statistics-advance-assignment-2

August 13, 2023

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

Probability Mass Function (PMF): The PMF is used for discrete random variables. It gives the probability that a discrete random variable takes on a specific value.

Example: Consider rolling a fair six-sided die. The random variable X represents the outcome of the roll. The PMF for X would be:

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 Each value has an equal probability of 1/6 of occurring.

Probability Density Function (PDF): The PDF is used for continuous random variables.

Example: Consider a continuous random variable Y that represents the height of individuals in a population (assuming heights are measured in a continuous manner). The PDF for Y might be represented by a bell-shaped curve, like the normal distribution.

The PDF provides information about the likelihood of observing heights within certain intervals, but it does not give the probability of a specific height occurring.

In summary, the PMF is used for discrete random variables, providing probabilities for specific values, while the PDF is used for continuous random variables, describing the distribution's shape and likelihood of values within intervals.

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

The Cumulative Distribution Function (CDF) is a concept used in probability and statistics 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 has several important properties:

The CDF is non-decreasing: As x increases, the cumulative probability can only stay the same or increase. The CDF is bounded: It ranges between 0 and 1, inclusive. The CDF approaches 0 as x approaches negative infinity and approaches 1 as x approaches positive infinity.

Example: Consider a simple example using a fair six-sided die. The random variable X represents the outcome of a single roll of the die. The CDF for X would be as follows:

F X (1)=P(X 1)=1/6
F X (2)=P(X 2)=2/6

F X (3)=P(X 3)=3/6
F X (4)=P(X 4)=4/6
F X (5)=P(X 5)=5/6
F X (6)=P(X 6)=6/6=1

The CDF tells us, for example, that the probability of getting a value less than or equal to 3 on a single roll of the die is 3 / 6 = 0.5 3/6=0.5

The CDF is a valuable concept in probability and statistics for several reasons:

Probability Calculation: The CDF provides an easy way to calculate the probability of a random variable falling within a specific range.

Comparisons: The CDF allows for quick comparisons of probabilities between different values or ranges of the random variable.

Quantile Calculation: The CDF can be used to find quantiles, which are values that divide the distribution into specific probabilities.

Visual Representation: The CDF can be visualized graphically to show how the probability accumulates as the value of the random variable increases

[ ]:

```
# 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.
```

. Here are some examples of situations where the normal distribution might be used as a model:

Height of Individuals: The heights of a large population tend to follow a normal distribution. While there may be some variations due to genetics and environmental factors, the distribution of heights often resembles a bell-shaped curve.

IQ Scores: IQ scores in a population tend to be normally distributed, with most people clustering around the average IQ score.

Stock Market Returns: Daily stock market returns are often modeled using a normal distribution for quantitative finance purposes, despite the fact that real financial data can exhibit heavier tails and volatility clustering.

The parameters of the normal distribution are the mean ( ) and the standard deviation ( ). These parameters play a crucial role in shaping the distribution:

Mean ( ): The mean determines the central location of the distribution. It is the value around which the data tend to cluster. The mean is also the point of symmetry for the normal distribution.

Standard Deviation ( ): The standard deviation measures the spread or dispersion of the data points around the mean. A larger standard deviation results in a wider distribution, while a smaller standard deviation results in a narrower distribution.

The shape of the normal distribution is influenced by both the mean and the standard deviation:

If the mean is shifted to the right, the entire distribution is shifted to the right. If the standard deviation is increased, the distribution becomes wider and flatter. If the standard deviation is decreased, the distribution becomes narrower and taller.

[ ]:

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

The normal distribution, also known as the Gaussian distribution or bell curve, holds significant importance in various fields due to its mathematical properties and its ability to model a wide range of real-world phenomena.

Here are some reasons why the normal distribution is important: Statistical Inference: The central limit theorem states that the sum (or average) of a large number of independent and identically distributed random variables approaches a normal distribution, regardless of the original distribution. This property is essential for statistical inference, hypothesis testing, and confidence interval estimation.

Simplification of Complex Systems: The normal distribution provides a simple and convenient way to approximate complex systems. Even when the real data distribution is not perfectly normal, the normal distribution can often serve as a reasonable approximation.

Real-life examples of situations where the normal distribution is commonly observed include:

Human Height: Heights of individuals in a large population tend to follow a normal distribution, with most people around the mean height.

Test Scores: Standardized test scores, such as SAT or IQ scores, often follow a normal distribution.

Error in Measurement: Measurement errors in scientific experiments often follow a normal distribution due to the influence of multiple random factors.

Body Weight: Body weights of a large and diverse population can approximate a normal distribution.

Temperature: Daily temperatures in a region can often be modeled using a normal distribution.

Income Distribution: In some cases, the distribution of income in a population may approximate a normal distribution.

[ ]:

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

A discrete probability distribution for a random experiment that has only two possible outcomesis known Bernoulli Distribution.

Example: India will win cricket world cup or not • It has only two possible outcomes • Success (1) • Failure (0)

Bernaulli Distribution Deals with the single trial event Binomial Deals with the outcome of Multiple trials of the single events

Bernaulli Distribution Has only two possible outcome 0 and 1 Binomial Sum of identically and independent distributed Bernoulli Random Variable

[ ]:

```
# 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
```

We can use the Z-score formula and the standard normal distribution table. The Z-score is a measure of how many standard deviations an observation is away from the mean.

The formula to calculate the Z-score is:

Z = (X -) /

Where:

X is the value we want to find the probability for (in this case, 60) is the mean of the dataset (50 in this case) is the standard deviation of the dataset (10 in this case)

Substitute the values:

Z = (60 - 50) / 10 Z = 1

Using the standard normal distribution table or a calculator, you'll find that the cumulative probability up to Z = 1 is approximately 0.8413. Therefore, the probability that a randomly selected observation will be greater than 60 is:

1 - 0.8413 0.1587

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

```
[ ]: # Q7: Explain uniform Distribution with an example
```

The uniform distribution, also known as the rectangular distribution, is a probability distribution that describes a situation where all values within a certain range are equally likely to occur. In other words, in a uniform distribution, every possible outcome has an equal probability of occurring.

Example: Rolling a Fair Die Consider a fair six-sided die. When you roll the die, each face (1, 2, 3, 4, 5, or 6) has an equal probability of landing face up. This situation can be described by a uniform distribution.

```
[ ]: # 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 how many standard deviations a particular data point is away from the mean of a dataset.

The formula to calculate the z-score for a data point "x" within a dataset with mean " " and standard deviation " " is:

z = (x -) /

The importance of the z-score lies in its ability to standardize data and provide a common basis for comparing different data points within the same dataset or between different datasets.

Normalization and Standardization: Z-scores transform raw data into a standard scale where the mean is 0 and the standard deviation is 1. This allows you to compare data from different distributions on a common scale, making it easier to identify outliers, trends, and patterns.

Identifying Outliers: Data points with z-scores far from zero (typically greater than ±2 or ±3) are considered outliers. Outliers could indicate errors in data collection or significant deviations from the norm.

Probability and Percentiles: Z-scores correspond to probabilities in the standard normal distribution. By looking up z-scores in a standard normal distribution table or using statistical software, you can determine the probability of a data point occurring within a certain range.

Hypothesis Testing: In hypothesis testing, z-scores help determine whether a sample mean is significantly different from a population mean. By comparing the calculated z-score with critical values, you can make informed decisions about hypotheses.

[ ]:

```
# 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 the sampling distribution of the sample means (or averages) of a large enough number of independent, identically distributed random variables will be approximately normally distributed, regardless of the distribution of the original population.

In simpler terms, the Central Limit Theorem says that when you take multiple random samples from any population, the distribution of the sample means will tend to follow a normal distribution, even if the original population's distribution is not normal.

Significance of the Central Limit Theorem:

Normality Assumption: The Central Limit Theorem is particularly important because it allows us to make assumptions about the distribution of sample means, even when we don't know the distribution of the population itself. This is crucial because the normal distribution is well-understood and has many convenient mathematical properties.

Inferential Statistics: The CLT underpins many statistical techniques and hypothesis tests. For example, when we perform hypothesis testing on sample means, we often assume that the distribution of the sample means is normal. The CLT validates this assumption, enabling us to use methods like Z-tests and t-tests.

Estimation: When estimating population parameters, such as the population mean, the CLT allows us to use the normal distribution as an approximation for the distribution of sample means. This is used in constructing confidence intervals for population parameters.

[ ]:

```
# Q10: State the assumptions of the Central Limit Theorem
```

The Central Limit Theorem (CLT) is a powerful statistical principle, but it comes with certain assumptions that need to be met in order for it to hold.

The data must adhere to the randomization rule. It needs to be sampled at random.

The samples should be unrelated to one another. One sample should not impact the others.

When taking samples without replacement, the sample size should not exceed 10% of the population.

When the population is symmetric, a sample size of 30 is generally considered reasonable.