Assignment Code: DA-AG-007 Statistics Advanced - 2| Assignment

Question 1: What is hypothesis testing in statistics?

Introduction:

Hypothesis testing is a **fundamental statistical technique** used to make decisions or inferences about population parameters based on sample data. It helps researchers determine whether there is enough evidence to support a specific claim or hypothesis about a population.

Definition:

Hypothesis testing is a **formal procedure** for comparing observed data with a statement (hypothesis) whose truth we want to assess. It involves making a claim (the hypothesis), collecting data, and then using statistical methods to decide whether to reject or fail to reject that claim.

Types of Hypotheses:

1. **Null Hypothesis (H_0):**
    - This is the **default or initial claim** that there is **no effect or no difference**.
    - It assumes that any observed difference is due to chance.
    - Example: H_0: $\mu = 50$
2. **Alternative Hypothesis (H_1 or Ha):**
    - This is what you **want to prove or support**.
    - It represents a new theory or effect.
    - Example: H_1: $\mu \neq 50$

Steps in Hypothesis Testing:

1. **State the hypotheses (H_0 and H_1).**
2. **Choose the level of significance (α):**
    - Common values: 0.05, 0.01
    - It represents the probability of rejecting H_0 when it is actually true (Type I error).
3. **Select the appropriate test statistic:**
    - e.g., z-test, t-test, chi-square test, etc.
4. **Determine the critical value or p-value.**

5. **Make a decision:**
    - If **p-value ≤ α**, reject H_0
    - If **p-value > α**, fail to reject H_0
6. **Interpret the results in the context of the problem.**

Types of Errors:

1. **Type I Error (α):**
    - Rejecting H_0 when it is actually true.
2. **Type II Error (β):**
    - Failing to reject H_0 when H_1 is true.

Example:

A factory claims that the average weight of its sugar packets is 1 kg. A consumer doubts this and tests 30 packets, finding a sample mean of 0.98 kg.

- H_0: $\mu = 1$ kg
- H_1: $\mu \neq 1$ kg
- After performing a t-test, the p-value is found to be 0.03.
- If $\alpha = 0.05$, then $0.03 < 0.05$ → **Reject H_0**
- Conclusion: There is sufficient evidence to doubt the factory's claim.

Importance of Hypothesis Testing:

- Supports **evidence-based decision making**
- Used in **scientific research**, **quality control**, **medicine**, **business**, and many other fields
- Helps reduce subjectivity and guesswork in data interpretation

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

Introduction:

In statistical hypothesis testing, we use two opposing hypotheses to make inferences about a population based on sample data: the **null hypothesis** and the **alternative hypothesis**. These form the foundation of hypothesis testing.

1. Definition of Null Hypothesis (H_0):

The **null hypothesis** is a **statement of no effect, no difference, or no relationship**. It represents the default or status quo assumption that any observed change or difference in data is due to **random chance** or natural variation.

- **Symbol:** H_0
- **Purpose:** To be tested and possibly rejected
- **Example:**
  - H_0: $\mu = 50$ (The population mean is 50)
  - H_0: There is no difference in test scores between two groups.

---

2. Definition of Alternative Hypothesis (H_1 or Ha):

The **alternative hypothesis** is the **statement you want to prove**. It suggests that there **is** an effect, difference, or relationship in the population.

- **Symbol:** H_1 or Ha
- **Purpose:** Competes with H_0. If H_0 is rejected, we accept H_1.
- **Example:**
  - H_1: $\mu \neq 50$ (The population mean is not 50)
  - H_1: There is a difference in test scores between two groups.

---

3. Key Differences Between Null and Alternative Hypotheses:

Feature	Null Hypothesis (H_0)	Alternative Hypothesis (H_1 or Ha)
Meaning	Assumes no effect or no difference	Assumes there is an effect or difference
Goal	To be tested and possibly rejected	To be accepted if H_0 is rejected
Symbol	H_0	H_1 or Ha
Example	H_0: $\mu = 100$	H_1: $\mu \neq 100$
Basis for Conclusion	Default assumption	New claim based on evidence
If p-value < α	Reject H_0	Accept H_1
If p-value > α	Fail to reject H_0	Cannot accept H_1

---

4. Types of Alternative Hypotheses:

- **Two-tailed test:** H_1: $\mu \neq \mu_0$
- **Left-tailed test:** H_1: $\mu < \mu_0$
- **Right-tailed test:** H_1: $\mu > \mu_0$

The choice depends on the research question.

---

5. Example:

Suppose a medicine company claims its drug cures 80% of patients. A scientist wants to test this claim.

- **H_0:** $p = 0.80$ (The cure rate is 80%)
- **H_1:** $p \neq 0.80$ (The cure rate is not 80%)

After testing a sample, if the results strongly differ from 80%, the scientist may **reject H_0** and **accept H_1**.

---

6. Importance of Both Hypotheses:

- Provides a **clear and testable structure** for statistical analysis.
- Ensures **objective decision-making** based on data.
- Prevents biased conclusions by requiring strong evidence to reject H_0.

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

Definition:

The **significance level** in hypothesis testing, denoted by **α (alpha)**, is the **probability of rejecting the null hypothesis when it is actually true**. It represents the **risk of committing a Type I error**, i.e., a **false positive**.

Common Significance Levels:

- **0.05 (5%)** → Most commonly used
- **0.01 (1%)**
- **0.10 (10%)**

Role in Hypothesis Testing:

1. **Threshold for Decision Making:**

- The **p-value** obtained from a test is compared to **α**.
- If **$p \leq \alpha$**, we **reject the null hypothesis** (statistically significant).
- If **$p > \alpha$**, we **fail to reject the null hypothesis** (not statistically significant).

2. **Controls Error Risk:**
   - A lower α reduces the risk of **Type I error** but increases the risk of **Type II error**.
3. **Interpreting Results:**
   - If $p = 0.03$ and $\alpha = 0.05$ → **Reject H_0**
   - If $p = 0.06$ and $\alpha = 0.05$ → **Fail to reject H_0**

Example:

A scientist tests whether a drug has an effect (H_1) vs. no effect (H_0).
Using $\alpha = 0.05$:

- **$p = 0.02$** → $0.02 < 0.05$ → **Reject H_0**, the drug is effective.
- **$p = 0.08$** → $0.08 > 0.05$ → **Fail to reject H_0**, not enough evidence.

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

Definition of Errors:

Error Type	Meaning	Probability	Consequence
Type I	Rejecting a true null hypothesis (False Positive)	α	Detecting an effect when there is none
Type II	Failing to reject a false null hypothesis (False Negative)	β	Missing a real effect

---

1. Type I Error (α):

- **Occurs when:** H_0 is true, but we reject it.
- **Example:** A COVID test says you have the virus (positive), but you're healthy.
- **Consequence:** False alarm – might lead to unnecessary treatment.

---

2. Type II Error (β):

- **Occurs when:** H_0 is false, but we fail to reject it.

- **Example:** A COVID test says you don't have the virus (negative), but you actually do.
- **Consequence:** Missed detection – might spread the disease unknowingly.

Visual Representation:

Reality / Decision	Accept H_0	Reject H_0
H_0 is true	✅ Correct	❌ Type I
H_0 is false	❌ Type II	✅ Correct

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

Z-Test vs T-Test

Feature	Z-Test	T-Test
Population SD known?	Yes	No
Sample Size (n)	Large ($n > 30$)	Small ($n \le 30$)
Formula	$Z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$	$t = (\bar{x} - \mu) / (s / \sqrt{n})$
Distribution Used	Standard normal (Z-distribution)	Student's t-distribution
Shape of Distribution	Fixed	Changes with degrees of freedom ($n - 1$)

When to Use:

- **Z-Test:**
  - Population standard deviation (σ) is **known**
  - **Large samples** (Central Limit Theorem applies)
- **T-Test:**
  - Population standard deviation is **unknown**
  - **Small sample sizes**
  - More conservative than Z-test

Example:

- Testing if a sample mean differs from a known population mean:

- **Population σ known → Z-test**
- **Population σ unknown → T-test**

Question 6: Python Program – Binomial Distribution (n=10, p=0.5)

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

# Generate binomial distribution data
n = 10
p = 0.5
size = 1000

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

# Plot histogram
plt.hist(data, bins=range(n+2), align='left', edgecolor='black')
plt.title('Binomial Distribution (n=10, p=0.5)')
plt.xlabel('Number of Successes')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

Question 7: Z-Test in Python (Sample Dataset)

```
import numpy as np
from scipy import stats

# Given 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
x_bar = np.mean(sample_data)
s = np.std(sample_data, ddof=1)
n = len(sample_data)

# Z-statistic
z = (x_bar - mu_0) / (s / np.sqrt(n))
p_value = 2 * (1 - stats.norm.cdf(abs(z)))

print(f"Sample mean = {x_bar:.2f}")
print(f"Z-statistic = {z:.3f}")
print(f"P-value = {p_value:.4f}")

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

Question 8: Simulate Normal Distribution & 95% Confidence Interval

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

```
# Simulate normal data
data = np.random.normal(loc=100, scale=15, size=100)

# Calculate mean and 95% CI
mean = np.mean(data)
se = stats.sem(data)
conf_int = stats.t.interval(0.95, len(data)-1, loc=mean, scale=se)

print(f"Mean = {mean:.2f}")
print(f"95% Confidence Interval = {conf_int}")

# Plot histogram with CI lines
plt.hist(data, bins=15, edgecolor='black', alpha=0.7)
plt.axvline(conf_int[0], color='red', linestyle='--', label='Lower 95% CI')
plt.axvline(conf_int[1], color='green', linestyle='--', label='Upper 95% CI')
plt.axvline(mean, color='blue', label='Mean')
plt.title("Simulated Normal Data with 95% CI")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

Question 9: Z-Score Function & Histogram

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

# Generate random data
data = np.random.normal(loc=50, scale=10, size=100)

# Calculate Z-scores
```

```
z_scores = zscore(data)

# Plot histogram
plt.hist(z_scores, bins=15, edgecolor='black', color='lightblue')
plt.title('Histogram of Z-Scores')
plt.xlabel('Z-Score')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()

# Explanation
print("Z-scores indicate how many standard deviations a value is from the mean.")
print("Z = 0 → mean, Z > 0 → above mean, Z < 0 → below mean.")
```