Assignment Code: DA-AG-006

Statistics Advanced - 1| Assignment

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

Total Marks: 200

Question 1: What is a random variable in probability theory?

Answer:

A random variable is a core concept in probability theory and statistics, frequently used in data science, machine learning, and quantitative analysis.

A **random variable** is a function that assigns a numerical value to each outcome of a random experiment. It's a key concept that bridges the gap between the outcomes of an experiment and the numerical values that we can use for mathematical analysis in probability and statistics.The random variable is the result of applying the random variable to an observed outcome of a random experiment. This is what the experimenter actually observes. The realization of a random variable is typically denoted using lowercase italicized Roman letters, e.g., x is a realization of X.

Example : Flipping a two-sided coin is a random process because we do not know if we will observe heads or tails. Flipping a one-sided coin is a deterministic process because we know that we will always observe heads.

Question 2: What are the types of random variables?

Answer:

There are two types of random variables:-

A **Discrete Random Variables** can take on a finite or countably infinite number of values. In advanced statistics, the study of discrete variables goes beyond simple probability distributions to include:

Probability Mass Functions: This function, denoted as P gives the probability that a discrete random variable X takes on a specific value x.

Common Distributions: We'll delve into the properties of specific discrete distributions like the Binomial, Poisson, and Geometric.

Expected Value and Variance: The mean and variance are crucial measures of central tendency and spread, calculated using the PMF.

A **Continuous Random Variables** can take on any value within a given interval. The key difference from the discrete case is that the probability of a continuous variable equaling a specific value is zero. Instead, we use:

Probability Density Functions (PDF): This function, f(x), describes the relative likelihood of a variable taking a certain value. The probability of the variable falling within an interval [a,b] is the integral of the PDF over that interval: $P = \int abf(x)dx$

Common Distributions: Advanced statistics explores continuous distributions such as the Normal, Exponential, and Uniform.

Question 3: Explain the difference between discrete and continuous distributions.

Answer:

The primary difference between discrete and continuous distributions lies in the nature of the data they model. Discrete distributions describe data that can only take on a finite or countable number of distinct values, while continuous distributions describe data that can take on any value within a given range.

Discrete Distributions

Data Type: Used for counting things. The outcomes are distinct and separate, often whole numbers.

Examples:

The number of defective items in a batch.

The result of rolling a die.

The number of cars that pass a checkpoint in an hour.

Probability Function: A Probability Mass Function is used to define the probability for each individual value. The sum of all probabilities for every possible outcome must equal 1. For example, the probability of rolling a 3 on a fair die is exactly 1/6.

Visualization: Often represented by a bar graph, where each bar's height shows the probability of a specific outcome.

Continuous Distributions

Data Type: Used for measuring things. The outcomes can be any value within a range, including decimals and fractions. There are infinitely many possible values.

Examples:

A person's height.

The amount of rainfall in a day.

The time it takes to run a race.

Probability Function: A Probability Density Function (PDF) is used. Because there are infinite possibilities, the probability of a continuous variable taking on any *exact* value is considered zero. Instead, we calculate the probability of the variable falling within a specific interval or range by finding the area under the PDF curve. The total area under the entire curve is always 1.

Visualization: Typically represented by a smooth curve, where the probability corresponds to the area under the curve.

Question 4: What is a binomial distribution, and how is it used in probability?
Answer:

A **binomial distribution** is a discrete probability distribution that models the number of successes in a fixed number of independent trials. Each trial must have only two possible outcomes, typically labeled "success" or "failure," and the probability of success must be the same for every trial.

The binomial distribution is defined by two parameters: the number of trials (n) and the probability of success on any single trial (p).

The binomial distribution is used to calculate the probability of getting exactly k successes in n trials. This is done using the probability mass function : $P(X = k) = (\frac{n}{k})p^{k}(1 - p)^{n-k}$

Here, (n/k) is the binomial coefficient, which represents the number of ways to choose k successes from n trials.

When we use binomial distribution :

We have a fixed number of trials.

Each trial is independent.

Each trial has only two outcomes.

The probability of success remains constant.

Question 5: What is the standard normal distribution, and why is it important?

Answer:

Standard normal distribution, also known as the z-distribution, is a special type of normal distribution. In this distribution, the mean is 0 and the standard deviation is 1. This creates a bell-shaped curve that is symmetrical around the mean.

The standard normal distribution is a special case of the normal distribution with:

Mean (μ) = 0

Standard deviation (σ) = 1

1. Universal Applicability

Many natural phenomena follow a normal distribution.

The Central Limit Theorem states that the sum of many independent random variables tends toward a normal distribution—even if the original variables aren't normal.

2. Z-Scores and Standardization :

It allows us to standardize and compare data from any normal distribution. By converting a value from any normal distribution into a z-score, we express it in terms of how many standard deviations it is away from the mean. The formula for a z-score is : $z = \frac{x-\mu}{\sigma}$

This allows comparison across different datasets and scales.

3. Hypothesis Testing & Confidence Intervals :

The standard normal distribution is the foundation for many statistical tests, such as the z-test. These tests rely on the properties of the standard normal distribution to make inferences about a population based on a sample. For instance, in hypothesis testing, we use the z-distribution to determine if a sample mean is significantly different from a population mean..

Question 6: What is the Central Limit Theorem (CLT), and why is it critical in statistics?

Answer:

The **Central Limit Theorem** (CLT) states that if we take a sufficiently large number of random samples from a population, the distribution of the sample means will be approximately normal, regardless of the original population's distribution. The larger the sample size, the more closely the distribution of the sample means will resemble a normal distribution.
The Central Limit Theorem is one of the most fundamental and powerful concepts in statistics for several reasons:

Foundation for Inferential Statistics: The CLT is the basis for most statistical inference, allowing us to make conclusions about a population based on a sample. It justifies the use of parametric tests like z-tests and t-tests, even when the underlying population data isn't normally distributed.

Simplifies Complex Calculations: Without the CLT, we'd need to know the exact distribution of the population to make accurate probability statements. By guaranteeing that the sample means will follow a normal distribution, the CLT allows us to use the properties of the standard normal distribution to calculate probabilities and construct confidence intervals, simplifying complex analyses.

Relationship Between Sample and Population: The theorem establishes a clear link between a population and its samples. It shows that the mean of the sampling distribution of the sample means will be equal to the population mean (μ), and the standard deviation of this sampling distribution will be the population standard deviation (σ) divided by the square root of the sample size ($\sqrt{n}$) this predictability is essential for estimating population parameters.

Question 7: What is the significance of confidence intervals in statistical analysis?

Answer:

Confidence intervals are crucial in statistical analysis because they provide a range of plausible values for an unknown population parameter, rather than a single point estimate. They are a fundamental tool for quantifying and communicating the uncertainty inherent in using sample data to make inferences about a larger population.

Quantifying Uncertainty: A confidence interval gives us a range of values where we can be confident the true population parameter lies. For example, a statement like "We are 95% confident that the true average height of students is between 155 cm and 165 cm" is far more informative than just saying "The average height is 160 cm". The confidence level represents the long-run probability that the interval making procedure would capture the true population parameter if the experiment were repeated many times.

Assessing Precision: The width of a confidence interval tells us about the precision of our estimate. A narrower interval suggests a more precise estimate, typically due to a larger sample size or lower data variability. A wider interval indicates more uncertainty and a less precise estimate.

Enhancing Decision-Making: Confidence intervals help in making better-informed decisions. For instance, if we're comparing two marketing campaigns, a confidence interval for the difference in their effectiveness can tell us if the difference is meaningful or likely just due to random chance. If the interval for the difference includes zero, it suggests there's no statistically significant difference between the campaigns.

Question 8: What is the concept of expected value in a probability distribution?

Answer:

The expected value (μ) is the weighted average of all possible values of a random variable, where the weights are their probabilities.

For a Discrete Random Variable:

If XXX is a discrete random variable that takes values x1,x2,x3,…with corresponding probabilities P(X=xi)=pi, then the expected value E[X] is:
$E(X) = \sum x \cdot P(X = x)$

For a Continuous Random Variable:

If X is a continuous random variable with probability density function (PDF) f(x), then the expected value is: $E(X) = \int_{-\infty}^{\infty} x.f(x)dx$

The expected value is a crucial concept in many fields, including:

Gambling and Finance: It helps determine the average payout or return on an investment over the long term. A positive expected value suggests a profitable venture, while a negative one indicates a potential long-term loss.

Decision Theory: It's used to make decisions under uncertainty by comparing the expected outcomes of different choices.

Statistics: It is one of the most important moments of a distribution, providing a simple, single number that summarizes the center of the data.

Question 9: Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution. (*Include your Python code and output in the code box below.*)

Answer:

```
import numpy as np
import matplotlib.pyplot as plt
```

```
mean = 50
std_dev = 5
n = 1000

random_numbers = np.random.normal(loc=mean, scale=std_dev, size=n)

computed_mean = np.mean(random_numbers)
computed_std = np.std(random_numbers)

print(f"Computed Mean: {computed_mean:.2f}")
print(f"Computed Standard Deviation: {computed_std:.2f}")

plt.hist(random_numbers, bins=30, color='skyblue', edgecolor='black',
density=True)
plt.title("Histogram of Normally Distributed Random Numbers")
plt.xlabel("Value")
plt.ylabel("Density")

plt.axvline(computed_mean, color='red', linestyle='dashed', linewidth=2,
label=f'Mean = {computed_mean:.2f}')
plt.legend()
plt.show()
```

OUTPUT

Computed Mean : 50.06

Computed Standard Deviation : 5.09

Question 10: You are working as a data analyst for a retail company. The company has collected daily sales data for 2 years and wants you to identify the overall sales trend. daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,

235, 260, 245, 250, 225, 270, 265, 255, 250, 260]

- Explain how you would apply the Central Limit Theorem to estimate the average sales with a 95% confidence interval.
- Write the Python code to compute the mean sales and its confidence interval. (*Include your Python code and output in the code box below.*)

Answer:

Estimating Average Sales Using CLT

1.Understand the Central Limit Theorem :

Use the normal distribution to estimate population parameters

Construct confidence intervals around the sample mean

2.Collect a Sample of Sales Data :

our sample from a larger population of daily sales.

3.Calculate the Sample Mean :

an estimate of the average daily sales.

4.Calculate the Sample Standard Deviation

5.Compute the Standard Error

6.Find the Z-Score for 95% Confidence

7.Compute the Confidence Interval

8.Interpret the Result

```
import numpy as np

from scipy.stats import norm



daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,235,
260, 245, 250, 225, 270, 265, 255, 250, 260]



mean_sales = np.mean(daily_sales)

std_sales = np.std(daily_sales, ddof=1)

n = len(daily_sales)
```

```
SE = std_sales / np.sqrt(n)
z_score = norm.ppf(0.975)  # 2-tailed, so 0.975

CI_lower = mean_sales - z_score * SE
CI_upper = mean_sales + z_score * SE

# Output results
print(f"Mean Daily Sales: {mean_sales:.2f}")
print(f"95% Confidence Interval: ({CI_lower:.2f}, {CI_upper:.2f})")
```

OUTPUT

Mean Daily Sales = 248.25

95% Confidence Interval (240.68 ,255.82)