DA: Unit-2

Part-1::Introduction to Probability:	Part-2::Sampling and Sampling Distribution:
1. Classical Probability, 2. Relative Frequency, 3. Sample Space, 4. Events, 5. Types of Probability, 6. Conditional Probability, 7. Bayesian Rule, 8. Relative frequency method, 9. Random Variable, 10. Distribution Function, 11. Density Function	12.Random vs Non Random Sampling 13.Simple random sampling 14.cluster sampling 15.concept of sampling distributions, Student's t-test 16. Chi-square and F- distributions 17.Central limit theorem and its application, confidence intervals

Introduction to Probability

Probability is a fundamental concept in mathematics that measures the likelihood of an event occurring

Why Probability Matters in Data Analytics

- Uncertainty Management**:**
- Inference and Prediction**:**
- Decision-Making**:**

Tools and Techniques

- **Python Libraries**: Libraries like NumPy, SciPy, and pandas provide robust tools for probability computations.
- **Visualization**: Use tools like Matplotlib or Seaborn to visualize probability distributions and relationships.
- **Monte Carlo Simulations**: Technique to estimate probabilities and outcomes through repeated random sampling.

Applications of Probability in Data Analytics

- **Predictive Modeling**: Estimating future trends based on historical data.
- **A/B Testing**: Comparing two versions of a product or campaign to determine which performs better.
- **Risk Analysis**: Quantifying and managing potential risks in business or operations.
- **Natural Language Processing (NLP)**: Probability underpins models for tasks like text classification and sentiment analysis.
- **Machine Learning**: Many algorithms, such as Naive Bayes and Hidden Markov Models, are rooted in probability.

Basic Terms in Probability

1. **Experiment**
    - A procedure or process that produces a definite outcome.
    - Example: Rolling a die or flipping a coin.
2. **Sample Space (S)**
    - The set of all possible outcomes of an experiment.
    - Example: For a coin toss, S={Heads, Tails}
3. **Event**
    - A subset of the sample space, representing outcomes of interest.
    - Example: Getting an even number when rolling a die ({2, 4, 6}).
4. **Outcome**
    - A single possible result of an experiment.
    - Example: Rolling a "4" in a die toss.
5. **Probability (P)**
    - A measure of how likely an event is to occur.

Formula

$$P(\text{Event}) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

1. Classical Probability

Classical probability is based on the assumption that all outcomes in a sample space are equally likely. The probability of an event is calculated as the ratio of the number of favorable outcomes to the total number of possible outcomes.

It is calculated using the formula:

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

Example-1: What is the probability of rolling a 3 on a standard six-sided die?

- **Favorable outcomes**: Rolling a 3 → 1 outcome.
- **Total outcomes**: Numbers on the die (1, 2, 3, 4, 5, 6) → 6 outcomes.

P(rolling a 3)=1/6

Example-2: if you toss a coin 100 times and it comes up heads 60 times, the relative frequency of getting heads is 60/100 or 0.6.

2. Relative Frequency

Definition: Relative frequency is the ratio of the number of times an event occurs to the total number of trials or observations. It represents empirical probability.

$$P(E) = \frac{\text{Frequency of the event}}{\text{Total number of trials}}$$

Example:

Problem: A coin is tossed 100 times, and it lands on heads 48 times. What is the relative frequency probability of getting heads?

Solution:

- **Frequency of heads**: 48
- **Total trials**: 100

$$P(\text{Heads}) = \frac{48}{100} = 0.48$$

3. Sample Space

The sample space is the set of all possible outcomes of a random experiment.

Problem: Find the sample space for tossing two coins.

Solution:

- Each coin can show **Heads (H)** or **Tails (T)**.
- Sample space SSS: {HH, HT, TH, TT}

4. Events

Definition: An event is any subset of the sample space. Events can be simple (single outcome) or compound (multiple outcomes).

Example:

Problem: In the sample space S= {HH,HT,TH,TT}, define the event A: "at least one tail".

Solution:

- Outcomes with at least one tail: {HT, TH, TT}
- A={HT,TH,TT}

5. Types of Probability

- **Classical Probability**
- **Empirical (Experimental) Probability**
- **Subjective Probability**

Types of Probability. Probability can be categorized into various types, each with its own approach to defining and calculating the likelihood of events. The three main types of probability are Classical Probability, Empirical (or Relative Frequency) Probability, and Subjective Probability

1. Classical Probability

Based on the assumption that all outcomes in a sample space are equally likely.

It is most applicable to situations where each outcome is equally likely, such as flipping a fair coin or rolling a fair die.

The classical probability of an event A is calculated as the ratio of the number of favorable outcomes to the total number of possible outcomes.

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Example: Tossing a fair coin.

- Sample space: {Head, Tail}
- Probability of getting a head: P(Head)=1/2

2. Empirical (Experimental) Probability

Definition: Based on experiments or historical data. Probability is calculated as the ratio of the number of times an event occurs to the total number of trials.

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

Example: Rolling a die 100 times and observing a 6 appears 20 times.

- Probability of rolling a 6:

$$P(6) = \frac{20}{100} = 0.2$$

3. Subjective Probability

- Based on personal judgment, experience, intuition, or opinion rather than objective data.
- Subjective probabilities are often used in decision-making and situations where objective data is unusual.
- There is no specific formula for subjective probability; individuals assign probabilities based on their perception, experience, or other subjective factors.

Example: Predicting a 70% chance of rain tomorrow based on weather patterns

6. Conditional Probability

Conditional probability is the probability of an event occurring given that another event has already occurred. It is written as:

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

Where:

- **P(A|B)** is the conditional probability of event A occurring given that event B has occurred.
- **P(A∩B)** is the probability that both events A and B occur.
- **P(B)** is the probability that event B occurs.

Key Concepts:

1. **Joint Probability**: The probability that both events A and B occur, denoted as P(A∩B).
2. **Marginal Probability**: The probability of a single event occurring, denoted as P(A) or P(B).
3. **Independent Events**: If two events A and B are independent, then P(A∩B)=P(A)×P(B), and in this case, conditional probability does not change, i.e., P(A|B)=P(A).

Example 1: A Simple Coin Toss

Suppose we toss two fair coins. Let:
Event A: "The first coin is heads."
Event B: "At least one coin is heads."

We want to find the conditional probability of A given B, i.e., P(A|B).

Step 1: Find P(A∩B)

For both events A and B to happen (first coin is heads, and at least one coin is heads), the outcomes must be:

- **HH (Head, Head)**

Thus, **P(A∩B)=P(HH)=1/4.**

Step 2: Find P(B)

Event B happens if at least one coin shows heads. The possible outcomes are:

- HH, HT, TH

Thus, **P(B)=3/4**.

Step 3: Calculate P(A|B)P(A|B)P(A|B)

Now we can apply the formula:
P(A|B)=P(A∩B)/P(B)=(1/4)/(3/4)=1/3
So, the conditional probability of event A given event B is 1/3.

Example-2

Problem: In a group of students, 40% study math, 30% study physics, and 20% study both. What is the probability that a student studies math given that they study physics?

Solution:

- $P(\text{Math}) = 0.4$, $P(\text{Physics}) = 0.3$, $P(\text{Math and Physics}) = 0.2$

$$P(\text{Math} \mid \text{Physics}) = \frac{P(\text{Math and Physics})}{P(\text{Physics})} = \frac{0.2}{0.3} = \frac{2}{3}$$

7. Bayesian Rule:

- Bayes' Theorem (or Bayes' Rule) is a fundamental concept in probability theory and statistics.
- It describes the probability of an event, based on prior knowledge of conditions that might be related to the event.
- Bayes' Theorem is particularly useful for updating the probability estimate for an event as more evidence becomes available.

Bayes' Theorem Formula:

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

Where:

- **P(A|B)** is the **posterior probability**: the probability of event A occurring given that B has occurred.
- **P(B|A)** is the **likelihood**: the probability of event B occurring given that A has occurred.
- **P(A)** is the **prior probability**: the initial probability of event A occurring, before considering B.
- **P(B)** is the **marginal likelihood**: the total probability of event B occurring (which is the sum of the probabilities of all ways B can happen).

Steps to Apply Bayes' Theorem:

1. **Identify the events**: Define the events you are dealing with.
2. **Determine the known probabilities**: Identify the prior probability P(A), the likelihood P(B|A), and the marginal probability P(B).
3. **Apply the formula**: Substitute the values into Bayes' Theorem to calculate the posterior probability P(A|B)

Explanation

Bayes' Theorem allows you to update your belief about a hypothesis A when new evidence B is introduced. It essentially combines:

1. **Prior Knowledge**: The initial belief or probability of A.
2. **New Evidence**: How likely B is if A is true.
3. **Normalization**: Ensures probabilities sum to 1 by dividing by the overall probability of B

Example: Medical Diagnosis

Suppose a patient is tested for a disease. Let:

- A: The patient has the disease.
- B: The test result is positive.

Given:

- The prevalence of the disease ($P(A)$) = 1% or 0.01.
- The test's sensitivity ($P(B|A)$) = 99% or 0.99 (true positive rate).
- The test's false positive rate ($P(B|\neg A)$) = 5% or 0.05.

Step 1: Calculate $P(B)$

$$P(B) = P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)$$

$$P(B) = (0.99 \cdot 0.01) + (0.05 \cdot 0.99) = 0.0099 + 0.0495 = 0.0594$$

Step 2: Use Bayes' Theorem to find $P(A|B)$

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

$$P(A|B) = \frac{0.99 \cdot 0.01}{0.0594} = 0.1666 \, (\approx 16.7\%)$$

Interpretation

Even with a positive test result, the probability of actually having the disease is only about 16.7%. This result highlights the importance of considering prior probabilities and test accuracy when interpreting diagnostic results.

Applications

- **Medical Diagnosis**: Evaluating the likelihood of a disease given symptoms or test results.
- **Spam Filters**: Determining whether an email is spam based on its content.
- **Machine Learning**: Bayesian methods are used in classification and decision-making models.
- **Risk Assessment**: Updating risks based on new information

8. Relative Frequency Method

The **Relative Frequency Method** is a statistical approach used to estimate the probability of an event by dividing the number of times the event occurs by the total number of observations. It is particularly useful when probabilities are not known in advance and must be inferred from experimental or historical data.

Formula:

$$P(E) = \frac{\text{Number of times event } E \text{ occurs}}{\text{Total number of observations}}$$

Where:

- P(E) = Probability of event E.
- Number of times event E occurs is the frequency of the event.
- Total number of observations is the total number of trials or outcomes considered.

Steps to Apply the Relative Frequency Method:

1. **Collect Data**: Gather historical or experimental data about the events of interest.
2. **Count Occurrences**: Determine how many times the event of interest occurred. Count the number of occurrences of each category in the dataset. This creates a frequency distribution, showing how often each category appears.
3. **Calculate Relative Frequency**: Divide the count of the event's occurrences by the total number of observations.

   Calculate the relative frequency for each category by dividing the frequency of that category by the total number of observations. Mathematically, it can be expressed as:

   $$\text{Relative Frequency of Category} = \frac{\text{Frequency of Category}}{\text{Total Number of Observations}}$$

4. **Interpret the Result**: The resulting value represents the estimated probability.
5. **Visualize the Distribution:** Represent the relative frequencies graphically using charts or graphs such as bar charts, pie charts, or histograms.

   Visualization aids in better understanding the patterns and trends in the data.

Example:

Let's consider a simple example where you have collected data on the favorite colors of 100 people

Blue: 30 people
Red: 20 people
Green: 15 people
Yellow: 10 people
Other: 25 people

Calculations:

$$\text{Relative Frequency of Blue} = \frac{30}{100} = 0.30$$
$$\text{Relative Frequency of Red} = \frac{20}{100} = 0.20$$
$$\text{Relative Frequency of Green} = \frac{15}{100} = 0.15$$
$$\text{Relative Frequency of Yellow} = \frac{10}{100} = 0.10$$
$$\text{Relative Frequency of Other} = \frac{25}{100} = 0.25$$

8. Relative Frequency Method

- The **Relative Frequency Method** is a statistical approach used to estimate the probability of an event by dividing the number of times the event occurs by the total number of observations.
- It is particularly useful when probabilities are not known in advance and must be inferred from experimental or historical data.

Formula:

$$P(E) = \frac{\text{Number of times event } E \text{ occurs}}{\text{Total number of observations}}$$

Where:

- P(E) = Probability of event E.
- Number of times event E occurs is the frequency of the event.
- Total number of observations is the total number of trials or outcomes considered.

Steps to Apply the Relative Frequency Method:

- **Collect Data**: Gather historical or experimental data about the events of interest.
- **Count Occurrences**: Determine how many times the event of interest occurred. Count the number of occurrences of each category in the dataset. This creates a frequency distribution, showing how often each category appears.
- **Calculate Relative Frequency**: Divide the count of the event's occurrences by the total number of observations.
    - Calculate the relative frequency for each category by dividing the frequency of that category by the total number of observations. Mathematically, it can be expressed as:

$$\text{Relative Frequency of Category} = \frac{\text{Frequency of Category}}{\text{Total Number of Observations}}$$

    -
- **Interpret the Result**: The resulting value represents the estimated probability.
- **Visualize the Distribution:** Represent the relative frequencies graphically using charts or graphs such as bar charts, pie charts, or histograms.
    - Visualization aids in better understanding the patterns and trends in the data.

Example:

Let's consider a simple example where you have collected data on the favorite colors of 100 people

Blue: 30 people
Red: 20 people
Green: 15 people
Yellow: 10 people
Other: 25 people

Calculations:

$$\text{Relative Frequency of Blue} = \frac{30}{100} = 0.30$$
$$\text{Relative Frequency of Red} = \frac{20}{100} = 0.20$$
$$\text{Relative Frequency of Green} = \frac{15}{100} = 0.15$$
$$\text{Relative Frequency of Yellow} = \frac{10}{100} = 0.10$$
$$\text{Relative Frequency of Other} = \frac{25}{100} = 0.25$$

Example 1: Rolling a Die

Suppose you roll a six-sided die 50 times and record the outcomes:

- Outcome 1: Occurred 10 times.
- Outcome 2: Occurred 8 times.
- Outcome 3: Occurred 7 times.
- Outcome 4: Occurred 9 times.
- Outcome 5: Occurred 6 times.
- Outcome 6: Occurred 10 times.

To estimate the probability of rolling a 1:

$$P(\text{Rolling a } 1) = \frac{\text{Frequency of } 1}{\text{Total rolls}} = \frac{10}{50} = 0.2$$

Thus, the probability of rolling a 1 is 0.2 or 20%.

Advantages:

- Simple and easy to compute with experimental or historical data.
- Flexible and applicable to various scenarios.

Limitations:

- **Requires sufficient data** for accurate estimation.
- Results depend heavily on the quality and quantity of the collected data.
- May not be reliable for predicting rare events if observations are limited.

9. Random Variable

Random Variable: Definition

A **random variable** is a numerical outcome of a random phenomenon or experiment. It is a function that assigns a real number to each possible outcome of a random process. Random variables are used in probability and statistics to quantify uncertainty and randomness.

Random variables can be classified into two types:

1. **Discrete Random Variable**: Takes on a countable number of distinct values (e.g., integers).
2. **Continuous Random Variable**: Takes on an infinite number of possible values within a given range (e.g., real numbers).

Example: Discrete Random Variable

Experiment:

Roll a six-sided die.

Random Variable X:

Let X represent the outcome of the roll.

- Possible values of X: $\{1, 2, 3, 4, 5, 6\}$
- $P(X = 1) = \frac{1}{6}$, $P(X = 2) = \frac{1}{6}, \ldots, P(X = 6) = \frac{1}{6}$

In this case, X is a **discrete random vari** because it has a finite number of outcomes.

Example: Continuous Random Variable

Experiment:

Measure the time it takes for a car to complete a lap on a track.

Random Variable Y:

Let Y represent the time (in seconds).

- Possible values of Y: Any positive real number (e.g., $57.3, 57.31, 57.312$, etc.)
- Probability is described using a probability density function (PDF), such as $f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ for a normal distribution.

In this case, Y is a **continuous random variable** because it can take any value within a range.

Key Characteristics

1. **Probability Distribution**: Describes the probabilities associated with each possible value of the random variable.
   - Discrete: Probability Mass Function (PMF).
   - Continuous: Probability Density Function (PDF).
2. **Expected Value (Mean)**: The average or central value of the random variable.

$$E(X) = \sum_i x_i P(X = x_i) \quad \text{(for discrete variables)}$$

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) \quad \text{(for continuous variables)}$$

10. Distribution Function

Cumulative Distribution Function (CDF): Definition

The **Cumulative Distribution Function (CDF)** of a random variable X gives the probability that X takes a value less than or equal to a certain number x. Formally, the CDF is defined as:

$$F_X(x) = P(X \leq x)$$

Where:

- $F_X(x)$ is the value of the CDF at x,
- X is the random variable.

The CDF applies to both **discrete** and **continuous** random variables, though it is computed differently in each case.

Properties of a CDF

1. $0 \leq F_X(x) \leq 1$: The CDF value is always between 0 and 1.
2. Non-decreasing: $F_X(x_1) \leq F_X(x_2)$ for $x_1 < x_2$.
3. Limits:
   - $\lim_{x \to -\infty} F_X(x) = 0$
   - $\lim_{x \to \infty} F_X(x) = 1$
4. For discrete random variables, the CDF is a step function.
5. For continuous random variables, the CDF is a smooth and continuous curve.

Example: Discrete Random Variable

Random Experiment:

Roll a six-sided die.

Random Variable X:

Let X be the outcome of the die roll ($X \in \{1, 2, 3, 4, 5, 6\}$).

CDF Calculation:

- $F_X(x) = P(X \leq x)$
- For specific values of x:
  - $F_X(1) = P(X \leq 1) = \frac{1}{6}$
  - $F_X(2) = P(X \leq 2) = P(X = 1) + P(X = 2) = \frac{2}{6}$
  - $F_X(3) = P(X \leq 3) = \frac{3}{6}$
  - $F_X(6) = P(X \leq 6) = 1$

The CDF for X is a step function that increases at each possible value of X.

Example: Continuous Random Variable

Random Experiment:

Measure the height (in cm) of randomly selected individuals in a population.

Random Variable X:

Let X represent the height. Assume X follows a normal distribution with a mean $\mu = 170$ and standard deviation $\sigma = 10$.

CDF Calculation:

The CDF is given by:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt$$

Where $f_X(x)$ is the probability density function (PDF).

For example:

- $F_X(160)$ represents the probability that a randomly selected individual has a height ≤ 160.
- $F_X(180)$ represents the probability that the height is ≤ 180.

Values of the CDF can be computed using statistical tables or software.

Graphical Representation

1. **Discrete CDF**: A step-like graph where the CDF increases at each possible value of X.
2. **Continuous CDF**: A smooth, non-decreasing curve that starts at 0 and approaches 1 as $x \to \infty$.

Uses of the CDF

- Determine probabilities: $P(a \leq X \leq b) = F_X(b) - F_X(a)$.
- Analyze random variable behavior across ranges.
- Compute percentiles and quantiles.

11. Density Function:

Probability Density Function (PDF): Definition

The **Probability Density Function (PDF)** describes the likelihood of a continuous random variable taking a specific value. While the value of the PDF itself does not represent a probability, the area under the curve of the PDF over a given interval represents the probability that the random variable falls within that interval.

Formal Definition

The PDF, $f_X(x)$, of a continuous random variable X is defined such that:

1. $f_X(x) \geq 0$ for all x,
2. The total area under the curve is 1:
$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1,$$
3. The probability that X falls within an interval $[a, b]$ is:
$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx.$$

Key Difference from Discrete Case:

For discrete random variables, we use the **Probability Mass Function (PMF)**, where probabilities are assigned to specific valu For continuous variables, probabilities are represented as areas under the PDF curve.

Example of a PDF

Random Experiment:

Measure the heights (in cm) of a group of people, which follows a normal distribution with a mean $\mu = 170$ cm and standard deviation $\sigma = 10$ cm.

Random Variable X:

Let X represent the height.

The PDF for a normal distribution is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For this example:

$$f_X(x) = \frac{1}{\sqrt{2\pi(10)^2}} e^{-\frac{(x-170)^2}{2(10)^2}} = \frac{1}{\sqrt{200\pi}} e^{-\frac{(x-170)^2}{200}}$$

Using the PDF:

- To find the likelihood that a person's height is between 160 cm and 180 cm:
$$P(160 \leq X \leq 180) = \int_{160}^{180} f_X(x)\, dx.$$
This requires calculating the area under the curve between $x = 160$ and $x = 180$.

Graphical Representation

- The PDF curve for a normal distribution is bell-shaped and symmetric about the mean (μ).
- The height of the curve represents the density of the random variable near a specific value.
- The area under the curve between two points gives the probability of the variable lying in that range.

Key Properties of a PDF

1. **Non-negativity**: $f_X(x) \geq 0$ for all x.
2. **Normalization**: The total area under the PDF curve is 1.
3. **No direct probability for a specific value**: For continuous variables, $P(X = x) = 0$