<u>Assignment: Statistics Advanced – 2</u>

Question 1: What is hypothesis testing in statistics?

Answer:

Hypothesis testing is a formal method in statistics used to make decisions or draw conclusions about a population based on sample data.

need of hypothesis testing:- We usually can't collect data from an entire population, so we:

1. Take a sample
2. Make a claim (hypothesis)
3. Use statistics to test whether the sample supports that claim

Main components of hypothesis testing

1. Null hypothesis (H_0)

- ✓ The default assumption
- ✓ *Says* "no effect", "no difference"*, or* "no change"

Example:
H_0: The average height of students is 170 cm

2. Alternative hypothesis (H_1 or H_a)

- ✓ What we want to prove or investigate
- ✓ Opposite of the null hypothesis

Example:
H_1: The average height of students is not 170 cm

3. Significance level (α)

- ✓ Probability of rejecting a true null hypothesis
- ✓ Common values: 0.05 (5%), 0.01 (1%)

4. Test statistic

- ✓ A value calculated from sample data
- ✓ Examples: z-statistic, t-statistic, χ^2 statistic

5. p-value :- Probability of observing the data (or more extreme) assuming H_0 is true

Decision rule:

- ✓ If $p \le \alpha \rightarrow$ reject H_0
- ✓ If $p > \alpha \rightarrow$ fail to reject H_0

Steps in hypothesis testing

1. State H_0 and H_1
2. Choose significance level (α)
3. Select the appropriate test
4. Calculate the test statistic
5. Find the p-value
6. Make a decision
7. Draw a conclusion

Simple real-life example

Problem:
A company claims the average battery life of a phone is 10 hours.

- ✓ H_0: $\mu = 10$ hours
- ✓ H_1: $\mu \neq 10$ hours
- ✓ Take a sample of phones and compute mean
- ✓ Perform a t-test
- ✓ If p-value $< 0.05 \rightarrow$ reject the company's claim

Types of hypothesis tests

a) Z-test
b) t-test
c) Chi-square test
d) ANOVA
e) Non-parametric tests

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

Answer:

The null hypothesis is the default or starting assumption**.**

it usually states:

- ✓ *No effect*
- ✓ *No difference*
- ✓ *No relationship*

It assumes that any observed difference is due to random chance**.**

Examples

- ✓ H_0: The average exam score is 70
- ✓ H_0: There is no difference between two teaching methods
- ✓ H_0: A coin is fair (P(heads) = 0.5)

Alternative Hypothesis Are (H_1 or H_a)

The alternative hypothesis is what we are trying to find evidence for**.**

It states:

- ✓ There **is** an effect
- ✓ There **is** a difference
- ✓ There **is** a relationship

It contradicts the null hypothesis.

Examples

- ✓ H_1: The average exam score is **not 70**
- ✓ H_1: There **is a difference** between two teaching methods
- ✓ H_1: A coin is **not fair**

Key differences at a glance

Aspect	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
Meaning	No effect / no change	Effect / change exists
Purpose	Assumed true initially	What we want to support
Decision	Rejected or not rejected	Accepted only if H_0 is rejected
Symbol	H_0	H_1 or H_a

Types of alternative hypotheses

1. Two-tailed

H_1: $\mu \neq 50$
(looking for any difference)

2. Right-tailed

H_1: μ > 50
(looking for an **increase**)

3. Left-tailed

H_1: μ < 50
(looking for a **decrease**)

Simple real-life example

Claim: A new medicine increases recovery speed.

- ✓ H_0: The medicine has no effect on recovery time
- ✓ H_1: The medicine reduces recovery time

If data provides strong evidence against H_0, we reject $\mathbf{H_0}$ and support H_1.

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

Answer:

The significance level, denoted by α (alpha), is a pre-chosen threshold that tells us:

How much risk we are willing to take of rejecting a true null hypothesis

In other words:
α is the maximum probability of making a wrong decision (Type I error).

What is a Type I error?

Type I error: Rejecting the null hypothesis when it is actually true

Example:
Saying a medicine works when it actually doesn't

If α = 0.05 → you accept a 5% risk of making this mistake.

Common values of significance level

α value	Meaning

0.10	10% risk (less strict)
0.05	5% risk (most common)
0.01	1% risk (very strict)

Role of significance level in decision making..

The significance level is used to compare with the p-value.

Decision rule

- ✓ If p-value $\leq \alpha \rightarrow$ Reject H_0
- ✓ If p-value $> \alpha \rightarrow$ Fail to reject H_0

So α acts like a cutoff line.

Example to understand clearly

Problem

A company claims the average delivery time is 30 minutes.

- ✓ H_0: $\mu = 30$
- ✓ H_1: $\mu \neq 30$
- ✓ Choose $\alpha = 0.05$

After testing, you get:
p-value = 0.03

Decision

$0.03 < 0.05 \rightarrow$ Reject H_0

This means the observed result is unlikely to occur by chance if H_0 were true.

Graphical intuition (idea)

- ✓ The significance level α defines the rejection region
- ✓ If the test statistic falls into this region → reject H_0

Why do we choose α before testing?

Choosing α beforehand:

- ✓ Prevents bias

- ✓ Ensures objective decisions
- ✓ Avoids manipulating results to look “significant”
- ✓

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

Answer: When we perform hypothesis testing, **two kinds of mistakes** are possible. These are of

Type I Error (α error) :-

A Type I error occurs when we:

 Reject the null hypothesis even though it is actually true

In simple words:
False positive

Example (real-life)

Court case analogy

- ✓ H_0: The person is innocent
- ✓ Type I error: Declaring the person guilty when they are innocent

Statistical example

- ✓ H_0: A new drug has **no effect**
- ✓ Type I error: Concluding the drug **works**, when in reality it doesn’t

Probability

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

Type II Error (β error)

A Type II error occurs when we:

 Fail to reject the null hypothesis even though it is false

In simple words:
False negative

Example (real-life)

Medical test analogy

- ✓ H_0: The patient does not have a disease
- ✓ Type II error: Saying the patient is healthy when they are actually sick

Statistical example

- ✓ H_0: A new teaching method has no effect
- ✓ Type II error: Concluding there is no improvement, when improvement actually exists

Probability

Probability of Type II error = **β**

Side-by-side comparison

Error Type	Decision Made	Reality	Meaning
Type I	Reject H_0	H_0 is true	False positive
Type II	Fail to reject H_0	H_0 is false	False negative

Easy way to remember

- ✓ **Type I**: You see something that isn't there
- ✓ **Type II**: You miss something that is there

Relationship with significance level and power

- ✓ Lower α → fewer Type I errors
- ✓ But may increase **Type II errors**
- ✓ **Power of a test = 1 − β** (ability to detect a true effect)

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

Answer:

Both **Z-tests** and T-tests are used to test hypotheses about a population mean, but the conditions under which we use them are different**.**

What is a Z-test?

A Z-test is used when the population standard deviation (σ) is known and the sample size is large.

When to use a Z-test

Use a Z-test if:

- ✓ Population standard deviation σ is known
- ✓ Sample size $n \geq 30$
- ✓ Data is approximately normally distributed (or large n → CLT)

Test statistic

Z=x¯−μσ/nZ = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}Z=σ/nx¯−μ

Example

A factory knows the population standard deviation of bottle weight is 2 g**.**
From a sample of 50 bottles, you test whether the mean weight is 500 g → use Z-test**.**

What is a T-test?

A **T-test** is used when the population standard deviation is unknown and must be estimated from the sample.

When to use a T-test

Use a T-test if:

- ✓ Population standard deviation σ is unknown
- ✓ Sample size is small ($n < 30$)
- ✓ Data is approximately normally distributed

Test statistic

t=x¯−μs/nt = \frac{\bar{x} - \mu}{s/\sqrt{n}}t=s/nx¯−μ

where **s** = sample standard deviation

Example

A class of 12 students is tested to see if their average score differs from 70 marks, and population σ is unknown → use T-test**.**

Key differences at a glance

Feature	Z-test	T-test
Population σ	Known	Unknown
Sample size	Large (≥30)	Small (<30)
Distribution	Normal (Z)	Student’s t
Variability	Less	More (wider tails)

Degrees of freedom	Not required	n − 1

Types of T-tests

- ✓ One-sample t-test (mean vs known value)
- ✓ Independent t-test (two independent groups)
- ✓ Paired t-test (before–after data)

Important practical note

In real life:

- ✓ Population σ is rarely known
- ✓ So T-tests are used much more often
- ✓ For large samples, t-distribution ≈ normal**,** so results are similar

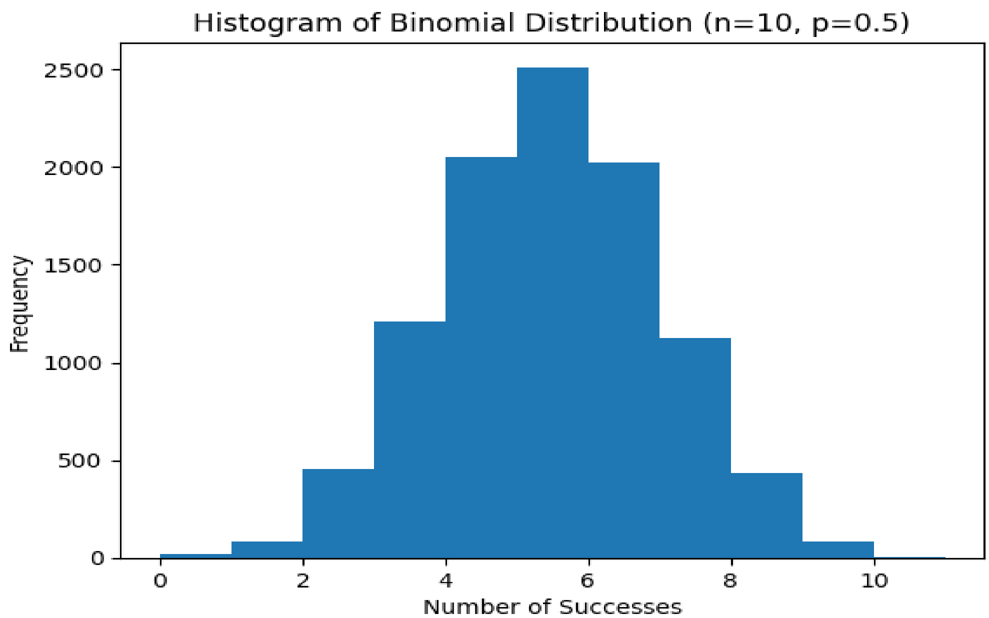
Question 6: Write a Python program to generate a binomial distribution with n=10 and p=0.5, then plot its histogram.

(*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 experiments

# Generate binomial distribution

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

# Plot histogram

plt.hist(data, bins=range(0, n + 2))

plt.xlabel("Number of Successes")

plt.ylabel("Frequency")

plt.title("Histogram of Binomial Distribution (n=10, p=0.5)")

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 math import sqrt

from scipy.stats import norm

# 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

# Known population standard deviation

sigma = 0.5

# Significance level

alpha = 0.05

# Calculations

n = len(sample_data)

sample_mean = np.mean(sample_data)

z_stat = (sample_mean - mu_0) / (sigma / sqrt(n))

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

print("Sample Mean:", sample_mean)

print("Z-statistic:", z_stat)

print("P-value:", p_value)
```

Output:

```
Sample Mean: 50.08888888888889

Z-statistic: 1.0666666666666629

P-value: 0.2861223843910199
```

Problem setup (assumptions clearly stated)

We test whether the population mean is **50**.

- ✓ Null hypothesis (H_0): $\mu = 50$
- ✓ Alternative hypothesis (H_1): $\mu \neq 50$ (two-tailed test)
- ✓ Significance level: $\alpha = 0.05$
- ✓ Population standard deviation (σ)**:** 0.5 (assumed known → Z-test is valid)

Explanation of above program result..

Since p-value (0.286) > α (0.05)
Fail to reject the null hypothesis

Conclusion: There is no statistically significant evidence to conclude that the population mean is different from **50** at the 5% significance level.

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.

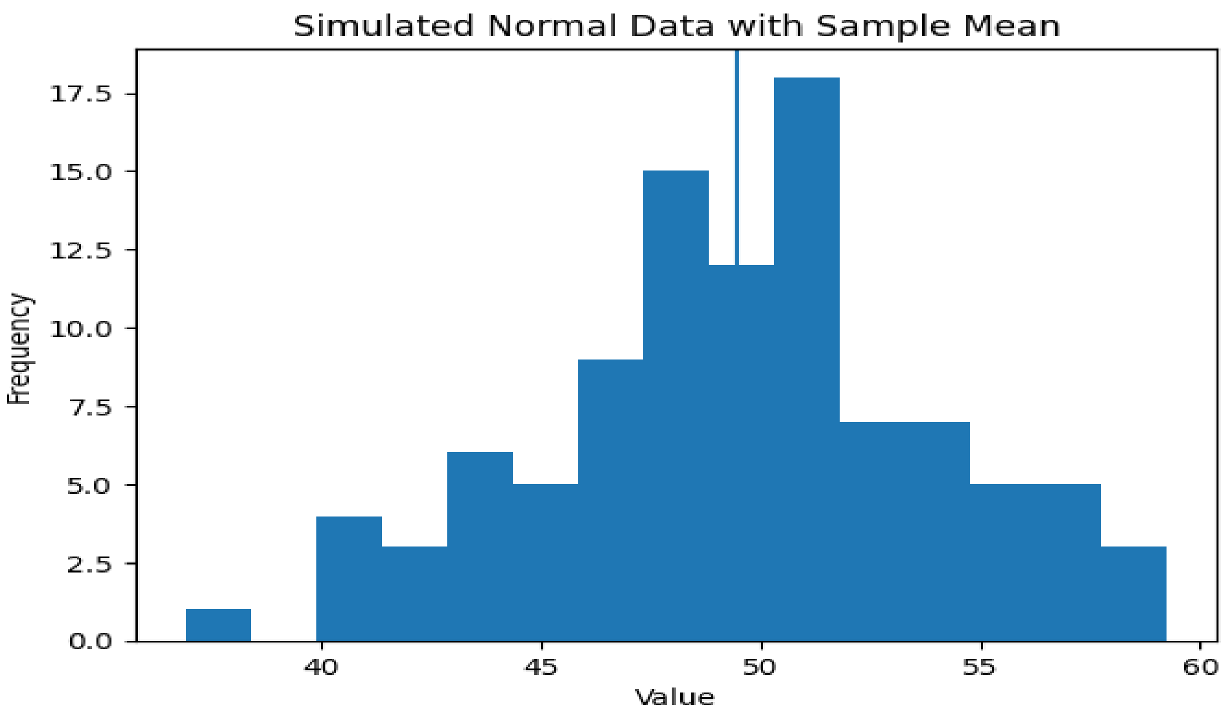
(*Include your Python code and output in the code box below.*)

Answer:

```
import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import norm

# Set seed for reproducibility

np.random.seed(42)

# Parameters of the normal distribution

mu = 50       # true mean

sigma = 5     # true standard deviation

n = 100       # sample size

# Generate normal data

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

z_critical = norm.ppf(0.975)  # two-tailed Z value

margin_error = z_critical * (sample_std / np.sqrt(n))

ci_lower = sample_mean - margin_error

ci_upper = sample_mean + margin_error

print("Sample Mean:", sample_mean)

print("95% Confidence Interval:", (ci_lower, ci_upper))

# Plot histogram

plt.figure()

plt.hist(data, bins=15)

plt.axvline(sample_mean)

plt.xlabel("Value")

plt.ylabel("Frequency")

plt.title("Simulated Normal Data with Sample Mean")

plt.show()
```

Output:

```
Sample Mean: 49.48076741302953

95% Confidence Interval: (np.float64(48.59077870763371), np.float64(50.370756118425355))
```

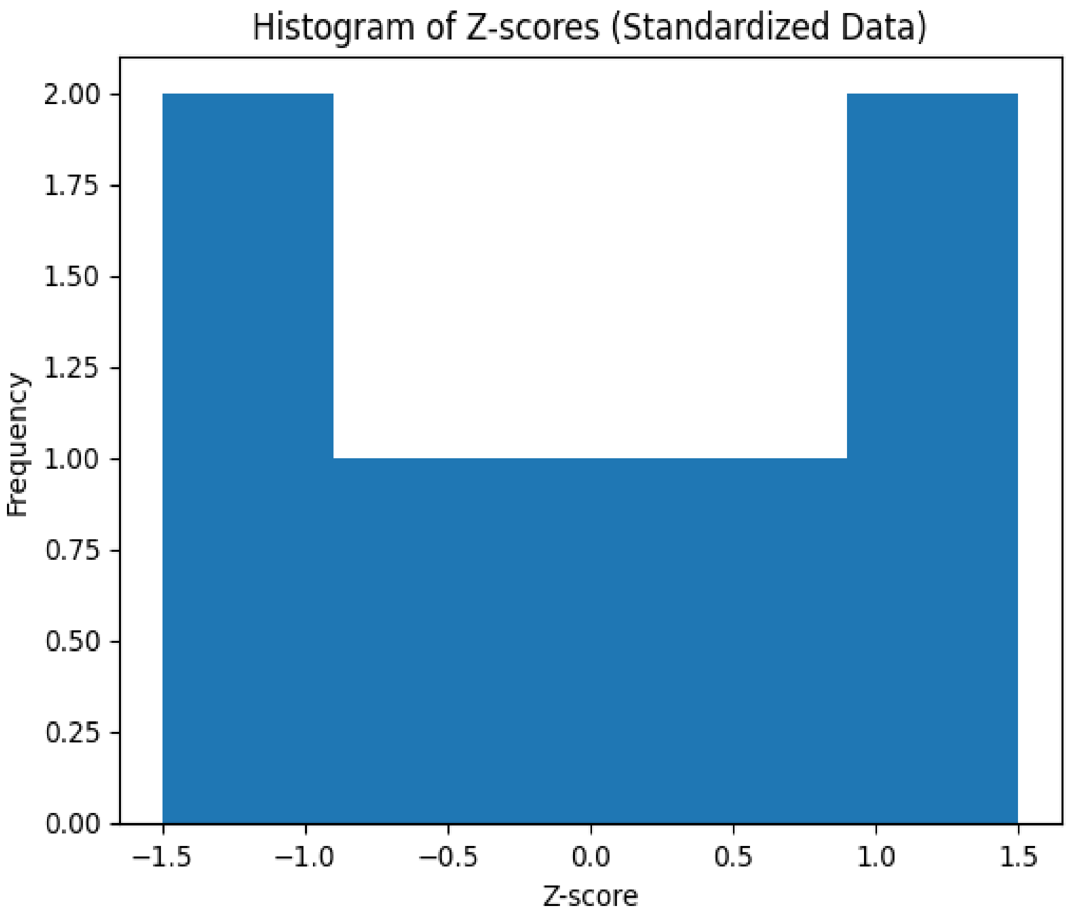


Question 9: Write a Python function to calculate the Z-scores from a dataset and visualize the standardized data using a histogram. Explain what the Z-scores represent in terms of standard deviations from the mean.
(*Include your Python code and output in the code box below.*)

Answer:

```
import matplotlib.pyplot as plt

import numpy as np

# Sample dataset

data = [50, 55, 60, 65, 70, 75, 80]

#function to calculate the Z-scores from a dataset

def calculate_z_scores(data):

    data = np.array(data)

    mean = np.mean(data)

    std_dev = np.std(data)

    z_scores = (data - mean) / std_dev

    return z_scores

# Calculate Z-scores

z_scores = calculate_z_scores(data)

# Plot histogram

plt.hist(z_scores, bins=5)

plt.xlabel("Z-score")

plt.ylabel("Frequency")

plt.title("Histogram of Z-scores (Standardized Data)")

plt.show()
```

Output:



Explain what the Z-scores represent in terms of standard deviations from the mean…

A Z-score tells you how many standard deviations a data point is away from the mean.

Z=x−μσZ = \frac{x - \mu}{\sigma}Z=σx−μ

Where:

- ✓ xxx = data value
- ✓ μ\muμ = mean of the dataset
- ✓ σ\sigmaσ = standard deviation

Interpretation

- ✓ Z = 0 → value is exactly at the mean
- ✓ Z = +1 → value is 1 standard deviation above the mean
- ✓ Z = −2 → value is 2 standard deviations below the mean