Assignment Code: DA-AG-006

Statistics Advanced - 1| Assignment

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

Answer:

In probability theory, a **random variable** is a function that assigns a numerical value to each outcome in a sample space of a random experiment. It provides a way to quantify outcomes and analyze them mathematically.

Types of Random Variables

1. **Discrete Random Variable**
   - Takes on a countable number of distinct values.
   - Examples: Number of heads in 3 coin tosses, number of students in a class.
2. **Continuous Random Variable**
   - Takes on an infinite number of possible values within a given range.
   - Examples: Height of a person, time taken to run a race.

Formal Definition

Let S be the sample space of a random experiment. A random

variable X is a function:

X: S \rightarrow \mathbb{R}

that maps each outcome s \in S to a real number X(s).

Example

Suppose you roll a fair six-sided die. The sample space is:

S = \{1, 2, 3, 4, 5, 6\}

Define a random variable X as the outcome of the roll. Then:

X(s) = s \quad \text{for each } s \in S

Why It Matters

Random variables allow us to:

- Calculate probabilities (e.g., P(X = 3))
- Define distributions (e.g., binomial, normal)
- Compute expected values, variances, and other statistical measures

---

Question 2: What are the types of random variables?

Answer:

Random variables are primarily classified into two main types based on the nature of the values they can take:

1**. Discrete Random Variable**

- Definition: A random variable that can take on a finite or countably infinite set of distinct values.
- Values: Typically integers or whole numbers.
- Examples:
  - Number of heads in 10 coin tosses.
  - Number of students present in a class.
  - Outcome of rolling a die: {1, 2, 3, 4, 5, 6}.
- Probability Distribution: Described by a probability mass function (PMF), which assigns probabilities to each possible value.

2. **Continuous Random Variable**

- Definition: A random variable that can take on any value within a given interval, including fractions and irrational numbers.
- Values: Uncountably infinite; typically real numbers.
- Examples:
  - Height of a person.
  - Time taken to complete a task.
  - Temperature at a given location.
- Probability Distribution: Described by a probability density function (PDF), and probabilities are calculated over intervals (e.g., P(a \leq X \leq b)).

The comparison between discrete and continuous random variables presented in line format:

- **Value Set**:
  - Discrete random variables take on countable values (either finite or countably infinite).
  - Continuous random variables take on uncountable values, typically any real number within an interval.

- **Examples**:
  - Discrete: Dice rolls, number of heads in coin tosses, number of students in a class.
  - Continuous: Height of a person, time taken to complete a task, temperature readings.

- **Probability Function**:
  - Discrete random variables use a **Probability Mass Function (PMF)** to assign probabilities to specific values.
  - Continuous random variables use a **Probability Density Function (PDF)** to define probabilities over intervals.

- **Probability of Exact Value**:
- For discrete random variables, the probability of a specific value can be non-zero.
- For continuous random variables, the probability of any exact value is zero; probabilities are calculated over ranges.

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

Answer:

The difference between **discrete** and **continuous distributions** lies in the type of random variable they describe and how probabilities are assigned.

Discrete Distribution

- **Definition**: Describes the probability distribution of a **discrete random variable**, which takes on countable values.
- **Probability Assignment**: Probabilities are assigned to **individual values**.
- **Probability Function**: Uses a **Probability Mass Function (PMF)**.
- **Key Property**:
- P(X = x) > 0 \quad \text{for some values of } x**Examples**:
- **Binomial Distribution**: Number of successes in a fixed number of Bernoulli trials.
- **Poisson Distribution**: Number of events occurring in a fixed interval of time or space.
- **Geometric Distribution**: Number of trials until the first success.

Continuous Distribution

- **Definition**: Describes the probability distribution of a **continuous random variable**, which takes on values from an interval of real numbers.
- **Probability Assignment**: Probabilities are assigned to **intervals**, not individual points.
- **Probability Function**: Uses a **Probability Density Function (PDF)**.
- **Key Property**:

- $P(X = x) = 0 \quad \text{for any specific value } x$Instead,
- $P(a \leq X \leq b) = \int_a^b f(x) \, dx$**Examples**:
- **Normal Distribution**: Bell-shaped curve describing many natural phenomena.
- **Exponential Distribution**: Time between events in a Poisson process.
- **Uniform Distribution**: Equal probability across a continuous interval.

Summary of Differences

- **Discrete distributions** deal with countable outcomes and assign probabilities to each one.
- **Continuous distributions** deal with uncountable outcomes and assign probabilities over ranges using integration.

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

Answer:

The **binomial distribution** is a discrete probability distribution that models the number of **successes** in a fixed number of **independent Bernoulli trials**, where each trial has only two possible outcomes:

- **Success** (e.g., heads, pass, win)
- **Failure** (e.g., tails, fail, lose)

Each trial must satisfy the following conditions:

- The number of trials n is fixed.
- Each trial is independent.
- The probability of success p remains constant across trials.

Parameters

- n: Total number of trials
- p: Probability of success in a single trial
- X: Random variable representing the number of successes

The probability of observing exactly k successes is given by the **binomial probability formula**:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}$$

Where:

- $\binom{n}{k}$ is the binomial coefficient, representing the number of ways to choose k successes from n trials.

Applications in Probability

The binomial distribution is used when:

- You are performing a fixed number of trials.

- Each trial has only two outcomes (success or failure).
- The trials are independent.
- The probability of success is constant.

Common Examples:

- Flipping a coin multiple times and counting the number of heads.
- Testing a batch of products and counting how many are defective.
- Surveying people and counting how many prefer a certain option.

Key Properties:

Property	Formula
Mean (Expected value)	\mu = np
Variance	\sigma^2 = np(1 - p)
Standard deviation	\sigma = \sqrt{np(1 - p)}

Example:

Suppose you flip a fair coin 5 times. What is the probability of getting exactly 3 heads?

- n = 5, p = 0.5, k = 3

P(X = 3) = \binom{5}{3} (0.5)^3 (0.5)^2 = 10 \cdot 0.125 \cdot 0.25 = 0.3125

So, there's a 31.25% chance of getting exactly 3 heads in 5 coin tosses.

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

Answer:

The **standard normal distribution** is a special case of the normal distribution that plays a central role in statistics and probability theory. Here's a clear breakdown:

What Is the Standard Normal Distribution?

The **standard normal distribution** is a **continuous probability distribution** with:

- **Mean** \mu = 0
- **Standard deviation** \sigma = 1

It is symmetric around the mean and follows the familiar **bell-shaped curve**.

The probability density function (PDF) is:

f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2 / 2}

This function describes how the values of a standard normal random variable Z are distributed.

Why Is It Important?

1. **Foundation for Statistical Inference**

Many statistical methods rely on the assumption of normality. The standard normal distribution is used to:

- Calculate **z-scores**
- Perform **hypothesis testing**
- Construct **confidence intervals**

2. **Z-Scores and Standardization**

Any normal distribution can be converted to the standard normal distribution using a **z-score**:

$$z = \frac{x - \mu}{\sigma}$$

This transformation allows comparisons across different normal distributions by placing them on a common scale.

3. **Central Limit Theorem (CLT)**

The CLT states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the original distribution. When standardized, this distribution becomes the standard normal.

4. **Probability Calculations**

The standard normal distribution is used to find probabilities and percentiles using **z-tables** or computational tools. For example:

- $P(Z < 1.96) \approx 0.975$
- $P(-1 < Z < 1) \approx 0.682$

These values are crucial in determining statistical significance.

Summary of Key Properties:

Property	Value
Mean	0
Standard deviation	1
Symmetry	yes(around 0)
Total area under curve	1
Shape	Bell curve

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

Answer:

The **Central Limit Theorem (CLT)** is one of the most powerful and foundational concepts in statistics. It explains why normal distributions appear so frequently in data analysis—even when the original data isn't normally distributed.

What Is the Central Limit Theorem?

The **Central Limit Theorem** states that:

When you take a large number of independent, identically distributed random samples from any population (with finite mean and variance), the distribution of the **sample means** will approximate a **normal distribution**, regardless of the shape of the original population.

Mathematically, if:

- X_1, X_2, ..., X_n are independent and identically distributed (i.i.d.) random variables with mean \mu and standard deviation \sigma,
- Then the sample mean \bar{X} will be approximately normally distributed as:

\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)

As n \to \infty, the approximation becomes increasingly accurate.

Why Is CLT Critical in Statistics?

1. **Enables Use of Normal Distribution**

Even if the original data is skewed or irregular, the CLT allows us to use the normal distribution to:

- Estimate probabilities
- Construct confidence intervals
- Perform hypothesis tests

2. **Justifies Many Statistical Methods**

Most parametric statistical techniques (like t-tests, ANOVA, regression) rely on the assumption of normality. CLT provides the theoretical basis for applying these methods to sample means.

3. **Simplifies Complex Problems**

It transforms problems involving unknown or complicated

distributions into manageable ones using the well-understood properties of the normal distribution.

4. **Supports Sampling Theory**

CLT explains why sample statistics (like the mean) are reliable estimators of population parameters, especially as sample size increases.

Example:

Suppose you measure the height of 10,000 people. The original distribution might be skewed due to outliers. But if you take repeated samples of 30 people and compute the mean height for each sample, the distribution of those sample means will be approximately normal—even if the original data isn't.

Key Takeaways:

Concept	Implication
Applies to sample means	Not necessarily to individual data points
Works for large n	Typically n\gep30 is sufficient
Requires finite variance	Infinite variance breaks the assumption
Enables inference	Makes statistical estimation possible

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

Answer:

Confidence intervals are a fundamental concept in statistical analysis because they provide a **range of plausible values** for an unknown population parameter, such as a mean or proportion, based on sample data.

What Is a Confidence Interval?

A **confidence interval (CI)** is a range of values, derived from sample statistics, that is likely to contain the true value of a population parameter with a specified level of confidence.

For example, a 95% confidence interval for a population mean might be:

\text{CI} = \bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right)

Where:

- \bar{x} is the sample mean
- z^* is the critical value from the standard normal distribution (e.g., 1.96 for 95% confidence)
- \sigma is the population standard deviation (or sample standard deviation if unknown)
- n is the sample size

Why Are Confidence Intervals Important?

1. **Quantify Uncertainty**

Confidence intervals express the uncertainty inherent in estimating population parameters from sample data. They show how precise or imprecise an estimate is.

2. **More Informative Than Point Estimates**

A single point estimate (like a sample mean) gives no sense of variability. A confidence interval provides a range, offering more context and reliability.

3. **Support Decision-Making**

Confidence intervals help assess whether observed effects are statistically significant. For example:

- If a 95% CI for a mean difference does **not** include zero, the difference is likely significant.
- If a CI for a proportion includes a critical threshold (e.g., 50%), it may affect conclusions.

4. **Used in Hypothesis Testing**

Confidence intervals are closely related to hypothesis tests. If a hypothesized value lies outside the interval, it is typically rejected at the corresponding confidence level.

<u>Interpretation</u>

A 95% confidence interval means:

If we were to take many samples and compute a confidence interval from each, approximately 95% of those intervals would contain the true population parameter.

It does **not** mean there's a 95% chance the true value is in the

interval for a single sample—once the interval is calculated, the true value either is or isn't in it.

Example:

Suppose you survey 100 people and find an average satisfaction score of 4.2 with a standard deviation of 0.5. A 95% confidence interval for the mean might be:

$$4.2 \pm 1.96 \left(\frac{0.5}{\sqrt{100}} \right) = 4.2 \pm 0.098$$

So the interval is (4.102, 4.298), meaning you're 95% confident the true average satisfaction lies within that range.

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

Answer:

The **expected value** (often denoted as E(X)) is the theoretical average of a random variable. It represents the value you would expect to observe on average if an experiment were repeated many times.

Mathematical Definition

For a **discrete random variable** X with possible values x_1, x_2, ...,

x_n and corresponding probabilities p_1, p_2, ..., p_n:

$$E(X) = \sum_{i=1}^{n} x_i \cdot p_i$$

For a **continuous random variable** with probability density function f(x):

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) \, dx$$

Intuition

The expected value is a **weighted average** of all possible outcomes, where each outcome is weighted by its probability. It does not necessarily correspond to a value the variable can actually take, but it gives a sense of the "center" of the distribution.

Example: Rolling a Fair Die

Let X be the outcome of rolling a fair six-sided die. Each outcome from 1 to 6 has a probability of $\frac{1}{6}$:

$$E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

Although you can never roll a 3.5, it represents the average result over many rolls.

Practical Application

Suppose a game costs ₹100 to play, and the payouts are:

- ₹500 with probability 0.1
- ₹200 with probability 0.2
- ₹0 with probability 0.7

Then the expected winnings are:

E(X) = (500 \times 0.1) + (200 \times 0.2) + (0 \times 0.7) = 50 + 40 + 0 = ₹90

Your expected profit is ₹90 − ₹100 = ₹−10, meaning you would lose ₹10 on average per game.

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.

Answer:

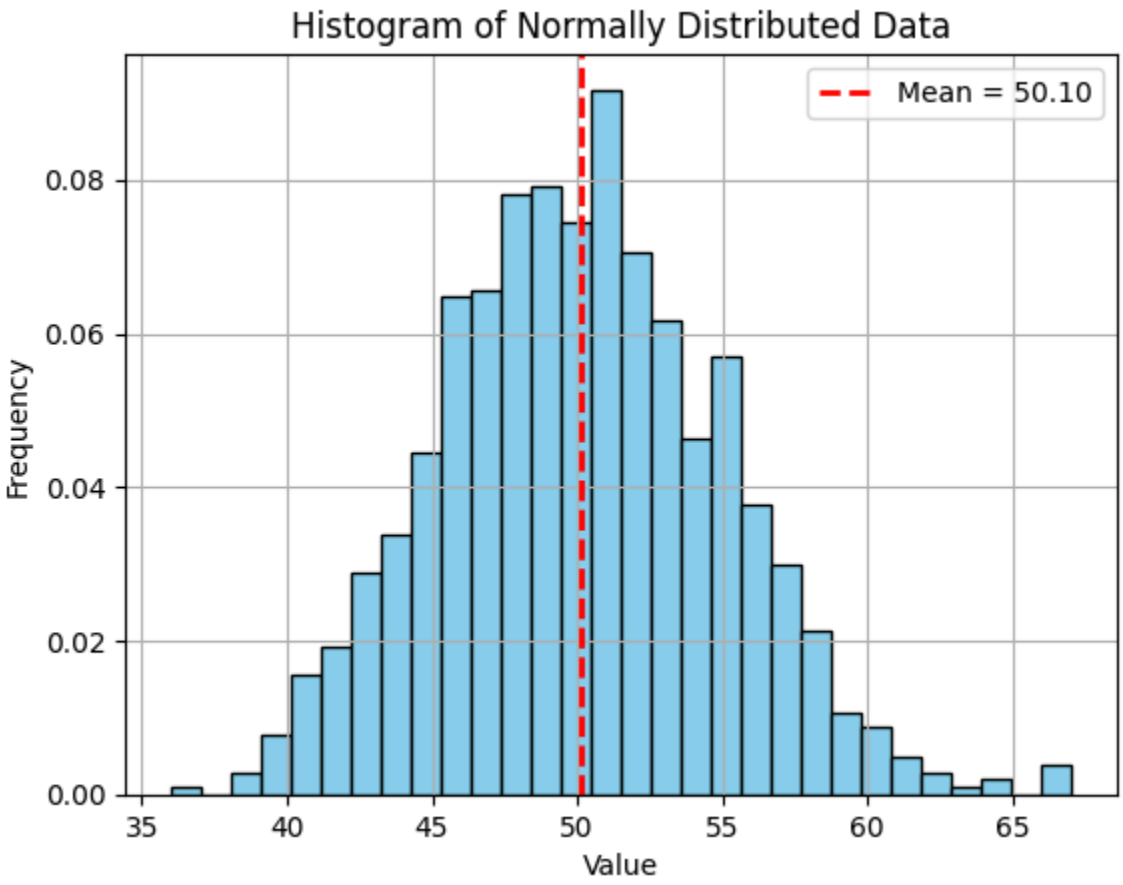
Here's a complete Python program that uses NumPy and Matplotlib to generate 1000 random numbers from a normal distribution with a mean of 50 and standard deviation of 5, computes the sample mean and standard deviation, and visualizes the distribution with a histogram:

Code:

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

# Set parameters
mean = 50
std_dev = 5
sample_size = 1000

# Generate random numbers from a normal distribution
data = np.random.normal(loc=mean, scale=std_dev, size=sample_size)

# Compute sample mean and standard deviation
computed_mean = np.mean(data)
computed_std = np.std(data)

# Display results
print(f"Sample Mean: {computed_mean:.2f}")
print(f"Sample Standard Deviation: {computed_std:.2f}")

# Plot histogram
plt.hist(data, bins=30, color='skyblue', edgecolor='black',
density=True)
plt.title('Histogram of Normally Distributed Data')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.axvline(computed_mean, color='red', linestyle='dashed', linewidth=2,
label=f'Mean = {computed_mean:.2f}')
plt.legend()
plt.grid(True)
plt.show()
```

Output:

```
Sample Mean: 50.10
Sample Standard Deviation: 4.83
```



What This Program Does:

- Uses `np.random.normal()` to generate 1000 values from a normal distribution.
- Computes the sample mean and standard deviation using `np.mean()` and `np.std()`.
- Plots a histogram with 30 bins to visualize the distribution.
- Adds a red dashed line to indicate the computed mean

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.

Answer:

Step-by-Step Explanation:

1. **Central Limit Theorem (CLT) Overview**

The CLT states that the sampling distribution of the sample mean will be approximately normal if:

- The sample size is sufficiently large (typically n \geq 30), or
- The population itself is normally distributed.

Even though your sample size is small (only 20 days), if the underlying sales data is roughly normal, CLT can still be applied cautiously.

2. **Estimate the Population Mean**

Use the sample mean as an estimate of the population mean.

3. **Compute the Standard Error (SE)**

Standard Error is calculated as:

SE = \frac{s}{\sqrt{n}}

Where:

- s is the sample standard deviation
- n is the sample size

4. **Determine the Critical Value**

For a 95% confidence level, the critical value from the **t-distribution** (since n < 30) is used. Degrees of freedom = n - 1.

5. **Compute the Confidence Interval**

\text{CI} = \bar{x} \pm t^* \cdot SE

```python
import numpy as np
import scipy.stats as stats

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

# Convert to NumPy array
sales_array = np.array(daily_sales)

# Sample statistics
sample_mean = np.mean(sales_array)
sample_std = np.std(sales_array, ddof=1)  # Use ddof=1 for sample
standard deviation
n = len(sales_array)

# Standard error
standard_error = sample_std / np.sqrt(n)

# t-critical value for 95% confidence interval
confidence_level = 0.95
degrees_freedom = n - 1
t_critical = stats.t.ppf((1 + confidence_level) / 2, df=degrees_freedom)

# Margin of error
margin_of_error = t_critical * standard_error

# Confidence interval
ci_lower = sample_mean - margin_of_error
ci_upper = sample_mean + margin_of_error

# Output results
print(f"Sample Mean Sales: {sample_mean:.2f}")
print(f"95% Confidence Interval: ({ci_lower:.2f}, {ci_upper:.2f})")
```

Output:

```
Sample Mean Sales: 248.25
95% Confidence Interval: (240.17, 256
```

##