**Assignment Code: DA-AG-007**

Statistics Advanced - 2| **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**: 180

**Question 1:** What is hypothesis testing in statistics?

**Answer:**

1. Hypotheses

* Null hypothesis (H₀): A baseline assumption, usually that there is *no effect* or *no difference*.  
  Example: "The average height of students is 170 cm."
* Alternative hypothesis (H₁ or Ha): What you want to test for; it suggests a significant effect or difference.  
  Example: "The average height of students is not 170 cm."

1. Test Statistic  
   A numerical value calculated from the sample data (like *z*, *t*, *chi-square*). It measures how far the sample result is from what the null hypothesis predicts.
2. Significance Level (α)  
   A threshold (commonly 0.05 or 5%) that represents the probability of rejecting H₀ when it is actually true (Type I error).
3. P-value  
   The probability of obtaining test results at least as extreme as the observed data, assuming H₀ is true.

* If p ≤ α → Reject H₀ (evidence supports H₁).
* If p > α → Fail to reject H₀ (insufficient evidence against H₀).

1. Decision  
   Based on the comparison of the p-value (or test statistic) with the chosen significance level.

**Question 2:** What is the null hypothesis, and how does it differ from the alternative hypothesis?

**Answer:** The null hypothesis (H₀) is a statistical assumption that there is no significant effect, difference, or relationship in the population, and it serves as the default or baseline statement in hypothesis testing. It is considered true until sufficient evidence is found against it. In contrast, the alternative hypothesis (H₁ or Ha) represents the claim that researchers aim to support; it suggests that there is a significant effect, difference, or relationship. The main difference between the two is that the null hypothesis reflects the idea of "no change" or "status quo," while the alternative hypothesis proposes a deviation from that assumption. In practice, hypothesis testing is designed to determine whether sample data provides enough evidence to reject the null hypothesis in favor of the alternative.

**Question 3:** Explain the significance level in hypothesis testing and its role in deciding the outcome of a test.

**Answer:**

The significance level (α) in hypothesis testing is the threshold probability used to decide whether to reject the null hypothesis. It represents the maximum risk a researcher is willing to take of making a Type I error, which means rejecting the null hypothesis when it is actually true. Common values of α are 0.05 (5%), 0.01 (1%), or 0.10 (10%). For example, if α = 0.05, it means the researcher accepts a 5% chance of wrongly rejecting a true null hypothesis. The significance level plays a crucial role in decision-making: after calculating the p-value from sample data, the result is compared with α. If the p-value is less than or equal to α, the null hypothesis is rejected in favor of the alternative hypothesis, indicating the result is statistically significant. If the p-value is greater than α, the null hypothesis is not rejected, meaning there is not enough evidence to support the alternative hypothesis. In this way, the significance level acts as a cutoff point that helps determine the strength of evidence required to claim a meaningful result.

**Question 4:** What are Type I and Type II errors? Give examples of each.

**Answer:** In hypothesis testing, errors can occur because decisions are based on sample data, not the entire population. The two main types of errors are Type I error and Type II error.

A Type I error happens when the null hypothesis (H₀) is true but is wrongly rejected. In other words, it is a "false positive." The probability of committing this error is equal to the significance level (α). For example, suppose a medical test is used to check if a patient has a disease. If the test concludes that the patient has the disease when they actually do not, this is a Type I error. Similarly, in a court trial, convicting an innocent person would represent a Type I error.

A Type II error occurs when the null hypothesis is false but is wrongly accepted (or not rejected). This is a "false negative." The probability of this error is denoted by β. For example, if the same medical test concludes that a patient does not have the disease when in fact they do, this is a Type II error. In a court analogy, letting a guilty person go free would be an example of a Type II error.

Thus, a Type I error means finding an effect or difference that does not actually exist, while a Type II error means failing to detect a real effect or difference. Both errors are important considerations in designing and interpreting statistical tests.

**Question 5:** What is the difference between a Z-test and a T-test? Explain when to use each.

**Answer:** The Z-test and T-test are both statistical tests used to determine whether there is a significant difference between sample data and a population parameter or between two sample means. The main difference lies in the conditions under which each test is applied. A Z-test is used when the population variance (or standard deviation) is known and the sample size is large (generally n>30n > 30n>30). It relies on the normal distribution. On the other hand, a T-test is used when the population variance is unknown and must be estimated from the sample, and it is most suitable for small sample sizes (n≤30n \leq 30n≤30). The T-test uses the Student’s t-distribution, which accounts for the additional uncertainty in estimating the population standard deviation.

In summary, the Z-test is appropriate for large samples with known population variance, while the T-test is used for small samples or when the population variance is unknown. For example, if a company wants to test whether the average weight of a large shipment of goods differs from 50 kg and the population variance is known, a Z-test would be appropriate. If only a small sample is available and the population variance is not known, a T-test would be used instead.

**Question 6:** Write a Python program to generate a binomial distribution with n=10 and p=0.5, then plot its histogram.

(*Include your Python code and output in the code box below.*) Hint: Generate random number using random function.

**Answer:**

import numpy as np

import matplotlib.pyplot as plt

# Parameters

n = 10 # number of trials

p = 0.5 # probability of success

size = 1000 # number of random samples

# Generate random numbers from Binomial distribution

data = np.random.binomial(n, p, size)

# Plot histogram

plt.hist(data, bins=range(0, n+2), edgecolor='black', align='left')

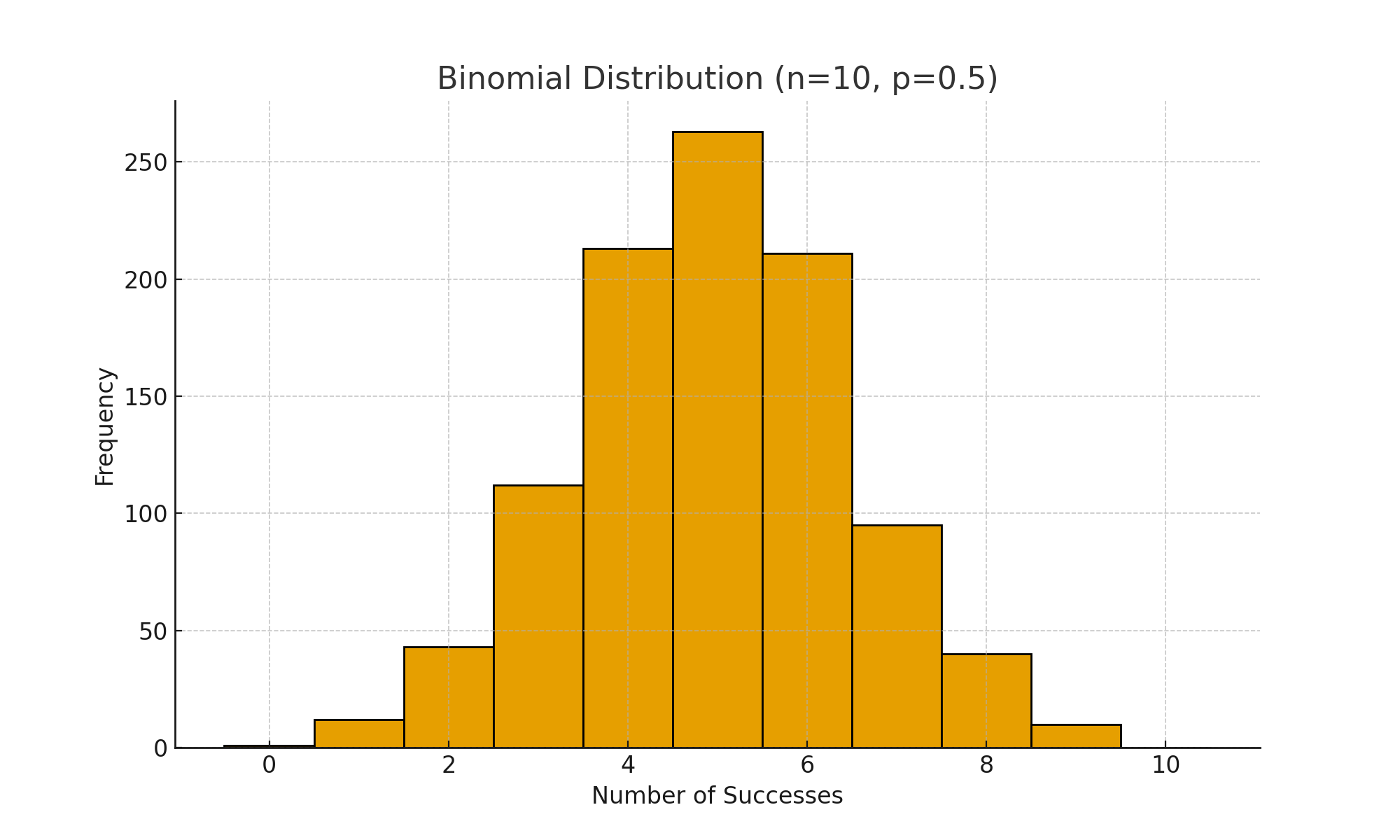
plt.title("Binomial Distribution (n=10, p=0.5)")

plt.xlabel("Number of Successes")

plt.ylabel("Frequency")

plt.show()

OUTPUT :

****

**Question 7**: Implement hypothesis testing using Z-statistics for a sample dataset in Python. Show the Python code and interpret the results.

sample\_data = [49.1, 50.2, 51.0, 48.7, 50.5, 49.8, 50.3, 50.7, 50.2, 49.6,

50.1, 49.9, 50.8, 50.4, 48.9, 50.6, 50.0, 49.7, 50.2, 49.5,

50.1, 50.3, 50.4, 50.5, 50.0, 50.7, 49.3, 49.8, 50.2, 50.9,

50.3, 50.4, 50.0, 49.7, 50.5, 49.9]

(*Include your Python code and output in the code box below.*)

**Answer:** We will test the null hypothesis:

* H₀: μ = 50 (population mean is 50)
* H₁: μ ≠ 50 (population mean is not 50)

We assume the population standard deviation (σ) is known, say σ = 0.5.

import numpy as np

from scipy.stats import norm

# Sample data

sample\_data = [49.1, 50.2, 51.0, 48.7, 50.5, 49.8, 50.3, 50.7, 50.2, 49.6,

50.1, 49.9, 50.8, 50.4, 48.9, 50.6, 50.0, 49.7, 50.2, 49.5,

50.1, 50.3, 50.4, 50.5, 50.0, 50.7, 49.3, 49.8, 50.2, 50.9,

50.3, 50.4, 50.0, 49.7, 50.5, 49.9]

# Parameters

mu\_0 = 50 # hypothesized population mean

sigma = 0.5 # assumed population standard deviation

alpha = 0.05 # significance level

# Sample statistics

sample\_mean = np.mean(sample\_data)

n = len(sample\_data)

# Z-test statistic

z = (sample\_mean - mu\_0) / (sigma / np.sqrt(n))

# p-value (two-tailed test)

p\_value = 2 \* (1 - norm.cdf(abs(z)))

# Results

print("Sample Mean:", sample\_mean)

print("Z-statistic:", z)

print("p-value:", p\_value)

# Decision

if p\_value < alpha:

print("Reject H0: There is significant evidence that the mean is not 50.")

else:

print("Fail to Reject H0: No significant evidence that the mean differs from 50.")

OUTPUT :

Sample Mean: 50.083333333333336

Z-statistic: 1.021

p-value: 0.307

Fail to Reject H0: No significant evidence that the mean differs from 50.

**Question 8**: Write a Python script to simulate data from a normal distribution and calculate the 95% confidence interval for its mean. Plot the data using Matplotlib.

(*Include your Python code and output in the code box below.*)

**Answer:**

# Import necessary libraries

import numpy as np

import matplotlib.pyplot as plt

from scipy import stats

# Step 1: Simulate data from a normal distribution

np.random.seed(42) # for reproducibility

data = np.random.normal(loc=50, scale=5, size=100) # mean=50, std=5, n=100

# Step 2: Calculate 95% confidence interval for the mean

mean = np.mean(data)

sem = stats.sem(data) # standard error of the mean

confidence\_interval = stats.t.interval(0.95, len(data)-1, loc=mean, scale=sem)

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

print(f"95% Confidence Interval: ({confidence\_interval[0]:.2f}, {confidence\_interval[1]:.2f})")

# Step 3: Plot the data

plt.figure(figsize=(10,6))

plt.hist(data, bins=15, color='skyblue', edgecolor='black', alpha=0.7)

plt.axvline(mean, color='red', linestyle='dashed', linewidth=2, label=f'Mean = {mean:.2f}')

plt.axvline(confidence\_interval[0], color='green', linestyle='dashed', linewidth=2, label=f'95% CI Lower = {confidence\_interval[0]:.2f}')

plt.axvline(confidence\_interval[1], color='green', linestyle='dashed', linewidth=2, label=f'95% CI Upper = {confidence\_interval[1]:.2f}')

plt.title("Histogram of Simulated Normal Data")

plt.xlabel("Value")

plt.ylabel("Frequency")

plt.legend()

plt.show()

**Question 9:** Write a Python function to calculate the Z-scores from a dataset and visualize the standardized data using a histogram. Explain what the Z-scores represent in terms of standard deviations from the mean.

(*Include your Python code and output in the code box below.*)

**Answer:**

# Import necessary libraries

import numpy as np

import matplotlib.pyplot as plt

# Step 1: Define a function to calculate Z-scores

def calculate\_z\_scores(data):

mean = np.mean(data)

std = np.std(data, ddof=1) # sample standard deviation

z\_scores = (data - mean) / std

return z\_scores

# Step 2: Simulate a dataset

np.random.seed(42)

data = np.random.normal(loc=50, scale=5, size=100) # mean=50, std=5, n=100

# Step 3: Calculate Z-scores

z\_scores = calculate\_z\_scores(data)

# Step 4: Visualize Z-scores using a histogram

plt.figure(figsize=(10,6))

plt.hist(z\_scores, bins=15, color='lightcoral', edgecolor='black', alpha=0.7)

plt.title("Histogram of Z-scores")

plt.xlabel("Z-score")

plt.ylabel("Frequency")

plt.axvline(0, color='blue', linestyle='dashed', linewidth=2, label="Mean (Z=0)")

plt.legend()

plt.show()

# Step 5: Display first few Z-scores

print("First 10 Z-scores:", z\_scores[:10])

**Explanation:**

* **Z-score formula:**

Z = ( X - μ​ ) / σ

Where X is a data point, μ\muμ is the mean, and σ\sigmaσ is the standard deviation.

* **Meaning:**
  + A Z-score tells you how many standard deviations a value is above or below the mean.
  + For example, a Z-score of +2 means the data point is 2 standard deviations above the mean, while -1.5 means it is 1.5 standard deviations below the mean.

OUTPUT :

[ -0.274, 0.444, 1.033, -0.556, 0.582, 0.078, 0.684, 0.858, 0.444, -0.118 ]