9 MARCH PW SKILLS ASSIGNMENT

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

The Probability Mass Function (PMF) and Probability Density Function (PDF) are both mathematical functions used to describe the probability distribution of a random variable.

1. Probability Mass Function (PMF):
The PMF is used for discrete random variables, which are variables that can only take on specific values. The PMF assigns probabilities to each possible value of the random variable. It gives the probability that the random variable takes on a particular value.

Example:
Consider a fair six-sided die. The random variable in this case is the outcome of rolling the die. The PMF of this random variable can be defined as follows:
$P(X = 1) = 1/6$
$P(X = 2) = 1/6$
$P(X = 3) = 1/6$
$P(X = 4) = 1/6$
$P(X = 5) = 1/6$
$P(X = 6) = 1/6$
Here, X represents the random variable, and $P(X = x)$ denotes the probability that the outcome is x. The PMF assigns an equal probability of 1/6 to each possible outcome, as the die is fair.

2. Probability Density Function (PDF):
The PDF is used for continuous random variables, which can take on any value within a given range. The PDF describes the relative likelihood of the random variable taking on different values. It gives the probability density at each point on the distribution.

Example:
Consider a standard normal distribution with mean 0 and standard deviation 1. The random variable in this case is a normally distributed variable. The PDF of this random variable is given by the following formula:
$f(x) = (1 / \sqrt{(2\pi)}) * e^{(-x^2/2)}$
Here, f(x) represents the PDF, and x is a particular value on the distribution. The PDF describes the shape of the bell curve and the relative likelihood of the random variable taking on different values. The area under the curve within a range represents the probability of the random variable falling within that range.

Q2: What is Cumulative Density Function (CDF)? Explain with an example. Why CDF is used?

The Cumulative Density Function (CDF) is a function used to describe the cumulative probability distribution of a random variable. It gives the probability that a random variable takes on a value less than or equal to a specific value.

The CDF is defined as follows:
$F(x) = P(X \leq x)$

Where F(x) is the CDF of the random variable X, and $P(X \leq x)$ represents the probability that X takes on a value less than or equal to x.

Example:
Let's consider a continuous random variable X that follows a standard normal distribution with mean 0 and standard deviation 1. The CDF of this random variable can be expressed using the following formula:
$F(x) = \int[-\infty, x] f(t)\, dt$

Here, f(t) represents the Probability Density Function (PDF) of the random variable, and the integral calculates the area under the PDF curve from negative infinity to x. The resulting value represents the cumulative probability up to x.

For example, if we want to find the probability that X is less than or equal to 1 (i.e., $P(X \leq 1)$), we can calculate the CDF as follows:
$F(1) = \int[-\infty, 1] f(t)\, dt$

The CDF provides several benefits:

1. Probability Calculation: The CDF allows us to calculate the probability that a random variable falls within a given range. For example, by subtracting the CDF value at one point from another, we can find the probability of the random variable lying within that range.

2. Quantile Calculation: The CDF enables us to determine the value of a random variable for a given probability. By inverting the CDF, we can find the quantiles or percentiles of a distribution.

3. Distribution Comparison: The CDF helps in comparing different probability distributions. By plotting the CDFs of different distributions on the same graph, we can visually analyze and compare their shapes and characteristics.

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.

The normal distribution, also known as the Gaussian distribution or bell curve, is a widely used statistical model in various fields due to its versatility and applicability to many real-world situations. Some examples where the normal distribution might be used as a model include:

1. Measurement Errors: When measuring physical quantities or experimental data, errors are often present. These errors are commonly assumed to follow a normal distribution.

2. IQ Scores: Intelligence quotient (IQ) scores are often modeled using a normal distribution, where the mean represents the average IQ and the standard deviation represents the variation around the mean.

3. Height and Weight: Human height and weight often follow a normal distribution, with the mean representing the average height or weight and the standard deviation indicating the variation in measurements.

4. Financial Markets: Stock prices and investment returns are often assumed to follow a normal distribution in financial modeling, allowing for risk analysis and portfolio optimization.

The parameters of the normal distribution are the mean (μ) and the standard deviation (σ). The mean determines the center of the distribution, representing the most likely or average value. The standard deviation determines the spread or dispersion of the data points around the mean. A smaller standard deviation results in a narrower, more peaked distribution, while a larger standard deviation leads to a wider, flatter distribution.

By adjusting the mean and standard deviation, the shape of the normal distribution can be modified. Shifting the mean changes the center point, while changing the standard deviation alters the spread of the data. Together, these parameters allow the normal distribution to adapt to various data patterns, making it a powerful tool for modeling and analysis.

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

The normal distribution is important in statistics and data analysis due to its many properties and applications. Real-life examples include IQ scores, heights and weights of populations, measurement errors, exam scores, and financial market returns. It provides a useful approximation for a wide range of phenomena, simplifies analysis, and enables accurate predictions.

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

The Bernoulli distribution is a discrete probability distribution that models a single binary event with two possible outcomes: success (typically denoted as 1) or failure (typically denoted as 0). It is characterized by a parameter p, which represents the probability of success in a single trial.

Example: Flipping a coin, where heads is considered a success (1) and tails is considered a failure (0), can be modeled using a Bernoulli distribution with p = 0.5.

The main difference between Bernoulli and Binomial distributions is that Bernoulli deals with a single trial, while the Binomial distribution deals with multiple independent and identically distributed Bernoulli trials. The Binomial distribution represents the number of successes in a fixed number of trials, each with the same probability of success.

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.

Mean = 50
Std = 10
Xi = 60

Z-score = 60-50/10 = 1

Area under the curve (x>=60)= 1-0.84134
= 15.866%

Q7: Explain uniform Distribution with an example.

The uniform distribution is a continuous probability distribution where all outcomes within a specified range are equally likely. It has a constant probability density function (PDF) over the range. An example is rolling a fair six-sided die, where each outcome has an equal probability of 1/6.

Q8: What is the z score? State the importance of the z score.

The z-score, also known as the standard score, is a statistical measure that quantifies the number of standard deviations a data point is away from the mean of a distribution. It is calculated by subtracting the mean from the data point and dividing the result by the standard deviation.

The importance of the z-score lies in its ability to standardize and compare data points from different distributions. It allows for meaningful comparisons, as it expresses a data point's relative position within a distribution. The z-score provides information about how unusual or typical a particular observation is compared to others in the dataset. It is widely used in hypothesis testing, identifying outliers, establishing confidence intervals, and making predictions based on standardized data.

Q9: What is Central Limit Theorem? State the significance of the Central Limit Theorem.

The Central Limit Theorem (CLT) is a fundamental concept in statistics that states that when independent random variables are summed or averaged, regardless of the shape of their original distributions, the resulting distribution approaches a normal distribution as the sample size increases.

The significance of the Central Limit Theorem is profound. It enables the use of the normal distribution as an approximation for the sampling distribution of various statistics, even if the population distribution is not normal. This is crucial because the normal distribution is well understood and has many useful properties. The CLT forms the foundation for inferential statistics, allowing us to make reliable inferences about a population based on a sample. It is widely applied in hypothesis testing, confidence intervals, regression analysis, and many other statistical techniques.

Q10: State the assumptions of the Central Limit Theorem.

The Central Limit Theorem (CLT) relies on the following assumptions:

1. Independence: The random variables being summed or averaged are independent of each other.
2. Finite Variance: The variables have a finite variance.
3. Sample Size: The sample size is sufficiently large (although there is no strict threshold, a general guideline is a sample size greater than 30).
4. No Extreme Outliers: The presence of extreme outliers or influential observations should be limited, as they can impact the convergence to a normal distribution.