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: 100

Question 1: What is hypothesis testing in statistics?

Answer:

Hypothesis testing in statistics is a method used to make decisions or draw conclusions about a population based on sample data.
It helps you determine whether there is enough statistical evidence to support a particular belief or claim (called a **hypothesis**) about a population parameter.

- **Key Concepts:**
  1. **Null Hypothesis (H_0):**
     The default or initial assumption — usually states that *there is no effect or no difference*.
     Example:

H_0: The average blood pressure of patients = 120 mmHg

  2. **Alternative Hypothesis (H_1 or Ha):**
     The claim you want to test — it states that *there is an effect or a difference*.
     Example:

H_1: The average blood pressure of patients ≠ 120 mmHg

- **Steps in Hypothesis Testing:**
  1. **State the hypotheses**
     - H_0: Null hypothesis
     - H_1: Alternative hypothesis
  2. **Choose a significance level (α)**
     - Common values: 0.05, 0.01, or 0.10
     - It represents the probability of rejecting H_0 when it is actually true (Type I error).
  3. **Select the appropriate statistical test**
     - Example: *t-test*, *z-test*, *chi-square test*, *ANOVA*, etc.
  4. **Calculate the test statistic and p-value**

- The **test statistic** measures how far the sample result is from what's expected under H_0.
- The **p-value** is the probability of obtaining a result as extreme as (or more extreme than) the observed one if H_0 is true.

5. **Make a decision:**
   - If **p-value ≤ α**, reject H_0 → significant result.
   - If **p-value > α**, fail to reject H_0 → not enough evidence against H_0.

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

Answer:

◇ **Null Hypothesis (H_0)**

The **null hypothesis** is a statement that assumes **no effect, no difference, or no relationship** exists between variables.
It represents the **status quo** or what we assume to be true until proven otherwise.
Purpose:
It provides a starting point for statistical testing. You test against it to see if there is enough evidence to reject it.
Example:
A company claims that their new diet pill does not change weight.

- H_0: The diet pill has **no effect** on weight (mean weight loss = 0).

◇ **Alternative Hypothesis (H_1 or Ha)**

The **alternative hypothesis** is the statement you want to **prove** or **find evidence for**.
It suggests that there **is an effect, a difference, or a relationship**.
Example:
Continuing the same case:

- H_1: The diet pill **does affect** weight (mean weight loss ≠ 0).

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

Answer:

◇ **Definition**

The **significance level (α)** is the **probability of rejecting the null hypothesis when it is actually true**.
👉 In other words, it's the **risk of making a Type I error** (a *false positive*).
It tells us how much evidence we require before deciding that the effect or difference we observed is *statistically significant*.

◇ **Common Significance Levels**

α (alpha)	Confidence Level	Typical Use
0.10	90%	More tolerance for error
0.05	95%	Most commonly used
0.01	99%	When high accuracy is needed (e.g., medical research)

◇ **Role in Hypothesis Testing**

1. **You choose α before the test begins** (e.g., 0.05).
2. After analyzing your sample data, you calculate a **p-value**.
3. **Compare p-value with α:**
    - If **p ≤ α**, → Reject H_0 → Evidence is strong enough → **Statistically significant** result.
    - If **p > α**, → Fail to reject H_0 → Evidence is not strong enough → **Not significant**.

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

Answer:

◇ **1. Type I Error (False Positive)**

Definition:
A **Type I error** occurs when we **reject the null hypothesis (H_0) even though it is true**.
👉 You're saying there **is an effect or difference**, but in reality, **there isn't**.
Probability of making this error = α (significance level).
Example:
A medical test for a disease shows that a healthy person *has the disease*.

- **H_0:** The person does **not** have the disease.
- **H_1:** The person **has** the disease.
- If the test incorrectly rejects H_0 → **Type I error.**

In short: You found a "difference" that doesn't actually exist.

◇ **2. Type II Error (False Negative)**

Definition:
A **Type II error** occurs when we **fail to reject the null hypothesis (H_0) even though it is false**.
👉 You're saying there **is no effect or difference**, but in reality, **there is**.
Probability of making this error = β (beta).
The **power of a test = 1 − β** (the probability of correctly detecting a true effect).
Example:
A medical test fails to detect a disease in a sick person.

- **H_0:** The person does **not** have the disease.
- **H_1:** The person **has** the disease.
- If the test incorrectly fails to reject H_0 → **Type II error.**

In short: You missed a real difference.

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

Answer:

Feature	Z-test	T-test
Population standard deviation (σ)	Known	Unknown (use sample SD s)
Sample size	Large (n > 30)	Small (n ≤ 30)
Distribution used	Standard normal (Z)	Student's t-distribution
Shape of distribution	Fixed	Depends on degrees of freedom (n – 1)
Variance estimation	Assumed accurate	Estimated from sample
Example	Quality control with large data	Comparing small sample means

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 required libraries
import numpy as np
import matplotlib.pyplot as plt

# Parameters for the binomial distribution
n = 10     # number of trials
p = 0.5    # probability of success
size = 1000  # number of random samples

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

# Print first few values
print("First 10 random samples from the Binomial distribution:")
print(data[:10])

# Plot histogram
plt.hist(data, bins=range(0, n+2), edgecolor='black', alpha=0.7)
plt.title('Binomial Distribution (n=10, p=0.5)')
plt.xlabel('Number of Successes')
plt.ylabel('Frequency')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

OUTPUT:

```
First 10 random samples from the Binomial distribution:
[4 5 7 6 3 4 5 6 5 4]
```

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]

# Step 1: Define known parameters
mu_0 = 50        # Population mean (hypothesized)
sigma = 0.5      # Population standard deviation
alpha = 0.05     # Significance level

# Step 2: Compute sample statistics
x_bar = np.mean(sample_data)
n = len(sample_data)

# Step 3: Compute Z statistic
z = (x_bar - mu_0) / (sigma / np.sqrt(n))

# Step 4: Compute p-value (two-tailed)
p_value = 2 * (1 - norm.cdf(abs(z)))

# Step 5: Print results
print(f"Sample mean (x̄): {x_bar:.3f}")
print(f"Z-statistic: {z:.3f}")
print(f"P-value: {p_value:.4f}")

# Step 6: Decision rule
if p_value < alpha:
    print("☑ Reject the null hypothesis (H₀). The sample mean is significantly different from 50.")
else:
    print("❌ Fail to reject the null hypothesis (H₀). No significant difference from 50.")
```

OUTPUT:

```
Sample mean (x̄): 50.066
Z-statistic: 0.793
P-value: 0.4280
❌ Fail to reject the null hypothesis (H₀). No significant difference 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 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
mean = 100                     # true population mean
std_dev = 15                   # true population standard deviation
n = 100                      # sample size

data = np.random.normal(mean, std_dev, n)

# Step 2: Compute sample statistics
sample_mean = np.mean(data)
sample_std = np.std(data, ddof=1)   # sample standard deviation

# Step 3: Compute 95% confidence interval for the mean
confidence_level = 0.95
alpha = 1 - confidence_level
t_critical = stats.t.ppf(1 - alpha/2, df=n-1)  # t-critical value
margin_of_error = t_critical * (sample_std / np.sqrt(n))
ci_lower = sample_mean - margin_of_error
ci_upper = sample_mean + margin_of_error

# Step 4: Print results
print(f"Sample Mean: {sample_mean:.2f}")
print(f"Sample Standard Deviation: {sample_std:.2f}")
print(f"95% Confidence Interval for Mean: ({ci_lower:.2f}, {ci_upper:.2f})")

# Step 5: Plot histogram of the data
plt.hist(data, bins=15, color='skyblue', edgecolor='black', alpha=0.7)
plt.axvline(sample_mean, color='red', linestyle='dashed', linewidth=2, label=f"Mean =
{sample_mean:.2f}")
plt.axvline(ci_lower, color='green', linestyle='dotted', linewidth=2, label=f"95% CI Lower =
{ci_lower:.2f}")
plt.axvline(ci_upper, color='green', linestyle='dotted', linewidth=2, label=f"95% CI Upper =
{ci_upper:.2f}")
```

```
plt.title('Normal Distribution Data with 95% Confidence Interval')
plt.xlabel('Values')
plt.ylabel('Frequency')
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()
```

OUTPUT:

Sample Mean: 98.99

Sample Standard Deviation: 14.86

95% Confidence Interval for Mean: (96.05, 101.92)

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

# Step 1: Define a function to calculate Z-scores
def calculate_z_scores(data):
    mean = np.mean(data)
    std_dev = np.std(data)
    z_scores = (data - mean) / std_dev
    return z_scores, mean, std_dev

# Step 2: Create a sample dataset (normal distribution)
np.random.seed(0)
data = np.random.normal(loc=50, scale=10, size=100)

# Step 3: Calculate Z-scores
z_scores, mean, std_dev = calculate_z_scores(data)

# Step 4: Print sample results
print(f"Original Mean: {mean:.2f}")
print(f"Original Standard Deviation: {std_dev:.2f}")
print("First 5 Z-scores:", np.round(z_scores[:5], 2))

# Step 5: Visualize the standardized data
```

```
plt.hist(z_scores, bins=15, color='lightcoral', edgecolor='black', alpha=0.7)
plt.axvline(0, color='blue', linestyle='dashed', linewidth=2, label="Mean (Z=0)")
plt.title('Histogram of Z-Scores (Standardized Data)')
plt.xlabel('Z-Score')
plt.ylabel('Frequency')
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()
```

OUTPUT:

```
Original Mean: 50.60
Original Standard Deviation: 9.74
First 5 Z-scores: [ 1.09  0.21  0.48  1.23  1.1 ]
```