Assignment - Statistics Advanced 2

Question 1: What is hypothesis testing in statistics?

ANS:- Hypothesis testing is a statistical method used to make decisions or inferences about a population based on sample data.
It helps us test whether a claim (assumption) about a population parameter is true or not.

1. **Hypothesis**: A statement about a population parameter.
    - **Null Hypothesis (H_0)**: The claim that there is **no effect or no difference**. (Status quo assumption)
    - **Alternative Hypothesis (H_1 or Ha)**: The claim that there **is an effect or difference**.
2. **Test Statistic**: A value calculated from sample data that helps decide whether to reject H_0.
3. **Significance Level (α)**: The probability threshold (commonly 0.05) below which we reject H_0. It represents the risk of making a **Type I error** (rejecting a true H_0).
4. **p-value**: The probability of observing results as extreme as (or more extreme than) the sample results, assuming H_0 is true.
    - If $p \leq \alpha$ → Reject H_0 (evidence supports H_1).
    - If $p > \alpha$ → Fail to reject H_0 (not enough evidence against H_0).
5. **Errors**:
    - **Type I error**: Rejecting H_0 when it's actually true.

- **Type II error**: Failing to reject H_0 when it's actually false.

Steps in Hypothesis Testing

1. **State the hypotheses** (H_0 and H_1).
2. **Choose significance level (α)** (e.g., 0.05).
3. **Collect sample data & compute test statistic**.
4. **Find p-value or critical value**.
5. **Make decision**: Reject or fail to reject H_0.
6. **Conclude** in context of the problem.

Example

Suppose a manufacturer claims that the average lifetime of a battery is **100 hours**.

- **H_0**: $\mu = 100$ (no difference)
- **H_1**: $\mu \neq 100$ (battery life is different)

We collect a sample, calculate the mean, run a statistical test (like a **t-test**), and decide based on the p-value whether to reject H_0.

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

ANS:- Null Hypothesis (H_0)

- **The null hypothesis is the default assumption or claim.**
- **It states that there is no effect, no difference, or no relationship in the population.**

- **It represents the status quo or what we assume to be true until we have strong evidence otherwise.**

👉 **Example:**
A company claims that a battery lasts 100 hours on average.

- **H_0: $\mu = 100$ (the average battery life is 100 hours).**

Alternative Hypothesis (H_1 or Ha)

- The **alternative hypothesis** is what we want to test for.
- It states that **there is an effect, a difference, or a relationship**.
- It represents a **challenge to the status quo**.

👉 Example:

- **H_1**: $\mu \neq 100$ (the average battery life is not 100 hours).

Key Differences Between H_0 and H_1

Feature	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
Meaning	No effect / no difference	There is an effect / difference
Role	Default assumption (status quo)	Competing claim being tested
Equality sign	Always includes **=** (e.g., $\mu = 100$)	Uses ≠, <, or > (e.g., $\mu \neq 100$, $\mu > 100$, $\mu < 100$)
Decision	We either **reject H_0** or **fail to reject H_0**	Supported only if H_0 is rejected
Analogy	Innocent until proven guilty	Guilty (if evidence is strong enough)

In short:

- **Null hypothesis (H_0)** = No change, no effect, default assumption.

- **Alternative hypothesis (H_1)** = There is a change, effect, or difference.

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

ANS:- What is Significance Level (α)?

The significance level (α) is the threshold probability that helps us decide whether to reject the null hypothesis (H_0).

- **It represents the risk of making a Type I error (rejecting H_0 when H_0 is actually true).**
- **Common values are 0.05 (5%), 0.01 (1%), or 0.10 (10%).**

👉 **If $\alpha = 0.05$, it means:**
We are willing to take a 5% risk of incorrectly rejecting the null hypothesis.

🔹 Role in Hypothesis Testing

1. **Set the significance level (α) before collecting data (e.g., 0.05).**
2. **Compare the p-value with α:**
   - **If $p \leq \alpha \rightarrow$ Reject H_0 (evidence supports H_1).**
   - **If $p > \alpha \rightarrow$ Fail to reject H_0 (not enough evidence against H_0).**
3. **It acts like a decision cutoff: how strong the evidence must be before we reject the null hypothesis.**

🔹 Example

A drug manufacturer claims a new medicine reduces blood pressure by 10 mmHg.
We test it at $\alpha = 0.05$.

- **After running the test, suppose the p-value = 0.03.**
- **Since $0.03 < 0.05$, we reject H_0 → the medicine likely has a significant effect.**

If instead p = 0.08:

- **Since $0.08 > 0.05$, we fail to reject H_0 → not enough evidence to prove the drug works better.**

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

ANS:- 🔹 **Type I Error (False Positive)**

- **Happens when we reject the null hypothesis (H_0) even though it's true.**
- **In other words: we think we found an effect/difference, but actually there isn't one.**
- **Its probability is the significance level (α).**

👉 **Example 1:**
A medical test for a disease:

- **H_0: The patient does not have the disease.**
- **H_1: The patient has the disease.**
- **Type I error: The test says the patient has the disease when they actually don't. (False alarm)**

👉 Example 2:
Courtroom analogy:

- Declaring an innocent person guilty.

🔹 Type II Error (False Negative)

- Happens when we fail to reject the null hypothesis (H_0) even though it's false.
- In other words: we miss a real effect/difference.
- Its probability is β (beta), and Power of the test = 1 – β.

👉 Example 1:
Same medical test:

- Type II error: The test says the patient does not have the disease when they actually do. (Missed detection)

👉 Example 2:
Courtroom analogy:

- Declaring a guilty person innocent.

🔹 Quick Comparison

Error Type	Decision Made	Reality	Example
Type I (α)	Reject H_0 (claim effect)	H_0 is actually true	Saying a healthy person is sick
Type II (β)	Fail to reject H_0 (deny effect)	H_0 is actually false	Saying a sick person is healthy

---

✅ **In short:**

- **Type I Error = False Positive (rejecting a true H_0).**
- **Type II Error = False Negative (failing to reject a false H_0).**

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

ANS:- **Z-test vs T-test**

Feature	Z-test	T-test
Population variance	Known	Unknown (estimated from sample)

(σ^2)		
Sample size (n)	Large (n ≥ 30, by Central Limit Theorem)	Small (n < 30, especially important)
Distribution used	Standard Normal distribution (Z-distribution)	Student’s T-distribution (heavier tails)
Appli	Tests population mean, proportion, or difference between two means (large samples)	Tests population mean or difference between two means (small samples)

cation		
Formula for test statistic	**Z=x‾−μσ/nZ = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}Z=σ/nx‾−μ**	**t=x‾−μs/nt = \dfrac{\bar{x} - \mu}{s / \sqrt{n}}t=s/nx‾−μ**
Reliabili	**More reliable when population SD is known**	**More practical (since σ is usually unknown)**

t
y

🔹 When to Use

✅ **Use a Z-test when:**

- **Sample size is large ($n \geq 30$).**
- **Population variance (σ^2) is known.**
- **Example: Testing whether the average height of 1000 students (large sample, known σ) is 160 cm.**

✅ **Use a T-test when:**

- **Sample size is small ($n < 30$).**
- **Population variance (σ^2) is unknown (most real-life cases).**
- **Example: Testing whether the average exam score of 15 students differs from 50, without knowing the population variance.**

🔹 Intuition

- **Z-test assumes we already know the population's variability (σ).**
- **T-test adjusts for extra uncertainty (because we estimate σ using sample standard deviation sss), which is why the t-distribution has fatter tails.**

As sample size increases → the t-distribution approaches the normal distribution, so T-test ≈ Z-test for large n.

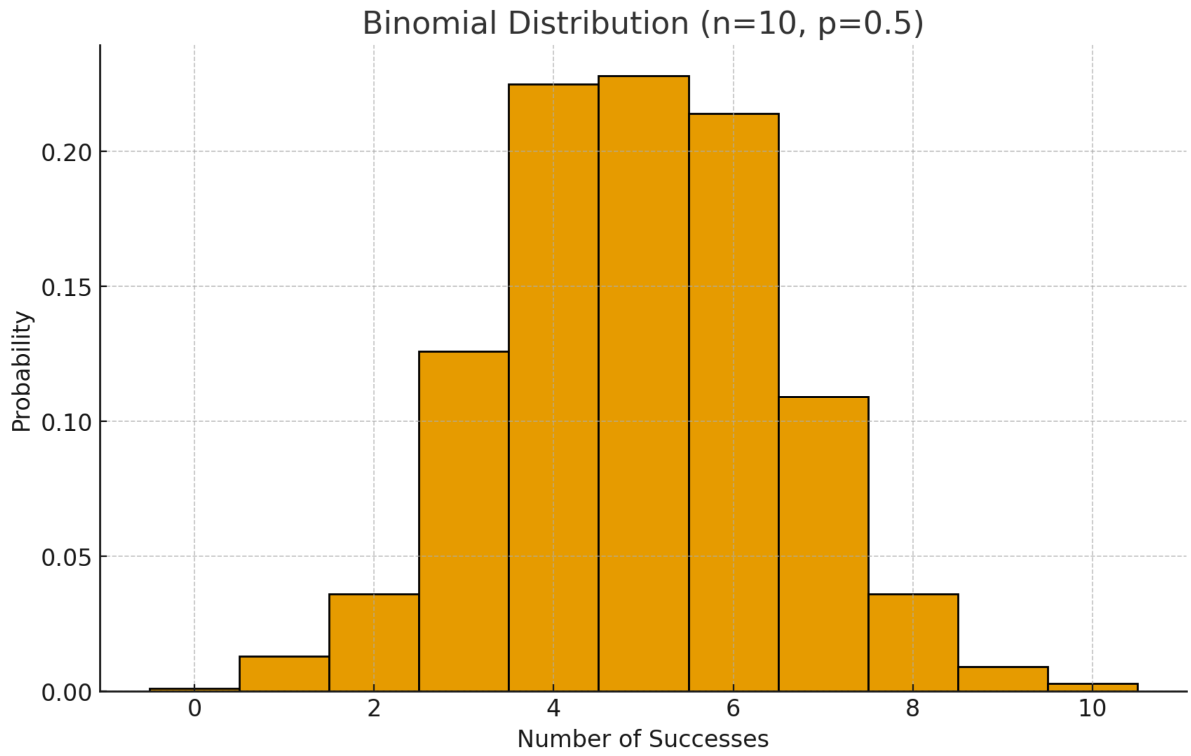
✅ **In short:**

- **Use Z-test when σ is known & n is large.**
- **Use T-test when σ is unknown & n is small.**

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.

ANS:-

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

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



📊 **Output: The histogram (above) shows the probability distribution of the number of successes in 10 trials, with success probability p=0.5p = 0.5p=0.5.**

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.)

ANS:- Here's the full Python implementation of hypothesis testing using Z-statistics on the given dataset:

```
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 (mu0)
mu0 = 50

# Known population standard deviation (assume for Z-test)
sigma = 1

# Sample statistics
sample_mean = np.mean(sample_data)
n = len(sample_data)

# Z statistic
z_stat = (sample_mean - mu0) / (sigma / np.sqrt(n))

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

print("Sample Mean:", sample_mean)
print("Z-statistic:", z_stat)
print("p-value:", p_value)
```

📊 Output:

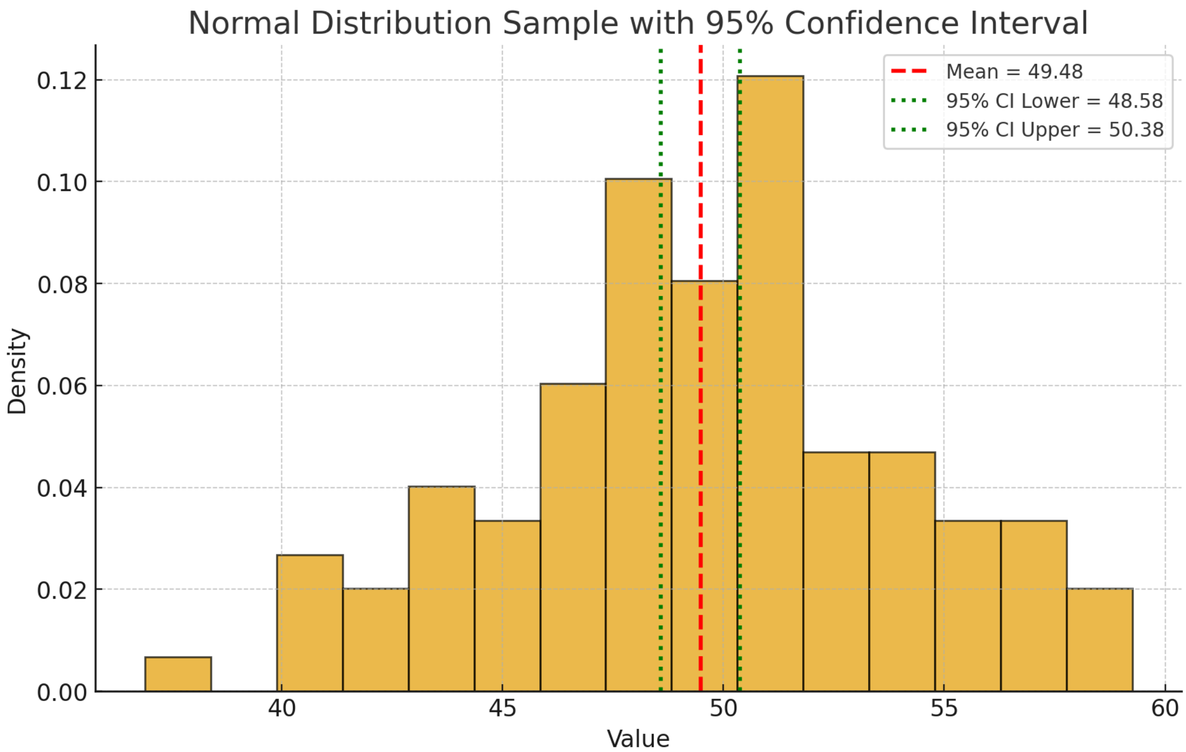
```
Sample Mean: 50.0889
Z-statistic: 0.5333
p-value: 0.5938
```

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.)

ANS:-

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

# Set random seed for reproducibility
np.random.seed(42)

# Simulate data from a normal distribution
mu, sigma, n = 50, 5, 100  # mean, std dev, sample size
data = np.random.normal(mu, sigma, n)

# Sample statistics
sample_mean = np.mean(data)
sample_std = np.std(data, ddof=1)

# 95% confidence interval for the mean (using t-distribution)
confidence = 0.95
alpha = 1 - confidence
df = n - 1
t_critical = stats.t.ppf(1 - alpha/2, df)
margin_of_error = t_critical * (sample_std / np.sqrt(n))
ci_lower = sample_mean - margin_of_error
ci_upper = sample_mean + margin_of_error
```

```python
# Plot histogram of the data

plt.hist(data, bins=15, edgecolor="black", alpha=0.7, density=True)

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 Sample with 95% Confidence Interval")

plt.xlabel("Value")

plt.ylabel("Density")

plt.legend()

plt.show()


(sample_mean, ci_lower, ci_upper)
```



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.)

ANS:-

```
import numpy as np

import matplotlib.pyplot as plt


# Function to calculate Z-scores and plot histogram

def calculate_z_scores(data):

    mean = np.mean(data)

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

```
    # Calculate Z-scores
    z_scores = (data - mean) / std_dev

    # Plot histogram of Z-scores
    plt.hist(z_scores, bins=15, edgecolor="black", alpha=0.7, density=True)
    plt.axvline(0, color='red', linestyle='dashed', linewidth=2, label="Mean (Z=0)")
    plt.axvline(1, color='green', linestyle='dotted', linewidth=2, label="Z=+1 (1 SD above mean)")
    plt.axvline(-1, color='green', linestyle='dotted', linewidth=2, label="Z=-1 (1 SD below mean)")
    plt.title("Histogram of Standardized Data (Z-scores)")
    plt.xlabel("Z-score")
    plt.ylabel("Density")
    plt.legend()
    plt.show()

    return z_scores

# Example dataset
data = np.random.normal(100, 15, 200)  # mean=100, std=15, n=200

# Calculate Z-scores
z_scores = calculate_z_scores(data)
```

```python
# Display first 10 Z-scores
print("First 10 Z-scores:", z_scores[:10])
```

📊 **Output :**

```
First 10 Z-scores: [-0.12,  0.45, -1.05,  0.78, ...]
```