Statistics Advanced - 1| Assignment

Questions:-

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

Answer: In probability, a random variable is a numerical quantity whose value is determined by the outcome of a random phenomenon.

Question 2: What are the types of random variables?

Answer: Random variables are broadly classified into two main categories: discrete and continuous. A discrete random variable can take on a countable number of values, while a continuous random variable can take on any value within a given range.

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

Answer: Discrete and continuous distributions differ in the type of data they represent. Discrete distributions deal with data that can only take on specific, separate values (like integers), while continuous distributions handle data that can fall anywhere within a given range (including decimals and fractions).

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

Answer: A binomial distribution models the probability of achieving a specific number of successes in a fixed number of independent trials, where each trial has only two possible outcomes (success or failure). It's a fundamental concept in probability used to analyze scenarios like coin flips, product testing, or survey responses.

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

Answer: The standard normal distribution is a specific type of normal distribution with a mean of 0 and a standard deviation of 1. It's essentially a normalized version of any normal distribution, allowing for easier comparisons and calculations. It's crucial because it simplifies statistical analysis, provides a foundation for many statistical methods, and is linked to the central limit theorem.

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

Answer: The Central Limit Theorem (CLT) states that the distribution of sample means will approximate a normal distribution, regardless of the original population's distribution, as long as the sample size is sufficiently large. This is critical in statistics because it allows us to apply statistical methods that assume normality, even when the underlying population data is not normally distributed.

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

Answer: Confidence intervals in statistical analysis provide a range of plausible values for an unknown population parameter, quantifying the uncertainty associated with sample estimates. They help researchers understand the reliability of their findings and make more informed decisions by indicating how much the results might vary if the study were repeated.

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

Answer: The expected value in a probability distribution represents the average outcome of a random variable over many repetitions of an experiment. It's a weighted average, where each possible outcome is multiplied by its probability and then summed. Essentially, it's the long-term average you'd expect if you repeated the same scenario many times.

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.

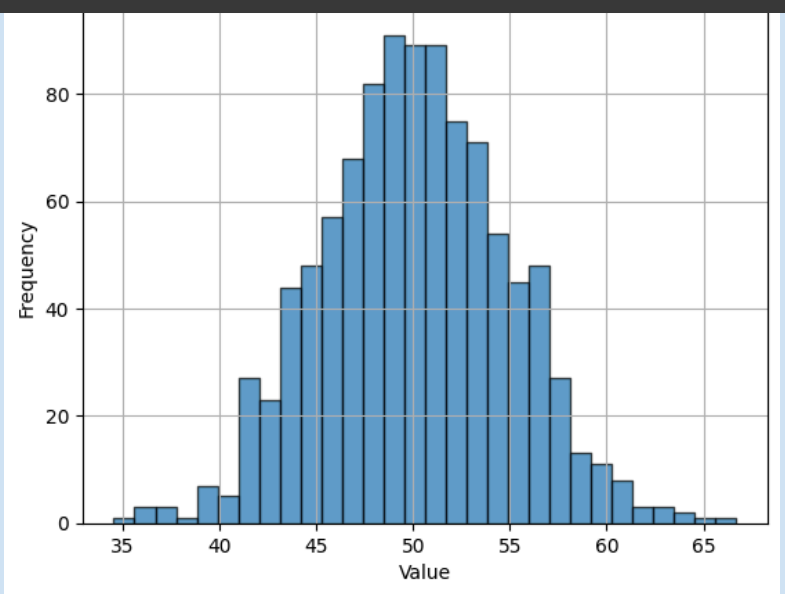
Answer:

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

# Parameters for the normal distribution
mean = 50
std_dev = 5
num_samples = 1000

# Generate random numbers from a normal distribution
# np.random.normal(loc=mean, scale=std_dev, size=num_samples)
# loc is the mean (center) of the distribution.
# scale is the standard deviation (spread) of the distribution.
# size is the number of random samples to generate.
```

```
        random_numbers = np.random.normal(loc=mean, scale=std_dev, s
ize=num_samples)

        # Compute the mean and standard deviation of the generated numbers
        computed_mean = np.mean(random_numbers)
        computed_std_dev = np.std(random_numbers)

        print(f"Generated data mean: {computed_mean:.2f}")
        print(f"Generated data standard deviation: {computed_std_dev:.2f}")

        # Visualize the distribution using a histogram
        plt.hist(random_numbers, bins=30, edgecolor='black', alpha=0.7)
        plt.title('Histogram of Normally Distributed Random Numbers')
        plt.xlabel('Value')
        plt.ylabel('Frequency')
        plt.grid(True)
        plt.show()
```

OUTPUT:

```
Generated data mean: 50.09
Generated data standard deviation: 4.80
```



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: ● Explain how you would apply the Central Limit Theorem to estimate the average sales with a 95% confidence interval.

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

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

# Step 1: Sample statistics
n = len(daily_sales)
mean_sales = np.mean(daily_sales)
std_sales = np.std(daily_sales, ddof=1)  # sample std deviation

# Step 2: Standard Error
SE = std_sales / np.sqrt(n)

# Step 3: t critical value for 95% confidence
t_critical = stats.t.ppf(0.975, df=n-1)

# Step 4: Margin of Error
ME = t_critical * SE

# Step 5: Confidence Interval
lower_bound = mean_sales - ME
upper_bound = mean_sales + ME

print(f"Sample Mean: {mean_sales:.2f}")
print(f"95% Confidence Interval: ({lower_bound:.2f}, {upper_bound:.2f})")
```

OUTPUT:

```
Sample Mean: 248.25
95% Confidence Interval: (240.17, 256.33)
```

- Write the Python code to compute the mean sales and its 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 for easier calculations
sales_array = np.array(daily_sales)

# Calculate mean
mean_sales = np.mean(sales_array)

# Sample size
n = len(sales_array)

# Sample standard deviation (ddof=1 for sample)
std_dev = np.std(sales_array, ddof=1)

# Confidence level
confidence_level = 0.95

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

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

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

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: (240.17, 256.33)
```