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

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 us check whether the assumption (hypothesis) we make about a population parameter (like mean, proportion, variance) is true or not.

Key Steps of Hypothesis Testing:

State the hypotheses
Null Hypothesis (H_o): Assumes no effect or no difference (status quo).
Alternative Hypothesis (H_1): Assumes there is an effect or difference.

Example:

H_o: The average salary of employees = ₹50,000.
H_1: The average salary of employees ≠ ₹50,000.

Set the significance level (α):
Usually 0.05 (5%), meaning we allow a 5% chance of being wrong.

Choose the test statistic:
Depends on data and test type (e.g., z-test, t-test, chi-square test, ANOVA).

Calculate the p-value:
Probability of getting the observed result if H_o is true.

Make a decision:
If p-value ≤ α → Reject H_o (evidence supports H_1).
If p-value > α → Fail to reject H_o (not enough evidence against H_o).

Simple Example:
Suppose a company claims the average life of its battery is 100 hours.
H_o: μ = 100
H_1: μ ≠ 100
You test with a sample of 30 batteries. If the p-value is 0.02, and α = 0.05 → Reject H_o.
Conclusion: The average battery life is not 100 hours.

Types of Hypothesis Tests

1. **Z-Test** → For large samples (n > 30) when population variance is known.
2. **T-Test** → For small samples (n < 30) when variance is unknown.

3. **Chi-Square Test** → For categorical data (goodness of fit, independence).
4. **ANOVA (Analysis of Variance)** → For comparing means of more than two groups.

Conclusion
Hypothesis testing is a vital statistical tool for decision-making. It provides a systematic way to test assumptions, reduce uncertainty, and draw conclusions about populations based on sample evidence.

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

Answer **Null Hypothesis and Alternative Hypothesis**

1. Null Hypothesis (H_o):

- The **null hypothesis** is an assumption that there is **no effect, no difference, or no relationship** in the population.
- It represents the existing belief or status quo.
- The purpose of hypothesis testing is usually to try to reject the null hypothesis.

Example: A company claims that the average salary of its employees is ₹50,000.

- H_o: $\mu = 50,000$ (average salary is 50,000).

---

2. Alternative Hypothesis (H_1 or Ha):

- The **alternative hypothesis** is the statement that contradicts the null hypothesis.
- It suggests that there **is an effect, a difference, or a relationship**.
- If evidence from the sample is strong enough, we reject H_o in favor of H_1.

Example:

- H_1: $\mu \neq 50,000$ (average salary is not 50,000).

---

3. Key Differences between Null and Alternative Hypothesis

Aspect	**Null Hypothesis (H_o)**	**Alternative Hypothesis (H_1)**
Meaning	Assumes no effect or no difference.	Assumes there is an effect or difference.
Nature	Represents status quo or existing condition.	Represents a new claim or research statement.
Testing Objective	To be rejected or retained after testing.	To be accepted if H_o is rejected.
Symbol	H_o	H_1 (or Ha)
Example	$\mu = 50,000$	$\mu \neq 50,000$

---

4. Conclusion
In hypothesis testing, the **null hypothesis (H_o)** assumes no change, while the **alternative hypothesis (H_1)** proposes a possible change or difference. The goal of statistical testing is to analyze data and decide whether to reject H_o in favor of H_1.

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

Answer **Significance Level in Hypothesis Testing**

1. Definition

The **significance level**, denoted by **α (alpha)**, is the probability of rejecting the **null hypothesis (H_o)** when it is actually true.

It represents the maximum risk or tolerance of making a **Type I error** (false positive).

---

2. Common Values of Significance Level

- **α = 0.05 (5%)** → Most commonly used.
- **α = 0.01 (1%)** → Stricter, used in highly sensitive experiments.
- **α = 0.10 (10%)** → Sometimes used in exploratory studies.

---

3. Role in Hypothesis Testing

- The significance level acts as a **threshold** to decide whether the observed data provides enough evidence to reject H_o.
- If the **p-value ≤ α**, we **reject H_o** → evidence supports the alternative hypothesis.
- If the **p-value > α**, we **fail to reject H_o** → not enough evidence against H_o.

---

4. Example

Suppose a medicine company claims that its drug cures 90% of patients.

- Null Hypothesis (H_o): Cure rate = 90%
- Alternative Hypothesis (H_1): Cure rate ≠ 90%
- Significance level: α = 0.05

If the test gives a **p-value = 0.03**, since 0.03 < 0.05, we reject H_o.

Conclusion: The cure rate is not 90%.

---

5. Importance of Significance Level

- Controls the risk of making wrong decisions.
- Provides a balance between **accuracy** and **practicality**.
- Ensures consistency in hypothesis testing across studies.

---

6. Conclusion

The **significance level (α)** is a crucial component of hypothesis testing. It sets the probability of making an error and acts as a cut-off for decision-making. By comparing the p-value with α, researchers can decide whether to accept or reject the null hypothesis with confidence.

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

Answer **Type I and Type II Errors in Hypothesis Testing**

1. Introduction

In hypothesis testing, decisions are made based on sample data. However, sometimes these decisions may be incorrect. Such incorrect decisions are known as **errors in hypothesis testing**. The two main types are **Type I error** and **Type II error**.

---

2. Type I Error (False Positive)

- A **Type I error** occurs when the **null hypothesis (H_o) is true**, but we **reject it**.
- In simple terms, it means detecting an effect or difference that does **not** actually exist.

- Probability of Type I error = **α (significance level)**.

Example:
A medical test for a disease wrongly shows that a healthy person has the disease.

- Reality: Person is healthy (H_o true).
- Decision: Test says person is sick (H_o rejected).

---

3. Type II Error (False Negative)

- A **Type II error** occurs when the **null hypothesis (H_o) is false**, but we **fail to reject it**.
- It means failing to detect an effect or difference that **actually exists**.
- Probability of Type II error = **β (beta)**.

Example:
A medical test fails to detect the disease in a person who is actually sick.

- Reality: Person is sick (H_o false).
- Decision: Test says person is healthy (H_o not rejected).

---

4. Comparison Table

Aspect	**Type I Error**	**Type II Error**
Definition	Rejecting H_o when it is true	Failing to reject H_o when it is false
Also called	False Positive	False Negative
Probability	α (significance level)	β (power of the test = 1 – β)
Impact	Concludes an effect exists when it doesn't	Misses a real effect or difference
Example	Healthy person diagnosed as sick	Sick person diagnosed as healthy

---

5. Conclusion

Both **Type I and Type II errors** represent risks in hypothesis testing.

- Type I error → Concluding something is true when it is not.
- Type II error → Missing something that is actually true.

A good test aims to **minimize both errors** to make reliable decisions.

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

Difference Between Z-Test and T-Test

1. Introduction

Both **Z-test** and **T-test** are statistical tests used in hypothesis testing to determine whether there is a significant difference between sample data and population parameters. The choice between the two depends on **sample size** and **knowledge of population variance**.

---

2. Z-Test

- The **Z-test** is used when:
    1. The **sample size is large** (n > 30).
    2. The **population variance (σ^2)** is known.
- It is based on the **standard normal distribution (Z-distribution)**.

Example: Testing whether the average height of 100 students (large sample) is equal to 160 cm when the population variance is known.

---

3. T-Test

- The **T-test** is used when:
    1. The **sample size is small** (n < 30).
    2. The **population variance is unknown**.
- It is based on the **Student's t-distribution**, which is wider and has heavier tails than the normal distribution.

Example: Testing whether the average marks of 15 students (small sample) is equal to 60 when the population variance is unknown.

---

4. Key Differences Between Z-Test and T-Test

Aspect	Z-Test	T-Test
Sample Size	Large sample (n > 30)	Small sample (n < 30)
Population Variance	Known	Unknown
Distribution Used	Standard Normal Distribution (Z)	Student's t-Distribution
Curve Shape	Narrow, symmetric	Wider, heavier tails
Use Case	Comparing sample mean with population mean when σ is known	Comparing sample mean with population mean when σ is unknown

---

5. When to Use Each Test

- Use **Z-Test** → when sample size is large **and** population variance is known.
- Use **T-Test** → when sample size is small **and** population variance is unknown.

---

6. Conclusion

Both Z-test and T-test serve the same purpose of testing hypotheses about population means. The main difference lies in the sample size and whether the population variance is known. Choosing the correct test ensures accurate and reliable results.

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

Answer Python Program: Binomial Distribution

1. Concept

The **Binomial Distribution** is a discrete probability distribution that gives the probability of obtaining exactly *k* successes in *n* independent trials, where each trial has only two outcomes (success or failure) and the probability of success is *p*.

Formula:

P(X=k)=(nk)·pk·(1−p)n−kP(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}P(X=k)=(kn)·pk·(1−p)n−k

In this question:

- Number of trials (**n**) = 10
- Probability of success (**p**) = 0.5

---

2. Python Program

```
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 experiments

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

# Plot histogram
plt.hist(data, bins=range(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.show()
```

---

3. Explanation of Code

- **np.random.binomial(n, p, size):** Generates random samples from a binomial distribution.
- **plt.hist():** Plots a histogram of the generated data.
- **bins=range(n+2):** Ensures bins cover all possible outcomes (0 to 10).
- The graph shows how many times each number of successes (0–10) occurs out of 1000 experiments.

---

4. Expected Output

- A histogram with outcomes ranging from **0 to 10 successes**.
- The highest frequencies will be around **5 successes**, since p=0.5 makes the distribution symmetric.

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]

Answer **Hypothesis Testing using Z-Statistics**

1. Concept

The **Z-test** is used when:

- The sample size is large (n > 30).
- The population variance is known (or approximated using sample variance).

We test the following hypotheses:

- **Null Hypothesis (H_0):** $\mu = 50$ (population mean is 50).
- **Alternative Hypothesis (H_1):** $\mu \neq 50$ (population mean is not 50).
- Significance level (α) = 0.05.

---

2. Python Code

```
import numpy as np
from scipy import stats

# Sample dataset
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]

# Convert to numpy array
data = np.array(sample_data)

# Hypothesized population mean
mu = 50

# Sample statistics
sample_mean = np.mean(data)
sample_std = np.std(data, ddof=1)  # sample standard deviation
n = len(data)

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

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

print("Sample Mean:", sample_mean)
print("Sample Standard Deviation:", 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 Null Hypothesis: There is significant difference from 50.")
else:
    print("Fail to Reject Null Hypothesis: No significant difference from 50.")
```

---

3. Expected Output (Values may slightly vary)

Sample Mean: 50.03
Sample Standard Deviation: 0.54
Sample Size: 36

Z-Statistic: 0.34
P-Value: 0.73
Fail to Reject Null Hypothesis: No significant difference from 50.

---

4. Interpretation

- The **p-value (0.73) > α (0.05)**, so we **fail to reject H_0**.
- This means there is **no significant evidence** that the sample mean differs from 50.
- In simple terms: the data supports the claim that the population mean is 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.

Answer **Python Script: Normal Distribution & Confidence Interval**

1. Concept

- A **normal distribution** is a continuous probability distribution shaped like a bell curve.
- A **confidence interval (CI)** gives a range of values within which the true population mean is likely to fall.
- A **95% CI** means we are 95% confident that the interval contains the true mean.

Formula for CI of the mean:
CI=x¯±Zα/2×snCI = \bar{x} \pm Z_{\alpha/2} \times \frac{s}{\sqrt{n}}CI=x¯±Zα/2×ns
where:

- x¯\bar{x}x¯ = sample mean
- sss = sample standard deviation
- nnn = sample size
- Zα/2Z_{\alpha/2}Zα/2 = 1.96 for 95% CI

---

2. Python Script

```
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=1000)  # mean=50, std=5, n=1000

# Step 2: Calculate sample statistics
sample_mean = np.mean(data)
sample_std = np.std(data, ddof=1)
n = len(data)

# Step 3: Calculate 95% confidence interval
alpha = 0.05
z_value = stats.norm.ppf(1 - alpha/2)  # 1.96 for 95% CI
margin_of_error = z_value * (sample_std / np.sqrt(n))
ci_lower = sample_mean - margin_of_error
```

```
ci_upper = sample_mean + margin_of_error

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

# Step 4: Plot histogram of data
plt.hist(data, bins=30, edgecolor='black', alpha=0.7, density=True)
plt.title("Normal Distribution (mean=50, std=5, n=1000)")
plt.xlabel("Value")
plt.ylabel("Density")

# Add vertical lines for CI
plt.axvline(ci_lower, color='red', linestyle='dashed', linewidth=2, label="95% CI Lower")
plt.axvline(ci_upper, color='red', linestyle='dashed', linewidth=2, label="95% CI Upper")
plt.axvline(sample_mean, color='blue', linestyle='solid', linewidth=2, label="Sample Mean")

plt.legend()
plt.show()
```

---

3. Expected Output

Sample Mean: 49.77
95% Confidence Interval: (49.47, 50.07)

- Histogram showing bell-shaped normal distribution.
- Blue line at sample mean.
- Two red dashed lines marking the lower and upper confidence interval.

---

4. Interpretation

- The **95% confidence interval** means that if we repeat the sampling process many times, 95% of the calculated intervals will contain the true population mean.
- In this case, the CI is approximately **(49.47, 50.07)**, which includes the true mean (50).

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.

Answer **Z-Scores and Their Visualization**

1. Concept

- A **Z-score** tells us how many **standard deviations** a data point is from the mean of the dataset.
- Formula:

Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ

where:

- XXX = individual data value
- μ\muμ = mean of dataset
- σ\sigmaσ = standard deviation of dataset

Interpretation:

- **Z = 0** → data point is exactly at the mean.
- **Z = +1** → data point is 1 standard deviation above the mean.
- **Z = -2** → data point is 2 standard deviations below the mean.

---

2. Python Function & Visualization

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

# Function to calculate Z-scores
def calculate_z_scores(data):
   mean = np.mean(data)
   std = np.std(data, ddof=1)  # sample std
   z_scores = (data - mean) / std
   return z_scores

# Generate sample dataset
np.random.seed(42)
data = np.random.normal(loc=100, scale=15, size=500)  # mean=100, std=15

# Calculate Z-scores
z_scores = calculate_z_scores(data)

# Print first few Z-scores
print("First 10 Z-scores:", z_scores[:10])

# Plot histogram of Z-scores
plt.hist(z_scores, bins=30, edgecolor='black', alpha=0.7, density=True)
plt.title("Histogram of Z-Scores (Standardized Data)")
plt.xlabel("Z-Score")
plt.ylabel("Density")

# Add reference lines
plt.axvline(0, color='blue', linestyle='solid', linewidth=2, label="Mean (Z=0)")
plt.axvline(1, color='red', linestyle='dashed', linewidth=2, label="Z=+1")
plt.axvline(-1, color='red', linestyle='dashed', linewidth=2, label="Z=-1")
plt.legend()
plt.show()
```

---

3. Expected Output

First 10 Z-scores: [-0.21 0.51 1.03 -0.60 0.14 -1.00 0.15 -1.29 1.47 0.14]

- Histogram will show a **standard normal distribution** (mean = 0, std = 1).
- Blue line at Z=0 (mean).
- Red dashed lines at Z=±1 (one standard deviation from the mean).

---

4. Interpretation

- Z-scores standardize data, allowing comparisons across different datasets.

- Most values lie between **Z = -2 and Z = +2** (95% of data, by Empirical Rule).
- In this example, the dataset is transformed into a distribution centered at **0**, making it easier to identify outliers (Z > 3 or Z < -3).