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 numerical values to the outcomes of a random experiment. It helps translate outcomes from a sample space into numbers so they can be analyzed mathematically. There are two main types: **discrete random variables**, which take on countable values (like the roll of a die), and **continuous random variables**, which can take on any value within a range (like measuring time or temperature). Despite the name, a random variable itself is not random—it's a rule that maps outcomes to numbers.

Question 2 :

What are the types of random variables?

ANSWER :

There are two main types of random variables in probability theory: **discrete** and **continuous**. A **discrete random variable** takes on a countable number of distinct values, such as the result of rolling a die or the number of heads in a series of coin tosses. These variables are typically associated with situations where outcomes can be listed or counted, and their behavior is described using a **probability mass function (PMF)**. On the other hand, a **continuous random variable** can take on an infinite number of possible values within a given range. Examples include measurements like time, height, or temperature. Continuous variables are described using a **probability density function (PDF)**, and the probability of the variable taking on any exact value is zero—instead, we look at the probability over intervals. Both types of random variables serve to model and analyze different kinds of random phenomena in quantitative terms.

Question 3 :

Explain the difference between discrete and continuous distributions.

ANSWER :

The key difference between discrete and continuous distributions lies in the type of values they represent and how probabilities are assigned. A **discrete distribution** deals with **countable values**, such as the outcome of rolling a die, where each specific value has a non-zero probability, described by a **probability mass function (PMF)**. In contrast, a **continuous distribution** involves **uncountable values** over intervals, like measuring height or time. Here, the probability of any exact value is zero, and probabilities are calculated over ranges using a **probability density function (PDF)**.

Question 4 :

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

ANSWER :

A **binomial distribution** is a type of discrete probability distribution that models 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 remains constant across trials. For example, it can be used to calculate the probability of getting a certain number of heads in a series of coin tosses. The distribution is defined by two parameters: **n** (number of trials) and **p** (probability of success in each trial).

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 is a symmetric, bell-shaped curve used to model many natural and social phenomena. Its importance lies in its role as a reference distribution in statistics—any normal distribution can be converted into the standard normal distribution using **z-scores**, which makes it easier to calculate probabilities and compare different data sets. It is widely used in hypothesis testing, confidence intervals, and many statistical methods.

Question 6 :

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

ANSWER :

The **Central Limit Theorem (CLT)** states that, regardless of the population's original distribution, the **sampling distribution of the sample mean** will approach a **normal distribution** as the sample size becomes large enough, typically $n \geq 30$. This is critical in statistics because it allows us to use **normal distribution methods** (like z-scores and confidence intervals) to make inferences about population parameters, even when the population itself is not normally distributed. It forms the foundation for much of inferential statistics.

Question 7 :

What is the significance of confidence intervals in statistical analysis?

ANSWER :

Confidence intervals are important in statistical analysis because they provide a **range of values** within which a population parameter (like the mean or proportion) is likely to fall, based on sample data. Instead of giving a single estimate, a confidence interval expresses the **uncertainty or reliability** of that estimate. For example, a 95% confidence interval means we can be 95% confident that the true population value lies within that range. This helps researchers make informed decisions and assess the **precision and credibility** of their results.

Question 8 :

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

ANSWER :

The **expected value** of a probability distribution is the **long-run average or mean value** you would expect to get if you repeated a random experiment many times. It is calculated by multiplying each possible outcome by its probability and summing these products. The expected value gives a measure of the **center or typical outcome** of the distribution, helping to summarize its overall behavior in a single number.

Question 9 :

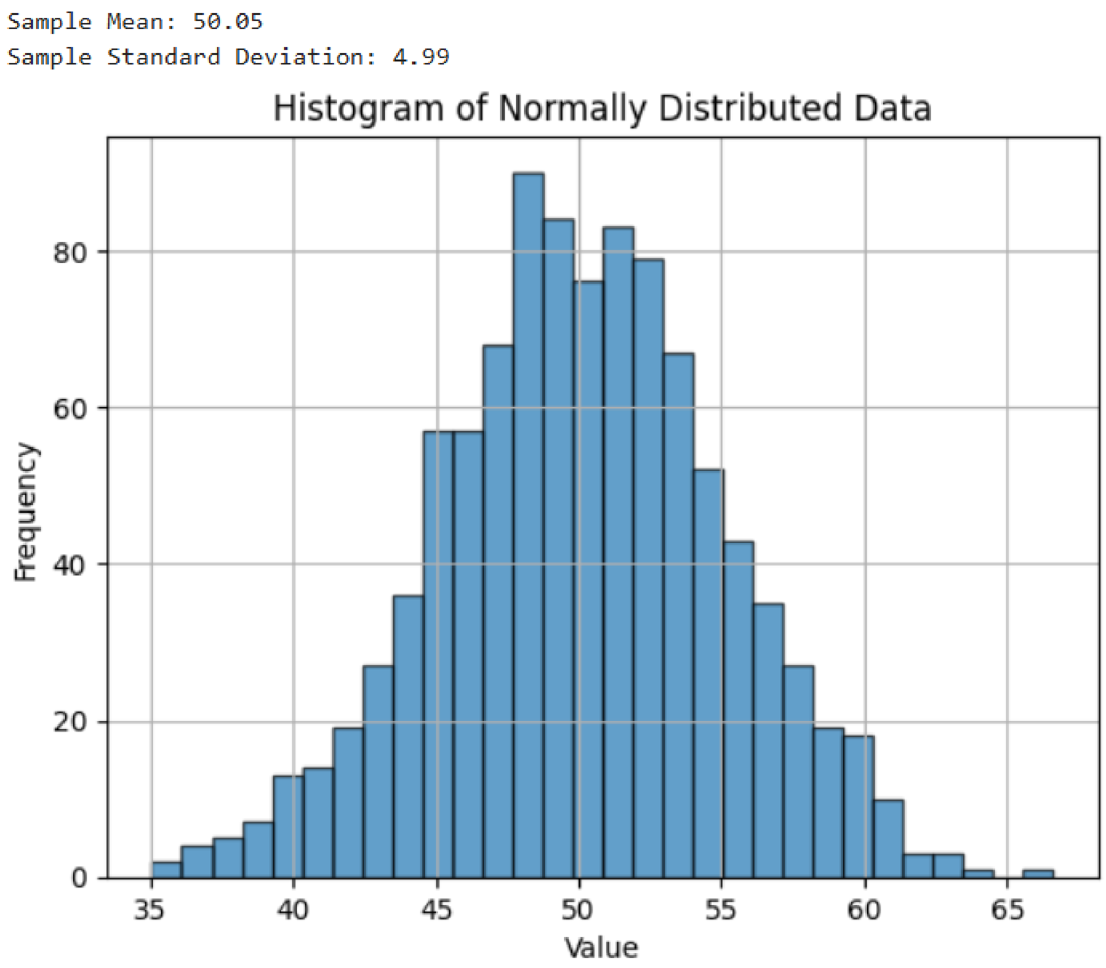
Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution.

ANSWER :

CODE :

```
import numpy as np

import matplotlib.pyplot as plt

# Parameters

mean = 50

std_dev = 5

sample_size = 1000

# Generate random numbers from normal distribution

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

# Compute mean and standard deviation

sample_mean = np.mean(data)

sample_std_dev = np.std(data)

print(f"Sample Mean: {sample_mean:.2f}")

print(f"Sample Standard Deviation: {sample_std_dev:.2f}")

# Plot histogram

plt.hist(data, bins=30, edgecolor='black', alpha=0.7)

plt.title('Histogram of Normally Distributed Data')

plt.xlabel('Value')
plt.ylabel('Frequency')

plt.grid(True)

plt.show()
```

OUTPUT :



```
Sample Mean: 50.05
Sample Standard Deviation: 4.99
```

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 :

Applying the Central Limit Theorem (CLT):

Since you have daily sales data for 20 days, the CLT tells us that if you take many samples of this size from the population, the **distribution of the sample means** will approximate a normal distribution, even if the original data isn't perfectly normal. This allows us to estimate the **average daily sales** (population mean) and construct a **95% confidence interval** around the sample mean. The confidence interval gives a range where we are 95% confident the true average daily sales lie.

CODE :

```
import numpy as np
import scipy.stats as stats
# Given 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
data = np.array(daily_sales)
# Sample size, mean and standard deviation
n = len(data)
mean_sales = np.mean(data)
std_dev = np.std(data, ddof=1)  # Sample standard deviation
# Confidence level and critical t-value
confidence = 0.95
alpha = 1 - confidence
t_crit = stats.t.ppf(1 - alpha/2, df=n-1)
# Margin of error
margin_of_error = t_crit * (std_dev / np.sqrt(n))
# Confidence interval
lower_bound = mean_sales - margin_of_error
upper_bound = mean_sales + margin_of_error
```

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

OUTPUT :

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