Assignment Code: DA-AG-006

Statistics Advanced - 1| Assignment

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

In probability theory, a random variable is a variable whose value depends on the outcome of a random experiment. It is used to assign a numerical value to each possible outcome of a random event. There are two main types of random variables: (1) a discrete random variable, which takes countable values such as the number of heads in three coin tosses, and (2) a continuous random variable, which can take any value within a given range, such as height, weight, or temperature. For example, when a coin is tossed, if we define X = 1 when the result is a head and X = 0 when the result is a tail, then X is a random variable because its value depends on the random outcome of the coin toss.

Question 2: What are the types of random variables?

There are mainly two types of random variables in probability theory: discrete random variables and continuous random variables. A discrete random variable can take only specific, countable values such as 0, 1, 2, 3, etc. Examples include the number of heads obtained when tossing a coin three times or the number of students present in a class. On the other hand, a continuous random variable can take any value within a certain range or interval. Its values are uncountable and often represent measurements like height, weight, time, or temperature. Thus, the key difference lies in whether the possible outcomes are countable (discrete) or uncountable (continuous).

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

The main difference between discrete and continuous distributions lies in the type of values their random variables can take. A discrete distribution represents situations where the random variable can take only specific, countable values. Examples include the number of students in a class or the number of heads obtained in coin tosses. The probabilities in a discrete distribution are described using a probability mass function (PMF).

In contrast, a continuous distribution represents situations where the random variable can take any value within a continuous range or interval. Examples include height, weight, or temperature. Probabilities in continuous distributions are represented using a probability density function (PDF), and the probability of the variable taking an exact value is zero; instead, probabilities are

calculated over intervals. Thus, discrete distributions deal with countable outcomes, while continuous distributions deal with uncountable, measurable outcomes**.**

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

A binomial distribution is a type of discrete probability distribution that describes the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes — success or failure. It is used when the probability of success (p) remains the same in every trial.

Mathematically, it is represented as: $P(X=k)=nCk\times p^{k}\times(1-p)^{n-k}$,
where n is the number of trials, k is the number of successes, p is the probability of success, and (1 − p) is the probability of failure.

The binomial distribution is widely used in probability to model situations such as flipping a coin multiple times, checking defective products in a batch, or predicting the number of students passing an exam. It helps in understanding how likely it is to get a certain number of successes in repeated independent experiments.

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

The standard normal distribution is a special case of the normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1. It is a continuous probability distribution that is symmetric about the mean and has a bell-shaped curve.

It is important because it serves as a reference distribution in statistics. Any normal distribution can be converted into a standard normal distribution using the z-score formula: $z = (X - \mu) / \sigma$,
where X is the data value, μ is the mean, and σ is the standard deviation.

The standard normal distribution is widely used in hypothesis testing, confidence intervals, and probability calculations, as it allows statisticians to compare different datasets on a common scale and easily find probabilities using standard normal tables (Z-tables).

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

The Central Limit Theorem (CLT) states that when a large number of independent random samples are taken from any population with a finite mean and variance, the sampling distribution

of the sample mean tends to become approximately normal, regardless of the shape of the original population distribution.

In simple terms, it means that as the sample size increases, the distribution of sample means approaches a normal distribution with mean μ (the population mean) and standard deviation σ/√n (where σ is the population standard deviation and n is the sample size).

The CLT is critical in statistics because it allows researchers to use normal distribution-based methods—like confidence intervals and hypothesis tests—even when the population itself is not normally distributed. It provides the theoretical foundation for making inferences about populations from sample data.

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

A confidence interval is a range of values used to estimate an unknown population parameter (like the mean or proportion) with a certain level of confidence. It provides both an estimate and a measure of uncertainty around that estimate.

The significance of confidence intervals in statistical analysis lies in the fact that they show how reliable an estimate is. For example, a 95% confidence interval means that if the same experiment were repeated many times, approximately 95% of the calculated intervals would contain the true population parameter.

Confidence intervals are important because they give more information than a single point estimate — they indicate the precision of the estimate and help in decision-making by showing the possible range of true values. They are widely used in research, surveys, and data analysis to assess the reliability of results.

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

The expected value in a probability distribution represents the average or mean outcome that one can expect from a random experiment if it were repeated many times. It gives a measure of the center of the probability distribution.

For a discrete random variable, the expected value (E[X]) is calculated as:
$E[X] = \Sigma [x \times P(x)]$,
where x represents each possible value of the random variable, and $P(x)$ is the probability of that value.

For a continuous random variable, it is calculated using an integral: $E[X] = \int x \times f(x)\, dx$, where *f(x)* is the probability density function.

The expected value is important because it helps in predicting long-term averages and making informed decisions in fields like economics, finance, and risk analysis. For example, in tossing a fair coin, if you win ₹10 for heads and lose ₹10 for tails, the expected value is ₹0 — meaning no gain or loss on average over many tosses.

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.)

Python program to generate 1000 random numbers from a normal distribution with mean 50 and standard deviation 5, compute the mean and standard deviation, and plot a histogram using **NumPy** and **Matplotlib**:

```
# Import necessary libraries

import numpy as np

import matplotlib.pyplot as plt


# Set parameters

mean = 50

std_dev = 5

num_samples = 1000


# Generate 1000 random numbers from normal distribution

data = np.random.normal(loc=mean, scale=std_dev, size=num_samples)
```

```
# Compute mean and standard deviation
calculated_mean = np.mean(data)
calculated_std = np.std(data)

print(f"Calculated Mean: {calculated_mean}")
print(f"Calculated Standard Deviation: {calculated_std}")

# Plot histogram
plt.hist(data, bins=30, color='skyblue', edgecolor='black')
plt.title("Histogram of Normal Distribution (mean=50, std=5)")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.show()
```

Explanation:

- np.random.normal(loc=mean, scale=std_dev, size=num_samples) generates 1000 random numbers from a normal distribution.
- np.mean() and np.std() calculate the mean and standard deviation of the generated data.
- plt.hist() draws a histogram to visualize the distribution.

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.)

Explanation using Central Limit Theorem (CLT):

The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean will be approximately normal, even if the underlying data is not perfectly normal, provided the sample size is reasonably large.

To estimate the average daily sales with a 95% confidence interval:

1. Compute the sample mean ($\bar{x}$) of the daily sales.
2. Compute the sample standard deviation (s).
3. Calculate the standard error (SE):
   $$SE = \frac{s}{\sqrt{n}}$$
   where *n* is the number of observations.
4. For a 95% confidence interval, the z-score is approximately 1.96.
5. Compute the confidence interval:
   $$CI = \bar{x} \pm (z \times SE)$$

Python Code:

```
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]


# Sample size

n = len(daily_sales)
```

```
# Sample mean and standard deviation
mean_sales = np.mean(daily_sales)
std_sales = np.std(daily_sales, ddof=1)  # ddof=1 for sample std deviation

# Standard error
se = std_sales / np.sqrt(n)

# 95% confidence interval
confidence_level = 0.95
z_score = stats.norm.ppf(0.975)  # two-tailed 95% CI
ci_lower = mean_sales - z_score * se
ci_upper = mean_sales + z_score * se

print(f"Mean Sales: {mean_sales}")
print(f"95% Confidence Interval: ({ci_lower}, {ci_upper})")
```

Explanation of the Code:

- np.mean() calculates the sample mean.
- np.std(ddof=1) calculates the sample standard deviation.
- stats.norm.ppf(0.975) gives the z-score for a 95% confidence interval (two-tailed).
- The confidence interval is then calculated using the formula mean ± z * SE.

This method allows you to estimate the average daily sales and understand the range in which the true mean likely lies with 95% confidence.