Assignment

Q1: What are the Probability Mass Function (PMF) and Probability Density Function (PDF)? Explain with
an example.

Ans:The Probability Mass Function (PMF) and Probability Density Function (PDF) are mathematical functions that describe the likelihood of a discrete or continuous random variable taking on specific values, respectively.

Probability Mass Function (PMF):

Definition:

- The PMF is a function that gives the probability of a discrete random variable taking on a specific value.
- It is often denoted as
- $\square(\square=\square)$
- $P(X=x)$, where
- $\square$
- X is the random variable and
- $\square$
- x is a particular value.
- The PMF satisfies two conditions: non-negativity and the sum of probabilities over all possible values is equal to 1.

Example:

- Consider a fair six-sided die. The PMF for this die is uniform, meaning that each outcome has an equal probability of
- 16
  - 6
  - 1
  -

. The PMF is expressed as:

-
- $\square(\square=\square)=16$
- $P(X=x)=$
  - 6
  - 1
  -

- for each

- $\square$
- x from 1 to 6.

Probability Density Function (PDF):

Definition:

- The PDF is a function that gives the probability density (likelihood per unit interval) of a continuous random variable taking on a specific value.
- It is often denoted as
- $\square(\square)$
- $f(x)$, where
- $\square$
- x is a particular value.
- The area under the PDF curve over a specific range represents the probability of the random variable falling within that range.

Example:

The standard normal distribution (bell curve) has a PDF given by:

-
- $\square(\square)=12\square\square-\square22$
- $f(x)=$
- 2π
-
  - 1
  -
- e
- $-$
  - 2
- x
- 2
  -

- This function describes the probability density for different values of
- $\square$
- x in a continuous distribution. However, the probability of obtaining a specific
- $\square$
- x is zero, and probabilities are calculated for intervals.

In summary, the PMF is used for discrete random variables and provides the probability of specific outcomes, while the PDF is used for continuous random variables and gives the probability density for different values. Both functions are fundamental in probability theory and statistical modeling.

Q2: What is Cumulative Density Function (CDF)? Explain with an example. Why CDF is used?
Ans:The Cumulative Distribution Function (CDF) is a function associated with a probability distribution. It describes the probability that a random variable takes on a value less than or equal to a given point. In other words, the CDF provides the cumulative probability up to a certain value.

Definition:

For a random variable

$\diamond$

X, the CDF, denoted as

$\diamond(\diamond)$

$F(x)$, is defined as:

$\diamond(\diamond)=\diamond(\diamond\leq\diamond)$

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

Example:

Let's consider a fair six-sided die. The CDF for this die can be calculated as follows:

- If
- $\diamond\leq 1$
- $x\leq 1$:
- $\diamond(\diamond)=\diamond(\diamond\leq 1)=16$
- $F(x)=P(X\leq 1)=$

- 6
- 1
-
-
- If
- $1<\text{□}\leq 2$
- $1<x\leq 2$:
- $\text{□}(\text{□})=\text{□}(\text{□}\leq 2)=26=13$
- $F(x)=P(X\leq 2)=$
  - 6
  - 2
  -
- =
  - 3
  - 1
  -
-
- If
- $2<\text{□}\leq 3$
- $2<x\leq 3$:
- $\text{□}(\text{□})=\text{□}(\text{□}\leq 3)=36=12$
- $F(x)=P(X\leq 3)=$
  - 6
  - 3
  -
- =
  - 2
  - 1
  -
-
- If
- $3<\text{□}\leq 4$
- $3<x\leq 4$:
- $\text{□}(\text{□})=\text{□}(\text{□}\leq 4)=46=23$
- $F(x)=P(X\leq 4)=$
  - 6
  - 4
  -
- =

- 3
- 2
-
-
- If
- $4<\text{□}\leq 5$
- $4<x\leq 5$:
- $\text{□}(\text{□})=\text{□}(\text{□}\leq 5)=56$
- $F(x)=P(X\leq 5)=$
  - 6
  - 5
  -
-
- If
- $5<\text{□}\leq 6$
- $5<x\leq 6$:
- $\text{□}(\text{□})=\text{□}(\text{□}\leq 6)=1$
- $F(x)=P(X\leq 6)=1$

Why CDF is Used?

Cumulative Information: The CDF provides cumulative information about the probability distribution up to a specific point. It gives the probability of the random variable being less than or equal to a given value.
Easier Probability Calculations: It simplifies the calculation of probabilities for intervals. The probability of the random variable falling within an interval
$[\text{□},\text{□}]$
$[a,b]$ is given by
$\text{□}(\text{□})-\text{□}(\text{□})$
$F(b)-F(a)$.
Quantifying Percentiles: The CDF allows us to determine percentiles. For example,
$\text{□}(0.5)$
$F(0.5)$ gives the median of the distribution.
Graphical Representation: The CDF can be graphically represented as a step function, providing a visual understanding of the cumulative probabilities.

In summary, the Cumulative Distribution Function is a crucial tool in probability theory and statistics, providing a comprehensive way to understand and work with the probabilities associated with a random variable.

Q3: What are some examples of situations where the normal distribution might be used as a model?
Explain how the parameters of the normal distribution relate to the shape of the distribution.
Ans:The normal distribution, also known as the Gaussian distribution or bell curve, is a versatile probability distribution commonly used to model a variety of natural phenomena. Some examples of situations where the normal distribution might be used as a model include:

Height of Individuals:
- The heights of a population tend to follow a normal distribution. The mean represents the average height, and the standard deviation indicates how much individuals' heights deviate from the mean.

IQ Scores:
- IQ scores are often modeled using a normal distribution. The mean IQ is set at 100, and the standard deviation indicates the variability in IQ scores.

Measurement Errors:
- Errors in measurement instruments, such as a ruler or scale, often follow a normal distribution. The mean represents the true value, and the standard deviation indicates the precision of the measurement instrument.

Blood Pressure:
- Blood pressure in a population often follows a normal distribution. The mean represents the average blood pressure, and the standard deviation indicates the variability around the mean.

Test Scores:
- In educational testing, the scores on standardized tests often approximate a normal distribution. The mean represents the average score, and the standard deviation indicates the spread of scores.

Natural Phenomena:
- Many natural phenomena, such as the distribution of particle velocities in a gas or the distribution of errors in a scientific experiment, can be modeled by a normal distribution.

Parameters of the Normal Distribution:

The normal distribution is characterized by two parameters:

Mean (


μ):

- The mean determines the location of the center of the distribution.
- The curve is symmetric, and the peak of the curve is at the mean.
- Shifting the mean to the right or left moves the entire distribution along the x-axis.

Standard Deviation (

σ):

- The standard deviation determines the spread or variability of the distribution.
- A larger standard deviation results in a wider, flatter curve, indicating greater variability.
- A smaller standard deviation results in a narrower, taller curve, indicating less variability.

In summary, the normal distribution is a versatile model that is often used in situations where data tends to cluster around a central value with a known level of variability. The mean and standard deviation are key parameters that determine the shape and characteristics of the distribution.

Q4: Explain the importance of Normal Distribution. Give a few real-life examples of Normal Distribution.

Ans:The normal distribution is of paramount importance in statistics and probability theory due to several key properties and its ubiquity in describing natural phenomena. Some reasons for the importance of the normal distribution include:

Central Limit Theorem (CLT):

- The normal distribution plays a central role in the Central Limit Theorem, which states that the sum (or average) of a large number of independent, identically distributed random variables, regardless of their original distribution, tends to follow a normal distribution.
- This makes the normal distribution relevant in many statistical analyses, as it provides a basis for making inferences about population parameters.

Statistical Inference:

- Many statistical methods and tests assume or work best when the underlying data follows a normal distribution. This includes hypothesis testing, confidence intervals, and regression analysis.

Parameter Estimation:

- In maximum likelihood estimation and other parameter estimation techniques, the normal distribution often serves as a convenient and mathematically tractable assumption.

Predictive Modeling:

- In machine learning and predictive modeling, the normal distribution is often assumed, and models like linear regression assume that the errors are normally distributed.

Quality Control:

- In manufacturing and quality control, the normal distribution is used to model variations in product specifications and to set tolerances. This is essential for ensuring product quality.

Biological and Social Phenomena:

- Many biological and social phenomena, such as height, weight, IQ scores, and test scores, tend to follow a normal distribution. This makes it a natural choice for modeling and analyzing these types of data.

Real-Life Examples of Normal Distribution:

Height of Individuals:

- The distribution of human heights tends to follow a normal distribution. The mean represents the average height, and most people cluster around this value, with fewer individuals at the extremes.

IQ Scores:

- IQ scores are designed to follow a normal distribution, with a mean of 100 and a standard deviation of 15. This allows for easy interpretation and comparison of an individual's cognitive abilities.

Temperature:

- Daily temperatures in a location over a long period often approximate a normal distribution. The mean temperature represents the typical climate, and variations from the mean follow a bell-shaped curve.

Blood Pressure:

- Blood pressure in a population tends to be normally distributed. The mean represents the average blood pressure, and most individuals fall within one standard deviation of the mean.

Exam Scores:

- Scores on standardized exams, such as SAT or GRE, are often designed to follow a normal distribution. The mean represents the average performance, and the spread of scores is characterized by the standard deviation.

Understanding and utilizing the normal distribution is crucial in various fields, enabling researchers, analysts, and decision-makers to make informed predictions, draw meaningful conclusions, and perform statistical analyses in a wide range of applications.

Q5: What is Bernaulli Distribution? Give an Example. What is the difference between Bernoulli Distribution and Binomial Distribution?

Ans:Bernoulli Distribution:

The Bernoulli distribution is a discrete probability distribution that models a random experiment with two possible outcomes, often labeled as "success" and "failure." It is named after the Swiss mathematician Jacob Bernoulli. The probability mass function (PMF) of the Bernoulli distribution is defined as:

$$\text{[illegible]}(\text{[illegible]}=\text{[illegible]})=\text{[illegible]}\text{[illegible]}\cdot(1-\text{[illegible]})1-\text{[illegible]}$$

$$P(X=k)=p$$

k

$$\cdot(1-p)$$

$1-k$

where:

- [illegible]
- X is the random variable representing the outcome (1 for success, 0 for failure),
- [illegible]
- k is the value that
- [illegible]
- X can take (either 0 or 1),
- [illegible]
- p is the probability of success.

Example of Bernoulli Distribution:

- Consider a single coin flip. Let
- [illegible]
- X be the random variable representing the outcome, where
- [illegible]$=1$
- $X=1$ if the coin lands heads (success) and
- [illegible]$=0$

- $X=0$ if it lands tails (failure).

The probability mass function is given by:

-
- $\text{\textdiamond}(\text{\textdiamond}=\text{\textdiamond})=\text{\textdiamond}\text{\textdiamond}\cdot(1-\text{\textdiamond})1-\text{\textdiamond}$
- $P(X=k)=p$
- k
- $\cdot(1-p)$
- $1-k$

- where
- $\text{\textdiamond}$
- p is the probability of getting heads (success).

Difference between Bernoulli Distribution and Binomial Distribution:

Number of Trials:
- Bernoulli Distribution: Represents a single trial or experiment with two possible outcomes.
- Binomial Distribution: Represents the number of successes in a fixed number of independent and identical trials.

Random Variable:
- Bernoulli Distribution: Has a single binary random variable (
- $\text{\textdiamond}$
- X) indicating success or failure.
- Binomial Distribution: Involves the sum of
- $\text{\textdiamond}$
- n independent Bernoulli trials, resulting in a binomial random variable representing the number of successes.

Probability Mass Function (PMF):
- Bernoulli Distribution:
- $\text{\textdiamond}(\text{\textdiamond}=\text{\textdiamond})=\text{\textdiamond}\text{\textdiamond}\cdot(1-\text{\textdiamond})1-\text{\textdiamond}$
- $P(X=k)=p$
- k
- $\cdot(1-p)$

- $1-k$
- , where
-
- k is 0 or 1.
- Binomial Distribution:
-
- $P(X=k)=($
  - k
  - n
  -
- $)p$
- k
- $(1-p)$
- $n-k$
- , where
-
- k can range from 0 to
-
- n.

Parameters:

- Bernoulli Distribution: Has a single parameter
-
- p representing the probability of success.
- Binomial Distribution: Has two parameters,
-
- n (number of trials) and
-
- p (probability of success in each trial).

Mean and Variance:

- Bernoulli Distribution: Mean (
-
- μ) is
-
- p, and variance (
- $\sigma 2$
- σ
- 2

-) is
- $\hat{\theta}(1-\hat{\theta})$
- $p(1-p)$.
- Binomial Distribution: Mean (
- $\hat{\theta}$
- μ) is
- $\hat{\theta}\hat{\theta}$
- np, and variance (
- $\hat{\theta}2$
- σ
- 2
-) is
- $\hat{\theta}\hat{\theta}(1-\hat{\theta})$
- $np(1-p)$.

In summary, the Bernoulli distribution is a special case of the binomial distribution where there is only one trial (

$\hat{\theta}=1$

$n=1$). The binomial distribution generalizes the concept to multiple independent trials, allowing

for the modeling of the number of successes in a sequence of experiments.

Q6. Consider a dataset with a mean of 50 and a standard deviation of 10. If we assume that the dataset
is normally distributed, what is the probability that a randomly selected observation will be greater
than 60? Use the appropriate formula and show your calculations.
ans:To find the probability that a randomly selected observation from a normally distributed dataset will be greater than 60, we can use the Z-score formula and standard normal distribution tables.

The Z-score is calculated as follows:

$\hat{\theta}=\hat{\theta}-\hat{\theta}\hat{\theta}$

$Z=$

σ

$X-\mu$

where:

- $\square$
- X is the value we're interested in (60 in this case),
- $\square$
- μ is the mean of the distribution (50),
- $\square$
- σ is the standard deviation of the distribution (10).

Calculating the Z-score for

$\square=60$

$X=60$:

$\square=60-5010=1$

$Z=$

10

$60-50$

$=1$

Now, we look up the probability associated with a Z-score of 1 in the standard normal distribution table. The standard normal distribution table typically provides the probability that a random variable from a standard normal distribution is less than the given Z-score. However, since we want the probability that the observation is greater than 60, we need to find

$\square(\square>1)$

$P(Z>1)$.

The standard normal distribution is symmetric, so

$$\text{\textdiamond}(\text{\textdiamond}>1)$$

$P(Z>1)$ is the complement of

$$\text{\textdiamond}(\text{\textdiamond}<1)$$

$P(Z<1)$.

From the standard normal distribution table,

$$\text{\textdiamond}(\text{\textdiamond}<1)$$

$P(Z<1)$ is approximately 0.8413.

Therefore,

$$\text{\textdiamond}(\text{\textdiamond}>1)$$

$P(Z>1)$ is approximately

$$1-0.8413=0.1587$$

$1-0.8413=0.1587$.

So, the probability that a randomly selected observation from the dataset will be greater than 60 is approximately 0.1587 or 15.87%.

Q7: Explain uniform Distribution with an example.
Ans:The uniform distribution is a probability distribution where all possible outcomes are equally likely. In other words, each value within a given range has an equal probability of occurring. The probability density function (PDF) of a uniform distribution is constant over the entire range, resulting in a rectangular-shaped probability density.

Probability Density Function (PDF) of Uniform Distribution:

The PDF of a continuous uniform distribution over the interval

[�,�]

$[a,b]$ is given by:

$$\text{[illegible]}(\text{[illegible]})=1\text{[illegible]}-\text{[illegible]}$$

$$f(x)=$$

$$b-a$$

$$1$$

where:

- [illegible]
- x is a random variable within the interval
- [illegible]
- $[a,b]$,
- [illegible]
- a is the lower bound of the interval,
- [illegible]
- b is the upper bound of the interval.

Example of Uniform Distribution:

Let's consider an example of a uniform distribution for the roll of a fair six-sided die. In this case:

- The outcomes are the numbers 1, 2, 3, 4, 5, and 6.
- Each outcome has an equal probability of
- 16
    - 6
    - 1
    -
- of occurring.

The probability density function for this uniform distribution is:

$$\text{□}(\text{□})=16$$

$$f(x)=$$

$$6$$

$$1$$

for

□

x in the range

$$[1,6]$$

$$[1,6].$$

Graphically, the PDF of a uniform distribution looks like a rectangle, where the height of the rectangle represents the constant probability density across the interval.

Properties of Uniform Distribution:

Constant Probability Density:

- The PDF is constant over the entire range, indicating that each value within the interval is equally likely.

Equal Probability:

- All values in the interval have the same probability of occurring.

Rectangular Shape:

- The probability density function graphically forms a rectangle, as opposed to the bell-shaped curve of the normal distribution.

Cumulative Distribution Function (CDF):

- The cumulative distribution function is a piecewise linear function that increases linearly within the interval.

Used in Probability Modeling:

- The uniform distribution is often used in probability modeling when there is no reason to believe that one outcome is more likely than another within a given range.

In summary, the uniform distribution is a straightforward and symmetric probability distribution where each outcome within a specified interval is equally likely. It is commonly used in situations where there is no preference for one outcome over another within the specified range.

Q8: What is the z score? State the importance of the z score.
Ans:Z-Score:

The Z-score, also known as the standard score or z-value, is a statistical measure that quantifies the number of standard deviations a data point is from the mean of a dataset. It is calculated using the formula:

$$Z = \frac{X - \mu}{\sigma}$$

where:

- Z is the Z-score,
- X is the individual data point,
- μ is the mean of the dataset,
- σ is the standard deviation of the dataset.

The Z-score indicates how many standard deviations a data point is from the mean and in which direction. A positive Z-score means the data point is above the mean, while a negative Z-score means the data point is below the mean.

Importance of Z-Score:

Standardization:

- Z-scores standardize data, allowing for the comparison of scores from different datasets with different units or scales. It transforms data into a common scale with a mean of 0 and a standard deviation of 1.

Identification of Outliers:

- Z-scores help identify outliers or extreme values in a dataset. Data points with Z-scores significantly larger or smaller than 0 may be considered outliers.

Probability and Normal Distribution:

- In a standard normal distribution (with mean 0 and standard deviation 1), Z-scores directly correspond to probabilities. The Z-score indicates the probability of a data point occurring within a certain range.

Data Analysis and Inference:

- Z-scores are used in hypothesis testing and statistical inference. They help determine how extreme an observed value is under a given hypothesis.

Quality Control:

- In quality control, Z-scores are used to assess whether a process is producing products within an acceptable range. Values outside a certain Z-score threshold may indicate a problem in the process.

Grade Comparisons:

- In education, Z-scores allow for the comparison of scores from different exams or classes. They provide information about a student's performance relative to the mean.

Normalization in Machine Learning:

- Z-scores are used in feature scaling and normalization in machine learning. Standardizing features ensures that no single feature dominates the learning algorithm.

Risk Assessment in Finance:

- In finance, Z-scores are used to assess the financial health and bankruptcy risk of a company. A low Z-score may indicate financial distress.

In summary, the Z-score is a valuable statistical tool that provides a standardized measure of how far a data point is from the mean, facilitating comparisons, identification of outliers, and various statistical analyses. It is widely used in various fields for data analysis, interpretation, and decision-making.

Q9: What is Central Limit Theorem? State the significance of the Central Limit Theorem.
Ans:Central Limit Theorem (CLT):

The Central Limit Theorem is a fundamental concept in probability and statistics that describes the shape of the sampling distribution of the sample mean (

$\bar{X}$

) for a large enough sample size, regardless of the shape of the original population distribution. It states that, as the sample size increases, the sampling distribution of the sample mean approaches a normal distribution, even if the population distribution is not normal.

The Central Limit Theorem is particularly powerful because it allows statisticians to make inferences about population parameters (such as the population mean) based on the distribution of sample means, assuming certain conditions are met.

Key Points of the Central Limit Theorem:

Large Sample Size:
- The Central Limit Theorem applies primarily to sufficiently large sample sizes. As a rule of thumb, a sample size of 30 or more is often considered "large enough" for the Central Limit Theorem to be effective.

Regardless of Population Distribution:
- The original population distribution does not have to be normal. The Central Limit Theorem holds for any population distribution, including those that are not symmetric or bell-shaped.

Sampling Distribution is Normal:
- The sampling distribution of the sample mean becomes approximately normal as the sample size increases, regardless of the shape of the original distribution.

Important for Inference:
- The Central Limit Theorem is crucial for making statistical inferences about population parameters. It allows for the use of normal distribution-based methods in hypothesis testing, confidence intervals, and other statistical analyses.

Mean and Standard Deviation:
- The mean of the sampling distribution of the sample mean (
- $\bar{X}$

- X
- ¯
-) is equal to the mean of the population (
-
- μ).
- The standard deviation of
- ¯
- X
- ¯
- (often called the standard error) is equal to the population standard deviation divided by the square root of the sample size (
- /
- σ/
- n
-
-).

Significance of the Central Limit Theorem:

Widespread Applicability:

- The Central Limit Theorem is applicable to a wide range of real-world situations, making it a fundamental tool in statistical analysis.

Inferential Statistics:

- It forms the basis for many inferential statistical techniques, allowing researchers and analysts to make population inferences based on sample data.

Simplifies Analysis:

- The theorem simplifies statistical analysis by allowing the use of normal distribution properties, even when dealing with non-normally distributed populations.

Sampling Distribution Stability:

- The Central Limit Theorem provides a stable and predictable shape for the sampling distribution of the sample mean, making it easier to make statistical predictions.

Foundation for Hypothesis Testing:

- Hypothesis testing often relies on the assumption of a normal distribution. The Central Limit Theorem allows for the use of normal distribution-based methods, enhancing the reliability of hypothesis tests.

In summary, the Central Limit Theorem is a cornerstone of statistical theory, providing a bridge between sample statistics and population parameters. It is a powerful tool that facilitates statistical inference and analysis in a wide range of practical scenarios.

Q10: State the assumptions of the Central Limit Theorem.
Ans:The Central Limit Theorem (CLT) is a powerful statistical concept, but it relies on certain assumptions to be valid. The key assumptions of the Central Limit Theorem are:

Random Sampling:

- The samples must be selected randomly from the population. This means that each member of the population has an equal chance of being selected. Non-random sampling methods can introduce biases that may affect the validity of the CLT.

Independence:

- The individual observations in the sample must be independent of each other. This means that the occurrence or value of one observation should not influence the occurrence or value of another observation. Independence is crucial for the reliability of the CLT.

Sample Size:

- The sample size should be sufficiently large. While there is no strict threshold, a commonly used guideline is that the sample size should be greater than 30. For smaller sample sizes, the CLT may not apply, and other methods may need to be considered.

Population Distribution Shape:

- The original population distribution (from which samples are drawn) does not have to be normal. However, for extremely skewed or heavily tailed distributions, larger sample sizes may be needed for the CLT to be effective. The CLT is most robust when the population distribution is not highly skewed.

Finite Variance:

- The population must have a finite variance (
- σ^2
-). If the population variance is infinite, the conditions for the CLT may not be met.

Same Finite Mean:

- The population must have a finite mean (
- μ). If the mean is not finite, the CLT may not hold.

It's important to note that while the CLT is robust and often works well even if some assumptions are not perfectly met, researchers and analysts should be aware of the assumptions and consider them in the context of their specific application.

If the sample size is small or the data violates the assumptions, alternative statistical methods may need to be considered. In practice, the CLT is a valuable guideline, but it's crucial to use judgment and explore other statistical approaches when necessary.