Statistics Advanced - 1| Assignment

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

Answer : A random variable is a variable that takes values from the possible outcomes of a random experiment.

Example: In tossing a coin, if we define Heads = 1 and Tails = 0, then the random variable can take the values {0, 1}

Question 2: What are the types of random variables?

Answer : Types of random variables:

1. Discrete Random Variable

    - It has countable values or it can be a whole number.

    Example: rolling a dice

2. Continuous Random Variable

    - It can take any value within a range.

    Example: The height of people → it can be 150.2 cm, 150.25 cm, etc.

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

Answer : the difference between discrete and continuous distribution are

Discrete Distribution

- It has countable outcomes.
- Values are specific like 0, 1, 2, 3…
- Example: Number of students in a class or number of emails received.

Continuous Distribution

- It has outcomes that can take any value within a range .
- Values are infinite between two points.
- Example: Height, weight, time.

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

Binomial Distribution

- It's a type of discrete distribution.
- It shows the probability of getting a certain number of successes in a fixed number of trials.
- Each trial has only two outcomes: success or failure.
- Example: Tossing a coin 5 times and counting how many heads appear.

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

Answer :

Standard Normal Distribution

- It’s a special type of continuous distribution.
- It is a normal distribution with a mean of 0 and a standard deviation of 1.
- The graph is a bell-shaped curve, symmetric around 0.

It is important because many of machine learning algorithms like regression ,logistic regression, clustering requires scaling.

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

Answer : central limit theorem states that if you have a population with mean and standard deviation take sufficiently large number of random from the population with replacement then the distribution of sample mean will be approximately normally distributed.

It is critical because it helps us understand and work with data easily. Even if the original data is not normal, the sample means will follow a normal pattern when the sample size is large. This allows us to use the normal distribution to make predictions, test ideas, and calculate probabilities in many real-life situations.

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

Answer : Significance of Confidence Intervals in Statistical Analysis are

- Confidence intervals (CI) show the range of values where the true population parameter (like the mean) is likely to be.

- It gives an idea of how reliable or uncertain your estimate is.

- A wider interval means more uncertainty, while a narrower interval means more confidence.

- It helps in making decisions and understanding the accuracy of results from sample data.

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

Answer : Expected Value in a Probability Distribution

- It is the average or mean value you expect from a random experiment.
- It’s like the long-term result if you repeat the experiment many times.
- Calculated by multiplying each possible outcome by its probability and then adding them all together.

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.)

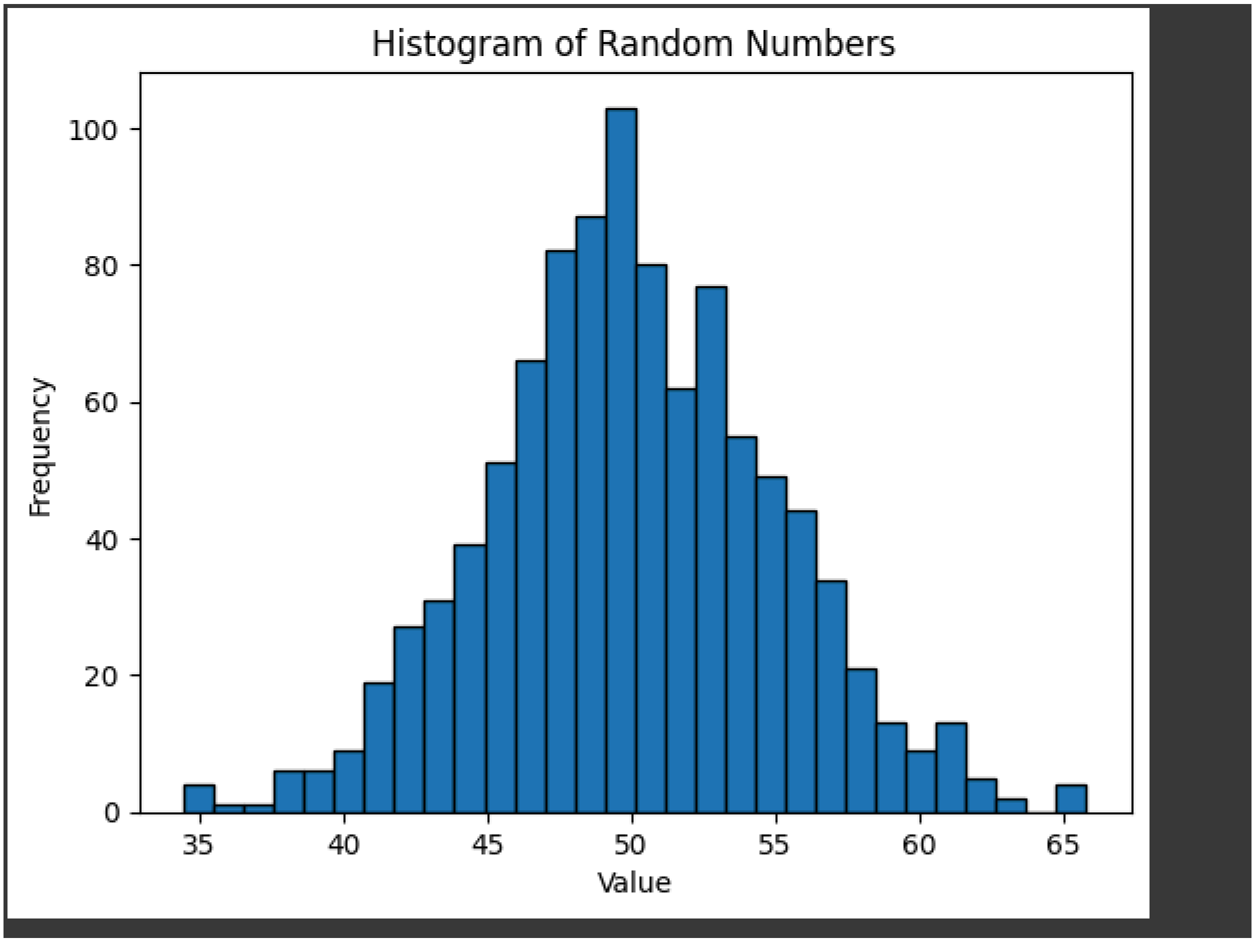
Answer :

Input :

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

mean = 50
std_dev = 5
random_numbers = np.random.normal(mean, std_dev, 1000)

calculated_mean = np.mean(random_numbers)
calculated_std = np.std(random_numbers)

print("Mean of generated data:", calculated_mean)
print("Standard deviation of generated data:", calculated_std)

plt.hist(random_numbers, bins=30, edgecolor='black')
plt.title("Histogram of Random Numbers")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.show()
```

Output :

```
Mean of generated data: 50.05558416816697
Standard deviation of generated data: 4.982419146573761
```



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.)

Input :

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

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_dev = np.std(daily_sales, ddof=1)
n = len(daily_sales)
std_error = std_dev / np.sqrt(n)

confidence_level = 0.95
confidence_interval = stats.t.interval(confidence_level, n-1,
loc=mean_sales, scale=std_error)

print("Mean of daily sales:", mean_sales)
print("95% confidence interval for the mean:", confidence_interval)
```

Output :

```
Mean of daily sales: 248.25
95% confidence interval for the mean: (np.float64(240.16957025147158),
np.float64(256.3304297485284))
```