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:

- **Formal Definition**

A **random variable (RV)** is a function

X:S→RX: S \rightarrow \mathbb{R}X:S→R

where:

- SSS is the **sample space** (set of all possible outcomes)
- R\mathbb{R}R is the set of **real numbers**

So, instead of working directly with outcomes like *Heads* or *Tails*, we work with numbers like *0* and *1*.

- **Example 1: Coin Toss**

Experiment: Toss a coin
Sample Space:

S={H,T}S = \{H, T\}S={H,T}

Define a random variable XXX:

- X(H)=1X(H) = 1X(H)=1
- X(T)=0X(T) = 0X(T)=0

Here, XXX converts outcomes into numbers.

🔹 **Types of Random Variables**

1️⃣ **Discrete Random Variable**

- Takes **countable values**
- Example: Number of heads in 3 coin tosses
  Possible values: {0,1,2,3}\{0,1,2,3\}{0,1,2,3}

2️⃣ **Continuous Random Variable**

- Takes **uncountable values** within an interval
- Example: Height of students, time taken to complete a task

Question 2: What are the types of random variables?

Answer:

In probability theory, **random variables** are mainly classified based on the values they can take. The two primary types are:

1️⃣ Discrete Random Variable

A **discrete random variable** takes **countable** values (finite or countably infinite).

Characteristics:

- Values can be listed
- Usually associated with **counting**
- Probability is given by a **Probability Mass Function (PMF)**

Examples:

- Number of heads when tossing 3 coins → {0, 1, 2, 3}
- Number of students in a class
- Number of defective items in a batch

2️⃣ Continuous Random Variable

A **continuous random variable** takes **uncountable** values within an interval.

Characteristics:

- Values are measured, not counted
- Can take infinitely many values in a range
- Probability is given by a **Probability Density Function (PDF)**

Examples:

- Height of students
- Time taken to complete a task
- Temperature of a city

3️⃣ Mixed Random Variable (Less Common)

A **mixed random variable** has **both discrete and continuous components**.

Example:

- A machine that either fails (discrete event) or runs for a continuous amount of time before failing

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

Answer:

1. Discrete Distribution

A **discrete distribution** deals with **countable values** (finite or countably infinite).

Key Characteristics

- Values are **separate and distinct**
- Can be **counted** (0, 1, 2, 3, …)
- Probability is given by a **Probability Mass Function (PMF)**
- Sum of all probabilities = 1

Example

- Number of students in a class
- Number of defective items in a batch
- Tossing a coin (0 heads, 1 head, 2 heads)

Example Distributions

- **Bernoulli Distribution**
- **Binomial Distribution**
- **Poisson Distribution**
- **Geometric Distribution**

Example

$P(X=x)=p(x)P(X = x) = p(x)P(X=x)=p(x)$

2. Continuous Distribution

A **continuous distribution** deals with **uncountable values** that lie within an interval.

Key Characteristics

- Values are **continuous** (can take any real value in a range)
- Cannot be counted individually
- Probability is given by a **Probability Density Function (PDF)**
- Probability at a single point is **zero**
- Area under the curve = 1

Example

- Height of a person
- Time taken to complete a task
- Temperature of a city

Example Distributions

- **Normal (Gaussian) Distribution**
- **Uniform Distribution**
- **Exponential Distribution**
- **Gamma Distribution**

Example

P(a≤X≤b)=∫abf(x) dxP(a \le X \le b) = \int_a^b f(x)\,dxP(a≤X≤b)=∫abf(x)dx

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

Answer:

📌 **What is a Binomial Distribution?**

A random variable **X** follows a binomial distribution if:

1. The experiment has a **fixed number of trials** nnn

2. Each trial has **only two possible outcomes**
   → *success* or *failure*

3. The **probability of success** ppp is the same for every trial

4. Trials are **independent**

It is written as:

X~Binomial(n,p)X \sim \text{Binomial}(n, p)X~Binomial(n,p)

📐 Probability Formula

The probability of getting exactly **k successes** in **n trials** is:

P(X=k)=(nk)pk(1−p)n−kP(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}P(X=k)=(kn)pk(1−p)n−k

Where:

- (nk)=n!k!(n−k)!\binom{n}{k} = \frac{n!}{k!(n-k)!}(kn)=k!(n−k)!n!
- ppp = probability of success
- 1−p1-p1−p = probability of failure

📊 Mean and Variance

- **Mean (Expected Value):**

E(X)=npE(X) = npE(X)=np

- **Variance:**

Var(X)=np(1−p)\text{Var}(X) = np(1 - p)Var(X)=np(1−p)

- **Standard Deviation:**

np(1−p)\sqrt{np(1-p)}np(1−p)

🎯 **Example**

Example:
A coin is tossed 10 times. What is the probability of getting exactly 6 heads?

- n=10n = 10n=10
- p=0.5p = 0.5p=0.5
- k=6k = 6k=6

P(X=6)=(106)(0.5)6(0.5)4=0.205P(X = 6) = \binom{10}{6} (0.5)^6 (0.5)^4 = 0.205P(X=6)=(610)(0.5)6(0.5)4=0.205

🧠 **How is Binomial Distribution Used?**

Binomial distribution is used when we want to model **yes/no outcomes**, such as:

- Number of **heads** in coin tosses
- Number of **defective items** in a batch
- Number of **students passing** an exam
- Number of **successful sales calls**
- **Medical trials** (success/failure of treatment)

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

Answer:

Standard Normal Distribution

The **standard normal distribution** is a special case of the normal (Gaussian) distribution with:

- **Mean (μ) = 0**
- **Standard deviation (σ) = 1**

It is denoted as **Z ~ N(0, 1)**.

Its probability density function (PDF) is:

f(z)=12πe−z22f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}f(z)=2π1e−2z2

The graph is a **bell-shaped curve**, symmetric about 0, where:

- Most values lie close to the mean
- The total area under the curve equals 1

Why Is the Standard Normal Distribution Important?

1. Simplifies Probability Calculations

Any normal distribution can be converted into the standard normal distribution using **z-scores**:

z=x−μσz = \frac{x - \mu}{\sigma}z=σx−μ

This allows us to use **standard normal tables (Z-tables)** to find probabilities easily.

2. Universal Reference Scale

Because all normal distributions can be standardized, the standard normal distribution acts as a **common reference** for:

- Exam scores
- Heights and weights
- Measurement errors

- Statistical tests

3. Foundation of Statistical Inference

It plays a key role in:

- Confidence intervals
- Hypothesis testing
- Regression analysis
- Central Limit Theorem (CLT)

4. Helps Compare Different Data Sets

Z-scores help compare values from different distributions:

- Example: Comparing marks from two different exams

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 important results in statistics. It states that:

> **If you take a sufficiently large random sample from any population (with finite mean and variance), the distribution of the sample mean will be approximately normal (bell-shaped), regardless of the original population's distribution.**

Key Points of CLT

- The population **does not need to be normally distributed**.
- The sample size should be **large enough** (commonly n≥30n \ge 30n≥30).
- The **mean of the sampling distribution** equals the population mean: μx¯=μ\mu_{\bar{x}} = \muμx¯=μ
- The **standard deviation of the sampling distribution** (standard error) is: σx¯=σn\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}σx¯=nσ

Why is CLT Critical in Statistics?

1. Enables Normal Approximation

Even if the original data is skewed or non-normal, sample means tend to follow a **normal distribution**, allowing us to use normal probability methods.

2. Foundation of Inferential Statistics

CLT forms the basis for:

- Confidence intervals
- Hypothesis testing
- z-tests and t-tests

3. Simplifies Real-World Analysis

Real-world data is often unknown or complex. CLT allows statisticians to:

- Make reliable inferences
- Estimate population parameters using samples

4. Works Across Disciplines

CLT is widely used in:

- Data science

- Machine learning
- Economics
- Engineering
- Social sciences

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

Answer:

What is a Confidence Interval?

A confidence interval is a **range of values** derived from sample data that is likely to contain the **true population parameter** (such as a mean or proportion) with a specified level of confidence (e.g., 95%).

> A 95% confidence interval means that if we repeated the sampling process many times, **95% of the calculated intervals would contain the true parameter**.

Significance of Confidence Intervals

1. Measure of Uncertainty

- A point estimate (like a sample mean) gives only one value.
- A confidence interval shows **how precise or uncertain** that estimate is.
- Narrow CI → high precision

- Wide CI → more uncertainty

2. Better Than Point Estimates Alone

Instead of saying:

> "The average score is 72"

You can say:

> "The average score is between 68 and 76 (95% CI)"

This provides **more reliable information**.

3. Helps in Decision Making

- Used in **medicine, business, economics, and engineering**.
- Helps assess whether an effect is **practically meaningful**, not just statistically significant.

Example:

- If a CI for a drug's effect does not include 0, the drug likely has a real effect.

4. Relationship with Hypothesis Testing

- Confidence intervals can be used to **test hypotheses**.
- If the null value (e.g., 0 difference or population mean) is **outside the CI**, the result is statistically significant at that confidence level.

5. Indicates Sample Size Adequacy

- Large sample size → narrower confidence interval
- Small sample size → wider confidence interval

Thus, CIs reflect **data quality and reliability**.

6. Applicable Across Many Statistics

Confidence intervals are used for:

- Means
- Proportions
- Regression coefficients
- Differences between groups

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

Answer:

Concept of Expected Value in a Probability Distribution

The **expected value (EV)**, also called the **mean** or **mathematical expectation**, represents the **long-run average outcome** of a random variable if an experiment is repeated many times.

In simple terms, it is the **weighted average of all possible values**, where the weights are their probabilities.

Formula

1. Discrete Random Variable

If XXX takes values x1,x2,...,xnx_1, x_2, ..., x_nx1,x2,...,xn with probabilities P(x1),P(x2),...,P(xn)P(x_1), P(x_2), ..., P(x_n)P(x1),P(x2),...,P(xn):

E(X)=∑xi⋅P(xi)E(X) = \sum x_i \cdot P(x_i)E(X)=∑xi⋅P(xi)

2. Continuous Random Variable

If XXX has probability density function f(x)f(x)f(x):

E(X)=∫−∞∞x⋅f(x) dxE(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dxE(X)=∫−∞∞x⋅f(x)dx

Why Expected Value Is Important

1. Measures Central Tendency

- It is the **mean** of the probability distribution.
- Gives a single representative value.

2. Used in Decision Making

- Widely used in **economics, finance, and game theory**.
- Helps evaluate risks and rewards.

Example:

- If EV > 0 → favorable decision
- If EV < 0 → unfavorable decision

3. Foundation for Other Concepts

Expected value is the basis for:

- Variance
- Standard deviation
- Law of Large Numbers
- Central Limit Theorem

4. Applies to Functions of Random Variables

If $Y=g(X)Y = g(X)Y=g(X)$, then:

$E(Y)=E[g(X)]E(Y) = E[g(X)]E(Y)=E[g(X)]$

This is useful in advanced statistical modeling.

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:

Here is a **complete Python program** that does exactly what you asked using **NumPy** and **Matplotlib**.

Steps covered:

- Generate 1000 random numbers from a normal distribution
- Compute mean and standard deviation
- Plot a histogram to visualize the distribution

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

# Set parameters
mean = 50
std_dev = 5
num_samples = 1000
```

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

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

print("Computed Mean:", computed_mean)
print("Computed Standard Deviation:", computed_std)

# Plot histogram
plt.hist(data, bins=30)
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.title("Histogram of Normally Distributed Data (Mean=50, Std=5)")
plt.show()
```

Explanation (for exams/interviews)

- np.random.normal() generates data following a normal distribution
- np.mean() and np.std() compute sample statistics
- Histogram shows the **bell-shaped curve**, confirming normality

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:

1️⃣ Applying the Central Limit Theorem (CLT)

Even though you have daily sales data, individual daily sales may not follow a normal distribution. The **Central Limit Theorem (CLT)** helps us here.

How CLT applies:

- According to CLT, **if the sample size is sufficiently large**, the **sampling distribution of the sample mean** will be approximately **normal**, regardless of the original data distribution.
- In practice, a sample size of **n ≥ 30** is ideal, but CLT can still be reasonably applied for smaller samples if the data is not extremely skewed.
- This allows us to:
    - Treat the mean sales as normally distributed
    - Estimate the **population mean sales**
    - Construct a **95% confidence interval**

95% Confidence Interval Formula:

CI=x¯±Z×sn\text{CI} = \bar{x} \pm Z \times \frac{s}{\sqrt{n}}CI=x¯±Z×ns

Where:

- $\bar{x}$ = sample mean
- s = sample standard deviation
- n = number of observations
- $Z = 1.96$ for 95% confidence level

Interpretation:

> We are **95% confident** that the true average daily sales lies within this interval.

2️⃣ Python Code to Compute Mean and 95% Confidence Interval

```
import numpy as np
from scipy import 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 = np.array(daily_sales)

# Sample statistics
mean_sales = np.mean(sales)
std_dev = np.std(sales, ddof=1)   # sample standard deviation
n = len(sales)

# Z-score for 95% confidence
z = 1.96

# Margin of error
margin_of_error = z * (std_dev / np.sqrt(n))

# Confidence interval
lower_bound = mean_sales - margin_of_error
upper_bound = mean_sales + margin_of_error
```

```
print("Mean Daily Sales:", mean_sales)
print("95% Confidence Interval:", (lower_bound, upper_bound))
```

3️⃣ Business Insight

- The **mean** represents the typical daily sales.
- The **confidence interval** provides a reliable range for planning inventory, forecasting revenue, and setting targets.
- Management can be confident that the true average sales falls within this range **95% of the time**.