the sample space representing a specific outcome or a combination of outcomes.	1. Event "getting an even number": {2, 4, 6} 2. Event "getting a prime number": {2, 3, 5} 3. Event: Getting Heads : {H} 4. Event : Getting a Heads or a Tail : {H,T}
Trial	A trial is a single performance of an experiment, activity, or process that can result in multiple possible outcomes.	1. A single flip of a fair coin represents one trial. 2. A single roll of a dice represents one trial.

Type of Events

DATAPLAY

Event Type	Definition	Example	Formula
Equally Likely Events	Events with equal probabilities of occurring.	Rolling a fair six-sided die, each outcome (1, 2, 3, 4, 5, 6) is equally likely.	$P(E_1) = P(E_2)$
Collectively/ Mutually Exhaustive Events	A set of events that together cover the entire sample space i.e. encompass all possible outcomes.	“H” and “T” are mutually exhaustive events in a coin toss. Red, blue, and patterned shirts cover all possibilities (exhaustive), but a shirt can belong to more than one category (not exclusive).	$\cup_{i=1 \text{ to } n}(E_i) = S$ The events need not be all mutually exclusive.
Mutually Exclusive Events	Events that cannot occur simultaneously; the occurrence of one event prevents the others from happening.	"Heads" and "Tails" are mutually exclusive outcomes while tossing a coin. “King” and “Hearts” while picking a card are not mutually exclusive.	$P(E_1 \text{ and } E_2) = 0$


Mutually exclusive
A
B
Collectively Exhaustive
A2
A3
A4
A5
A1
Both Mutually Exclusive and
Collectively Exhaustive
A2
A6
A4
A5
A1

Type of Events

Dependent Events	Events where the outcome of one event affects the outcome of another.	Drawing cards without replacement: The probability of the second card depends on the outcome of the first.	$P(E_1 \cap E_2) = P(E_2)*P(E_1\|E_2)$ Used in conditional probability to find the probability of one event given the occurrence of another.
Independent Events	Events where the outcome of one event does not affect the outcome of another.	Tossing a coin: Multiple coin tosses are typically independent events. Drawing Cards with replacement.	Two events E and F are said to be independent, if $P(F\|E) = P(F)$ provided $P(E) \neq 0$ and $P(E\|F) = P(E)$ provided $P(F) \neq 0$ $P(E_1 \cap E_2) = P(E_1)*P(E_2)$

Additive Rule of Probability (Using Set Theory)

DATAPLAY

A dice is rolled once and we want to calculate the probability of occurrence of 2 or even in this trial.

S = {1,2,3,4,5,6}
E1 = {2}
E2 = {2,4,6}
P(E1 or E2) = ?

P(E1) = 1/6
P(E2) = 3/6 = 1/2

E1 or E2 = E1 U E2
= E1 + E2 - (E1 ∩ E2)
E1 or E2 = {2} + {2,4,6} - {2}

P(E1 or E2) = P{E1} + P{E2} - P{E1 ∩ E2}
= 1/6 + 1/2 - 1/6
= 1/2

E1 or E2 = E1 U E2
= E1 + E2 - (E1 ∩ E2)
P(E1 or E2) = P{E1} + P{E2} - P{E1 ∩ E2}

Multiplicative Rule of Probability (Using Set Theory)

DATAPLAY

A dice is rolled twice and we want to calculate the probability of occurrence of 2 and 3 respectively in consecutive rolls.

$S1 = \{1,2,3,4,5,6\}$ -> $E1 = \{2\}$ -> $P(E1) = 1/6$
$S2 = \{1,2,3,4,5,6\}$ -> $E2 = \{3\}$ -> $P(E2) = 1/6$

$P(E1 \cap E2) = ?$

Total Sample space of 2 trials = S1*S2 =
{(1,1),(1,2),(1,3),......(3,3),(3,4),.......(6,4),(6,5),(6,6)}
= 36 possible outcomes

Desired Outcome = {(2,3)}

So $P(E1 \cap E2) = 1/36$
$= 1/6 * 1/6$
$= P(E1)*P(E2)$

$E1 \cap E2 = \{E1\} * \{E2\}$
$P(E1 \cap E2) = \{E1\} * \{E2\} / \{S1\}*\{S2\}$
$= P(E1) * P(E2)$
Note : E1 and E2 are independent

Q. A dice is rolled twice. Calculate the probability of occurrence of even and odd respectively in consecutive rolls.

Additive and Multiplicative Rule of Probability

DATAPLAY

Rule	Description	When to Use	Formula	Example
Additive Rule	Probability of "either or both" events	Events can be overlapping (not mutually exclusive)	P(A or B) = P(A) + P(B) - P(A and B)	Finding the probability of rolling a 2 or an even number on a dice (2, 4, or 6).
Multiplicative Rule	Probability of "both" independent events	Events are independent (occurring one doesn't affect the other)	P(A and B) = P(A) * P(B)	Finding the probability of drawing a red card and then an ace from a deck with replacement.

Example of Additive and Multiplicative Rule

Example 1: Let's say you are rolling a standard six-sided die twice. You want to find the probability of rolling either a 3 or a 4 (event A) and then rolling an even number (event B).

- **Additive Rule:** To find the probability of event A i.e. probability of rolling either a 3 or a 4, you can use the additive rule:
  P(3 or 4) = P(3) + P(4) = 1/6 + 1/6 = 1/3

- **Multiplicative Rule:** To find the probability of both event A and event B occurring, you can use the multiplicative rule:
  P((3 or 4) and even) = P(3 or 4) * P(even) = 1/3 * 3/6 = 1/6

Using set theory :
P(3 or 4) = P(3) U P(4) = {3,4} / {1,2,3,4,5,6} = 1/3

P(even) = P(2,4,6) = {2,4,6} / {1,2,3,4,5,6} = 1/2

P(3 or 4) and P(even) = {(3,2),(3,4),(3,6),(4,2),(4,4),(4,6)} / {(1,1),(1,2),(1,3)......(4,1),(4,2),.....(5,5),(5,6)} = 6/36 = 1/6

Example 2: Drawing Cards from a Deck Suppose you have a standard deck of 52 playing cards.

You want to find the probability of drawing a red card lets say event A (hearts or diamonds) and then drawing a face card lets say event B(king, queen, or jack) with replacement.

- To find the probability of event A occurring, you can use the additive rule:
  P(A) = P(hearts or diamonds) = P(hearts) + P(diamonds) - P(hearts and diamonds)

  P(hearts or diamonds) = 13/52 + 13/52 = ½

  P(B) = P(king, queen, or jack) = 12/52

- To find the probability of both event A and event B occurring (A and B are independent events), you can use the multiplicative rule:

- P(A and B) = P(A) * P(B)
  P(red and face card) = P(red) * P(face card) = (1/2) * (12/52) = 6/52 ≈ 0.115

  Solve this using set theory too. Calculate for without replacement case as well.

Practice

Q1. Suppose you have a box with 3 blue marbles, 2 red marbles, and 4 yellow marbles. You are going to pull out one marble, record its color, put it back in the box and draw another marble. What is the probability of pulling out a red marble followed by a blue marble?

Q2. A coin is flipped three times. What is the probability of getting heads on the first flip, tails on the second flip, and heads on the third flip?

Q3. You have a bag containing 5 green candies and 3 red candies. What is the probability of randomly selecting either a green candy or a red candy from the bag?

Q4. Bag A contains 3 Red Marbles and 4 Blue Marbles. Bag B contains 5 Red Marbles and 3 Blue Marbles. A marble is taken from each bag in turn. What is the probability of getting a marble of each colour?

Q5. In a deck of 52 playing cards, what is the probability of drawing either an Ace or a King, and then a Queen or a Jack, if the cards are drawn without replacement?

Q6. In a drawer, there are 8 white socks, 4 black socks, and 3 blue socks. If you randomly pick 2 socks without replacement, what is the probability that you pick at least one black sock? (Do without using P&C)

Conditional Probability

DATAPLAY

Conditional probability measures the likelihood of an event occurring given that another event has already occurred.

Generic formula for conditional probability is: $P(A|B) = P(A \cap B) / P(B)$ {Can be explained by Set theory}

Where:

$P(A|B)$ represents the conditional probability of event A occurring given that event B has already occurred.

$P(A \cap B)$ denotes the probability of both events A and B occurring simultaneously.

$P(B)$ represents the probability of event B occurring.

Intuition:

We restrict our attention to the sample space where event B has occurred and then calculate the probability of A within that restricted space. Conditional probability allows us to adjust probabilities based on new information, making it a fundamental concept in probability theory and statistics. It's widely used in various applications, including machine learning, decision making, and real-world problem-solving.

Conditional Probability is the base for Naive Bayes Algorithm where $P(A|B)$ posterior probability and $P(A)$ is prior probability.

Example of Conditional Probability

Drawing Marbles from a Bag

Find the probability of drawing a blue marble (B) from a bag with three red marbles (R) and two blue marbles (B) after already drawing a red marble, without replacement.

- P(R) is the probability of drawing a red marble: $P(R) = 3/5$.
- P(B) is the probability of drawing a blue marble: $P(B) = 2/5$.

$$P(B \mid R) = 2/4 = 0.5 \text{ (intuitively)}$$

P(B and R): The probability of drawing a red marble (R) and then a blue marble (B) in that order. Since you've already drawn the red marble, this is the probability of drawing a blue marble second.

$$P(B \mid R) = P(B \cap R) / P(R)$$

$$P(B \cap R) = P(B)*P(R|B) = (\tfrac{2}{5})*(\tfrac{3}{4}) = 3/10.$$

$$P(B|R) = (3/10)/(3/5) = 0.5$$

Example

DATAPLAY

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(\mathrm{A}) = 0.65$. B = the event Carlos is successful on his second attempt. $P(\mathrm{B}) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

a. What is the probability that he makes both goals?
b. What is the probability that Carlos makes either the first goal or the second goal?
c. Are A and B independent?
d. Are A and B mutually exclusive?

Solutions

a. The problem is asking you to find $P(\mathrm{A\ AND\ B}) = P(\mathrm{B\ AND\ A})$. Since $P(\mathrm{B}|\mathrm{A}) = 0.90 : P(\mathrm{B\ AND\ A}) = P(\mathrm{B}|\mathrm{A})P(\mathrm{A}) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

b. The problem is asking you to find $P(\mathrm{A\ OR\ B})$.

$$P(\mathrm{A\ OR\ B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A\ AND\ B}) = 0.65 + 0.65 - 0.585 = 0.715 \tag{4.3.3}$$

Carlos makes either the first goal or the second goal with probability 0.715.

c. No, they are not, because $P(\mathrm{B\ AND\ A}) = 0.585$.

$$P(\mathrm{B})P(\mathrm{A}) = (0.65)(0.65) = 0.423 \tag{4.3.4}$$

$$0.423 \neq 0.585 = P(\mathrm{B\ AND\ A}) \tag{4.3.5}$$

So, $P(\mathrm{B\ AND\ A})$ is **not** equal to $P(\mathrm{B})P(\mathrm{A})$.

d. No, they are not because $P(\mathrm{A\ and\ B}) = 0.585$.

To be mutually exclusive, $P(\mathrm{A\ AND\ B})$ must equal zero.

Factorial

The number of arrangements of n different objects taken r at a time, where $0<r\leq n$ and the objects do not repeat is $n(n–1)(n–2)...(n–r+1)$.

There will be as many arrangements as there are ways of filling in r vacant places by the n objects.

The first place can be filled in n ways; following which, the second place can be filled in $(n – 1)$ ways, following which the third place can be filled in $(n – 2)$ ways,..., the rth place can be filled in $(n – (r – 1))$ ways.

Therefore, the number of ways of filling in r vacant places in succession is $n(n–1)(n–2)... (n–(r–1))$ or $n (n – 1) (n – 2) ... (n – r + 1)$

This expression is cumbersome and we need a notation which will help to reduce the size of this expression.

n! (factorial n or n factorial) = $n (n – 1) (n – 2) ... (n – r + 1)...... 3 * 2 * 1$ = number of ways for arranging n items

Therefore, the number of ways of filling in r vacant places in succession is $n! / (n-r)!$

=> Permutation formula

The number of arrangements of n different objects taken r at a time, where repetition is allowed, is n^r.

Note : $0!=1$

Permutation and Combination

DATAPLAY

Criteria	Permutation	Combination
Definition	A permutation is an arrangement in a definite order of a number of objects taken some or all at a time. Combination+Arrangement	Combinations are selections of objects in which order does not matter.
Formula	P(n,r) = n! / (n - r)! , 0<r≤n	C(n,r) = n! / (r! * (n - r)!) = nPr / r! , 0<r≤n
Example	You have three different toys: A, B, and C. You want to explore the different ways you can arrange 2 toys out of them on a shelf in a specific order. $^{3}P_{2} = 3! / (3-2)! = 6$ AB, BC, AC, BA, CB, CA are the possible ordered arrangements.	You have three different toys: A, B, and C. You want to explore the different ways you can arrange 2 toys out of them on a shelf without considering order or you can say you have to select two toys out of given 3. $^{3}C_{2} = 3! / ((3-2)! * 2!) = 3$ AB, BC, AC are the possible arrangements.
Key Words	Arrange Line Up Order	Choose Select Pick

Formula Derivation

1. $^nC_a = {^nC_b}$

   $\Rightarrow a=b$ or $a=n-b$,

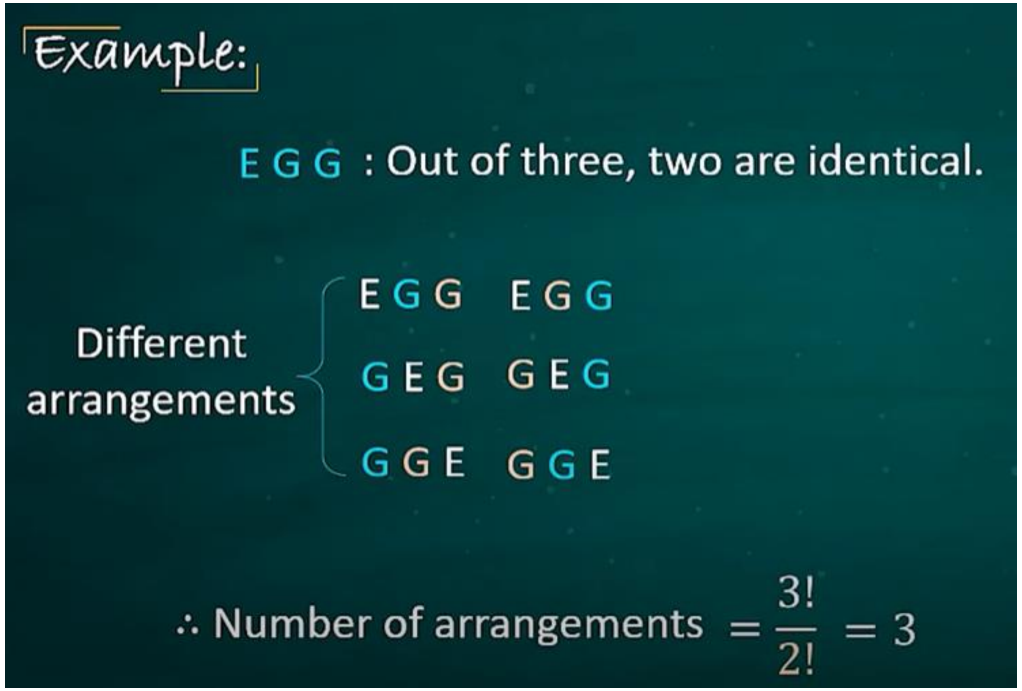
   i.e.,$n=a+b$

1. $^nC_r + {^nC_{r-1}} = {^{n+1}C_r}$

Possible ways of Permutation in case of repetition

The number of permutations of “n” objects, “r” of which are alike, “s” of which are alike, ‘t” of which are alike, and so on, is given by the expression

$$\boxed{\frac{n!}{r! \times s! \times t! \ldots}}$$



Permutation with fixed positions

Example : In how many ways can a group of 8 people line up for a photo if 3 of them must be at the ends of the line?

Possible permutations = (n-k)! * k!

= (8-3)! * 3!

= 5! * 3!

Permutation with restricted positions

Example : In how many ways can 5 men and 3 women be arranged in a row if no two women are standing next to one another?

Possible permutations = _ M1 _ M2 _ M3 _ M4 _ M5 _

= 5! * 6P3

= 5! * (6! / 3!)

Example : In how many ways can the letters in the word SUCCESS be arranged if no two S's are next to one another?

Possible Permutations = _ U _ C _ C _ E _

= (4! / 2!) * (5C3)

= (4! / 2!) * (5! / (2!*3!))

Permutation and Combination Practice

DATAPLAY

1. Given 4 flags of different colours, how many different signals can be generated, if a signal requires the use of 2 flags one below the other?

1. How many 2 digit even numbers can be formed from the digits 1, 2, 3, 4, 5 if the digits can be repeated?

1. Find the number of different signals that can be generated by arranging at least 2 flags in order (one below the other) on a vertical staff, if five different flags are available.
2. How many 3-digit even numbers can be formed from the digits 1, 2, 3, 4, 5, 6 if the digits can be repeated/can't be repeated?

3. How many 4-letter code can be formed using the first 10 letters of the English alphabet, if no letter can be repeated?

4. How many 5-digit telephone numbers can be constructed using the digits 0 to 9 if each number starts with 67 and no digit appears more than once?

5. A coin is tossed 3 times and the outcomes are recorded. How many possible outcomes are there?

6. Twelve persons meet in a room and each shakes hand with all the others. Determine the number of handshakes.

7. Seven points lie on a circle. How many chords can be drawn by joining these points pairwise?

8. What is the probability that two friends have different birthdays in a normal year?

9. Discuss Orientation Test Paper

Gender Prediction using Naïve Bayes Algorithm

DATAPLAY

Step 1: Calculate Prior Probabilities

- P(Male)=Number of Males / Total Number of Samples
- P(Female)=Number of Females / Total Number of Samples

Step 2: Calculate Likelihood

- For each last letter, calculate the likelihood for each gender:

Last Letter = a, b, c, d……,z

P(Last Letter | Male)

P(Last Letter | Female)

Step 3: Prediction Using Posterior Probability

To predict the gender based on a new name, say "Linda":

- Last letter is 'a'.

Calculate the posterior probabilities using Bayes' theorem:

$P(Male|a) = (P(a|Male) \times P(Male)) / P(a)$
$P(Female|a) = (P(a|Female) \times P(Female)) / P(a)$

Where $P(a) = P(a|Male) \times P(Male) + P(a|Female) \times P(Female)$