Probability and Statistics

Prepared by Syed Mohammad Hassan

Student Data Science and Artificial Intelligence Batch-II

Saylani Mass I.T Training Program

Introduction to Probability

- What is Probability?
  - Probability is the branch of mathematics that deals with that how likely an event is to occur.
  - In simple words chances of an event to occur. Whether something will happen or not.
- Rules and Conventions:
  - Probability of an event can range from 0 to 1.
  - Sum of probabilities of all Sample points in a Sample space is equal to 1.
  - Probability of event A is P(A).
- Interpretations of P(A):
  - If P(A) = 0, Event A will definitely not occur.
  - If P(A) is close to Zero, There are low chances that the Event A will occur.
  - If P(A) = 0.5, Chances are 50,50 that the Event A will occur.
  - If P(A) is close to 1, There are high chances that the Event A will occur.
  - If P(A) = 1, Event A will definitely occur.

Introduction to Probability

- How to Calculate the Probability?
  - $P(A) = \frac{Number\ of\ favourable\ Outcomes}{Total\ Number\ of\ outcomes}$
    - Example: Tossing a coin:
    - There are two possible outcomes, Heads or Tails, so the probability of Heads would be:

      P(Heads) = 1/2 = 0.5 and Probability of Tails would be P(Tales) = 1- P(Heads)
    - Example 2: Rolling a dice and getting a number greater than 4:
    - There are 2 possible outcomes {5, 6} and total 6 possible outcomes are {1, 2, 3, 4, 5, 6}
    - So the probability of P(Dice>4) = 2/6 = 0.33

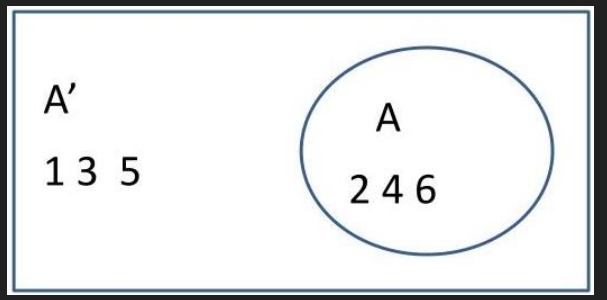
Introduction to Probability

- Sample Space:
  - The set of all possible outcomes of a statistical experiment is called the **Sample Space**. It is represented by the symbol "**S**".
  - Example: Experiment of rolling a Dice
    - Sample Space of rolling a Dice would be $S = \{1, 2, 3, 4, 5, 6\}$

- Sample Point:
  - It is one of the possible results from the Sample space
  - Example: rolling a dice
    - While rolling a dice the result would be one of the outcomes already specified in the Sample Space

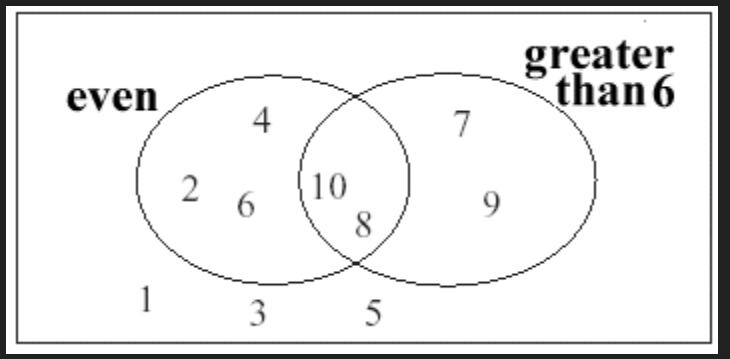
Introduction to Probability

- Events:
  - Event is a subset of Sample Space.
  - Example: Rolling a dice with even number.
    - If we are interested in only even numbers the Event would be $A = \{2, 4, 6\}$ which is a subset of $S = \{1, 2, 3, 4, 5, 6\}$.
  - Complement:
    - The **complement** of an event *A* with respect to any sample space is the subset of all elements of *Sample space* that are not in *A*. It is denoted by symbol *A'*.
    - For any event A there is another event exists which represents the remaining elements of the Sample space S
    - Probability of compliment is $P(A') = 1 - P(A)$ or $P(A) + P(A') = 1$
    - Example: As discussed in the previous example where we were interested in the Even numbers while rolling a dice were A = {2, 4, 6} where S = {1, 2, 3, 4, 5, 6} the complement would be A' = {1, 3, 5}.

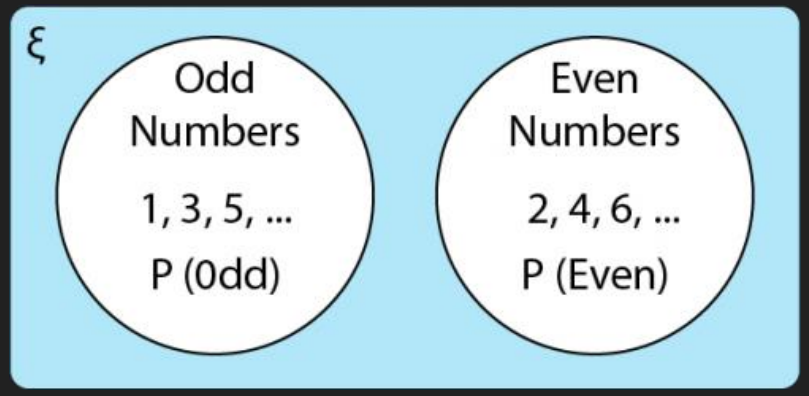


Introduction to Probability

- Events:
  - Intersection:
    - The **intersection** of two events *A* and *B*, denoted by the symbol $A \cap B$, is the event containing all elements that are common to *A* and *B*.
    - The intersection symbol **(∩)** is used to represent "AND" in probability
    - Probability of Intersection is $P(A \cap B) = P(A)*P(B)$
    - *Example: If S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} and A = {2, 4, 6, 8, 10} and B = {7 ,8 ,9 10} then $A \cap B$ = {8, 10}*
  - *Union:*
    - The **union** of the two events *A* and *B*, denoted by the symbol $A \cup B$, is the event containing all the elements that belong to *A* or *B* or both.
    - The Union symbol **(*∪*)** is used to represent "OR" in probability
    - Probability of Intersection is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - *Example: If S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} and A = {2, 4, 6, 8, 10} and B = {7 ,8 ,9 10} then $A \cup B$ = {2, 4, 6, 7, 8, 9, 10}*



Introduction to Probability



- Events:
  - Mutually Exclusive or Disjoint Events:
    - Two events *A* and *B* are **mutually exclusive**, or **disjoint**, if $A \cap B = \{\}$, that is, if *A* and *B* have no elements in common.
    - Two events that are mutually exclusive cannot occur at the same time. Then $P(A \cap B)=0$ and $P(A \cup B)=P(A)+P(B)$
    - *Example: Even and Odd Numbers, Outcomes of tossing a coin*

Introduction to Probability

- Conditional Probability:
  - The probability of an event *B* occurring when it is known that some event *A* has occurred is called a **conditional probability** and is denoted by $P(B \mid A)$. The symbol $P(B \mid A)$ is usually read as "the probability of *B*, given *A*."
  - $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$, Where $P(A) > 0$
  - Example: In a class, Probability of selecting a student for Internship who is Undergraduate given that the student is unemployed.

    It will be denoted as P(Undergrad | unemployed) = P(Undergrad ∩ unemployed)/ P(unemployed)

    Or Selecting a Student who is unemployed given that he is undergrad too will be denoted as:

    P(unemployed | Undergrad) = P(Undergrad ∩ unemployed)/ P(Undergrad)
  - Example 2: In a lot of T-Shirts, Probability of T-shirt whose print is defective given that the size of the T-shirt is small

Introduction to Probability

- Independent Events:
  - Two events are independent if the occurrence of one does not affect the probability of occurrence of the other.
  - Two events *A* and *B* are **independent** if and only if "$P(B|A) = P(B)$" or "$P(A|B) = P(A)$", *Otherwise the Events are* ***dependent***.
  - Example: Student of Data science and Owning a Pet
  - Example2: Owning a laptop and reaching home on time
  - Example3: Rolling a Dice twice and getting 2 on first roll and getting a 6 on second roll.

Introduction to Statistics

- Measure of Central Tendency:
  - A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.
  - The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.
- Mean:
  - The mean (or average) is the most popular and well known measure of central tendency.
  - The mean is equal to the sum of all the values in the data set divided by the number of values in the data set.
  - Sample Mean is Denoted by x-bar and Population Mean is denoted by Mu
  - Example: Average Age of Students in a class.
  - Example2: Average Runs in an over in cricket match.
  - Example3: Average Marks of Students in a test.

Sample Mean	Population Mean
$\bar{x} = \frac{\Sigma x}{n}$	$\mu = \frac{\Sigma x}{N}$

where Σx is sum of all data values

N is number of data items in population

n is number of data items in sample

Introduction to Statistics

- Measure of Central Tendency:
  - Median:
    - One important measure is the **sample median**. The purpose of the sample median is to reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliers.
    - The median is the middle score for a set of data that has been arranged in ascending order.

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{if } n \text{ is even.} \end{cases}$$

    - Example: In a class of matric students mostly students are aged around 15,16 years but there are two students aged around 20 years who are outliers because of them Mean would not be able to depict the actual center point for which we use median to find the center point of the ages of matric students.

Introduction to Statistics

- Measure of Central Tendency:
  - Mode:
    - The mode is the most frequent tally in our data set.
    - In simple words most repeated data point in our data set.
    - Example1: Number of Shirts sold in different collar sizes, Helps in restocking with the most sold collar size
    - Example2: Number of Air conditioners sold in summer season, Helps in restocking the air conditioner.
    - Example3: Number of customers from different areas, Helps marketing teams to target the areas where customer number is low.