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:

Hypothesis testing is a statistical method used to decide whether an assumption (claim) about a population is true or not, based on sample data.

- **Null Hypothesis (H_0): The default assumption (e.g., "no difference" or "no effect").**
- **Alternative Hypothesis (H_1): The statement we want to test (e.g., "there is a difference" or "there is an effect").**

Steps:

1. **Set up H_0 and H_1.**
2. **Choose a significance level (α, usually 0.05).**
3. **Collect data and calculate a test statistic and p-value.**
4. **Make a decision:**
    - **If $p \leq \alpha \rightarrow$ Reject H_0 (evidence supports H_1).**
    - **If $p > \alpha \rightarrow$ Fail to reject H_0 (not enough evidence).**

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

Answer:

Null Hypothesis (H_0)

- **The default assumption in hypothesis testing.**
- **It usually states that there is no effect, no difference, or no relationship.**
- **Example:** ***"The average height of men = 170 cm."***

Alternative Hypothesis (H_1 or Ha)

- **The statement we want to test/prove.**
- **It usually states that there is an effect, a difference, or a relationship.**
- **Example:** ***"The average height of men ≠ 170 cm."***

Key Difference

- **H_0 (Null): "Nothing new is happening" → baseline claim.**
- **H_1 (Alternative): "Something different is happening" → the research claim.**

Hypothesis testing works like a trial:

- **H_0 = innocent until proven guilty.**
- **H_1 = guilty (only accepted if evidence is strong enough).**

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

Answer:

PW SKILLS

Significance Level (α)

- The **significance level (α)** is a threshold we set **before** doing a hypothesis test.
- It represents the probability of making a **Type I error** → rejecting the null hypothesis (H_0) when it is actually true.
- Common choices: **0.05 (5%)**, **0.01 (1%)**, **0.10 (10%)**.

Role in Hypothesis Testing

1. **Set the cutoff:**
   - If the **p-value ≤ α**, we **reject H_0** → evidence supports H_1.
   - If the **p-value > α**, we **fail to reject H_0** → not enough evidence.
2. **Control risk:**
   - Smaller α → stricter test, less chance of false positives, but higher chance of missing a real effect (Type II error).
   - Larger α → more chance of detecting effects, but higher risk of false positives.

Example

Suppose α = 0.05.

- A drug trial gives a p-value of 0.03.

- Since **$0.03 \le 0.05$**, we reject H_0 → the new drug likely has an effect.
- Here, $\alpha = 0.05$ means we accept a 5% risk of wrongly concluding the drug works when it actually doesn't.

In short: **Significance level is the "decision cutoff" that balances the risk of making errors in hypothesis testing.**

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

Answer:

Type I Error (False Positive)

- **Happens when we reject H_0 even though it is true.**
- **In simple words: we think there is an effect/difference, but actually, there isn't.**
- **Probability of this error = α (significance level).**

Example:

- **Court case: An innocent person (H_0 true) is declared guilty → Type I error.**
- **Medicine trial: We conclude a new drug works better, but in reality, it doesn't.**

Type II Error (False Negative)

- **Happens when we fail to reject H_0 even though H_1 is true.**
- **In simple words: we miss a real effect/difference.**
- **Probability of this error = β (related to test's power = $1 - \beta$).**

Example:

- **Court case: A guilty person (H_1 true) is declared innocent → Type II error.**
- **Medicine trial: We conclude the new drug has no effect, but in reality, it works.**

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

Answer:

Difference between Z-test and T-test

Point	Z-test	T-test
Population standard deviation (σ)	**Known**	**Unknown (estimated from sample)**
Sample size (n)	**Large (n > 30, approx.)**	**Small (n ≤ 30, usually)**
Distribution used	**Normal distribution (Z)**	**Student's t-distribution**
Curve shape	**Fixed, narrower**	**Depends on n, heavier tails for small samples**
Use cases	**- Mean test (σ known, large n) - Proportion test**	**- Mean test (σ unknown, small n) - Comparing two means**

When to Use Each

- **Z-test:**
  - **Large sample size**
  - **Population standard deviation (σ) is known**
  - **Example: Testing average IQ of 500 people when σ is given.**
- **T-test:**
  - **Small sample size**
  - **Population standard deviation (σ) is unknown**

- **Example: Testing average marks of 15 students without knowing σ.**

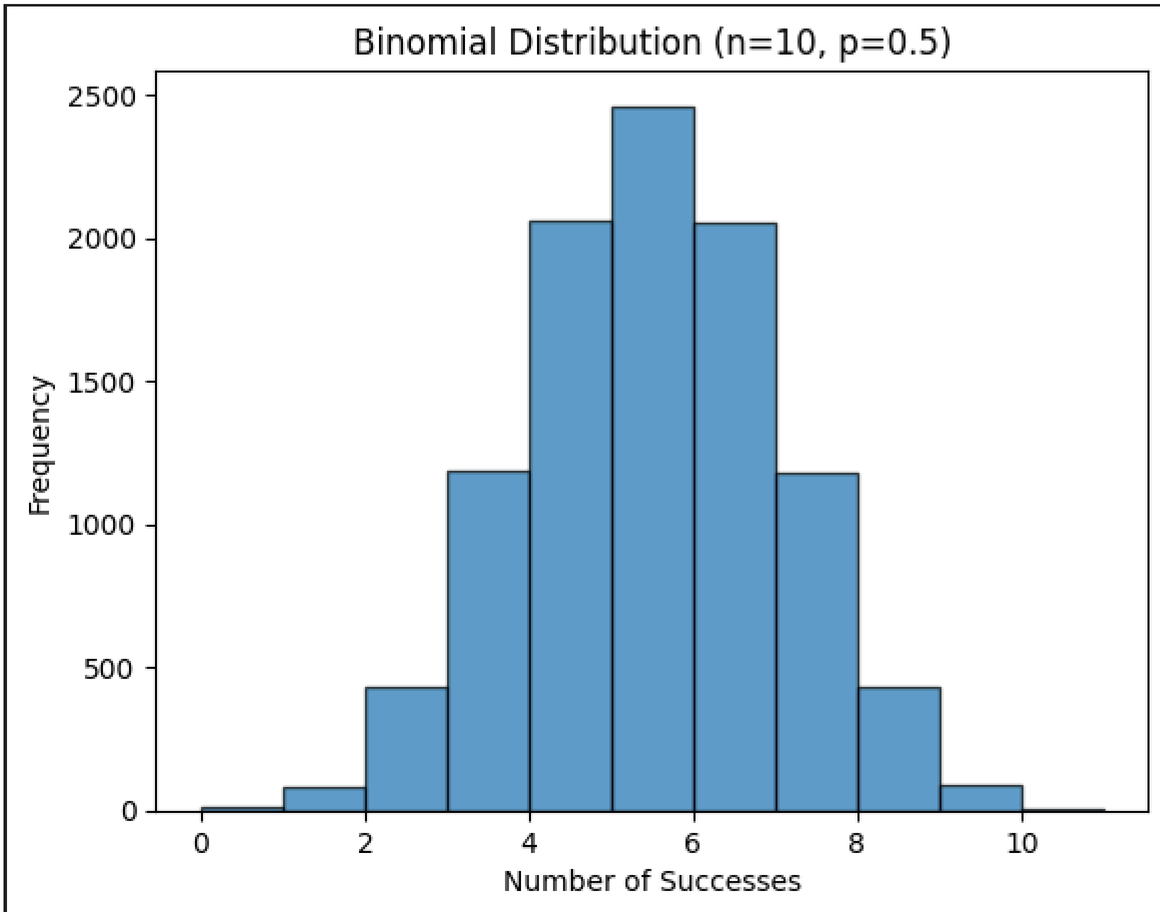
Question 6: Write a Python program to generate a binomial distribution with n=10 and p=0.5, then plot its histogram.

(*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 = 10000  # number of samples

# Generate binomial distribution samples
data = np.random.binomial(n, p, size)

# Plot histogram
plt.hist(data, bins=range(n+2), edgecolor='black', alpha=0.7)
plt.title(f'Binomial Distribution (n={n}, p={p})')
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:

```
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]

# Hypothesized population mean
mu_0 = 50

# Sample statistics
sample_mean = np.mean(sample_data)
sample_std = np.std(sample_data, ddof=1)    # sample std (unbiased)
n = len(sample_data)

# Z-statistic
z_stat = (sample_mean - mu_0) / (sample_std / np.sqrt(n))

# Two-tailed p-value
p_value = 2 * (1 - norm.cdf(abs(z_stat)))

print("Sample Mean:", sample_mean)
print("Sample Std Dev:", sample_std)
print("Sample Size:", n)
print("Z-statistic:", z_stat)
print("p-value:", p_value)

# Decision
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis (H0).")
else:
    print("Fail to reject the null hypothesis (H0).")

#output:-
Sample Mean: 50.08888888888889
Sample Std Dev: 0.5365379910807955
Sample Size: 36
Z-statistic: 0.9940271559503017
p-value: 0.3202096468890012
Fail to reject the null hypothesis (H0).
```

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.

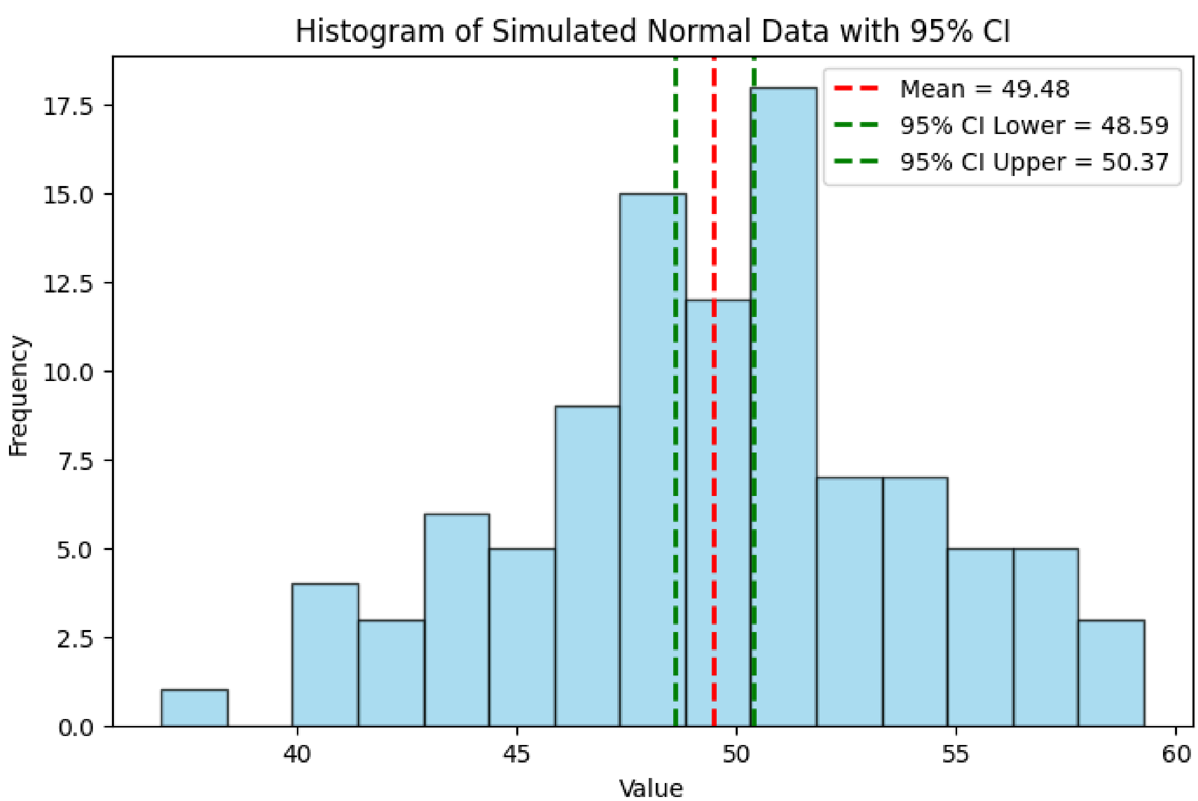
(*Include your Python code and output in the code box below.*)

Answer:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# Step 1: Simulate data
np.random.seed(42)  # for reproducibility
mu, sigma = 50, 5   # true mean and std deviation
n = 100             # sample size
data = np.random.normal(mu, sigma, n)

# Step 2: Calculate 95% confidence interval for the mean
sample_mean = np.mean(data)
sample_std = np.std(data, ddof=1)

confidence_level = 0.95
alpha = 1 - confidence_level
z_critical = stats.norm.ppf(1 - alpha/2)  # Z critical value

margin_of_error = z_critical * (sample_std / np.sqrt(n))
ci_lower = sample_mean - margin_of_error
ci_upper = sample_mean + margin_of_error

print(f"Sample Mean = {sample_mean:.2f}")
print(f"95% Confidence Interval = ({ci_lower:.2f}, {ci_upper:.2f})")

# Step 3: Plot the data
plt.figure(figsize=(8,5))
plt.hist(data, bins=15, edgecolor='black', alpha=0.7, color='skyblue')
plt.axvline(sample_mean, color='red', linestyle='dashed', linewidth=2,
```

```
label=f"Mean = {sample_mean:.2f}")
plt.axvline(ci_lower, color='green', linestyle='dashed', linewidth=2,
label=f"95% CI Lower = {ci_lower:.2f}")
plt.axvline(ci_upper, color='green', linestyle='dashed', linewidth=2,
label=f"95% CI Upper = {ci_upper:.2f}")

plt.title("Histogram of Simulated Normal Data with 95% CI")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

#output:-
Sample Mean = 49.48
95% Confidence Interval = (48.59, 50.37)



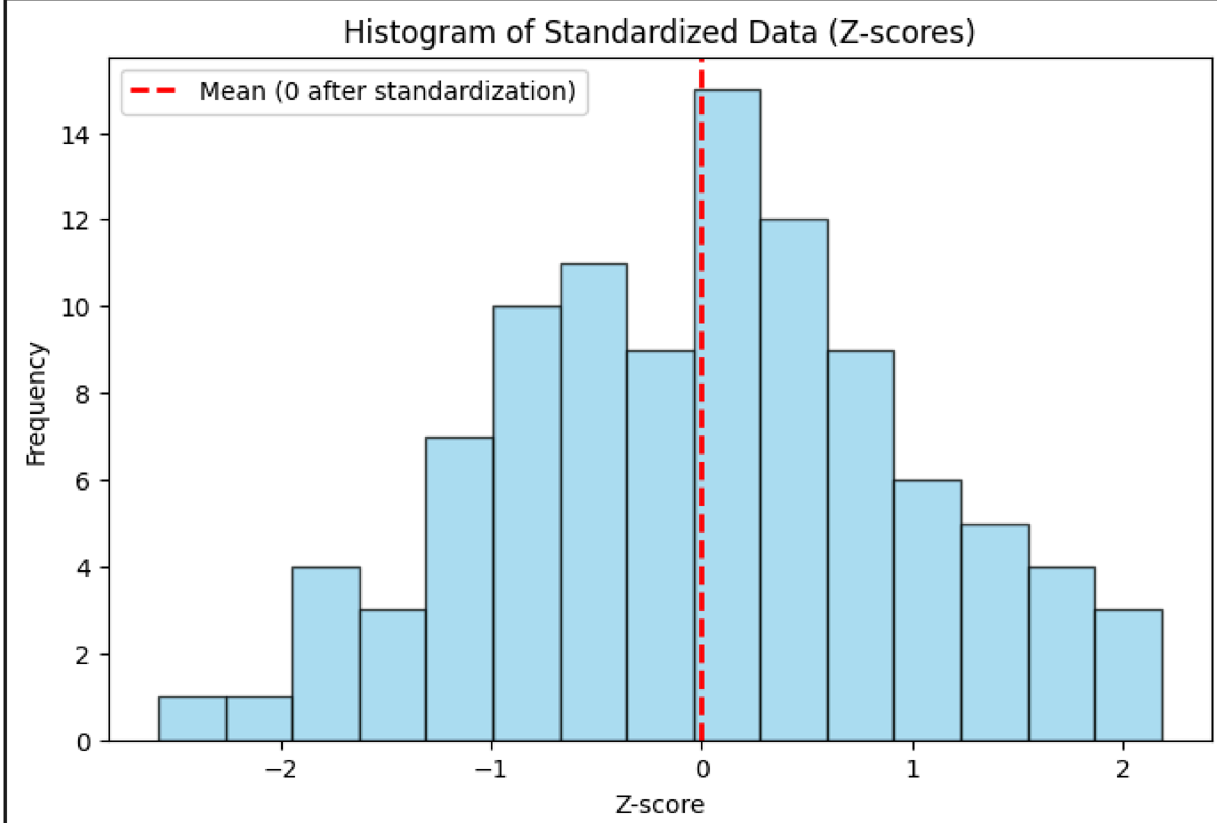
Question 9: Write a Python function to calculate the Z-scores from a dataset and visualize the standardized data using a histogram. Explain what the Z-scores represent in terms of standard deviations from the mean.

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

Answer:

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

def calculate_z_scores(data):
    """
    Calculate Z-scores for a dataset.
    Z = (x - mean) / std
    """
    mean = np.mean(data)
    std = np.std(data, ddof=1)  # sample standard deviation
    z_scores = (data - mean) / std
    return z_scores, mean, std

# Example dataset
np.random.seed(0)
data = np.random.normal(loc=50, scale=5, size=100)  # N(50, 5^2)

# Calculate Z-scores
z_scores, mean, std = calculate_z_scores(data)

print("Original Mean:", mean)
print("Original Std Dev:", std)
print("First 5 Z-scores:", z_scores[:5])

# Plot histogram of standardized data
plt.figure(figsize=(8,5))
plt.hist(z_scores, bins=15, color='skyblue', edgecolor='black',
alpha=0.7)
plt.axvline(0, color='red', linestyle='dashed', linewidth=2, label="Mean
(0 after standardization)")
plt.title("Histogram of Standardized Data (Z-scores)")
plt.xlabel("Z-score")
plt.ylabel("Frequency")
plt.legend()
plt.show()
#output:-
Original Mean: 50.29904007767244
Original Std Dev: 5.064798846342509
First 5 Z-scores: [1.68244029 0.33599478 0.90717321 2.15318046
```

```
1.78462169]
```



Histogram of Standardized Data (Z-scores)
Mean (0 after standardization)
Frequency
Z-score
0
2
4
6
8
10
12
14
−2
−1
0
1
2