Statistics Advanced - 2

Question 1: What is hypothesis testing in statistics?
Ans. **Hypothesis Testing in Statistics**

Hypothesis testing is a statistical method used to make decisions or inferences about a population parameter based on sample data.

- It begins with an **assumption (hypothesis)** about the population.

- Then we use sample data and probability theory to decide whether there is enough evidence to **reject that assumption** or not.

Steps in Hypothesis Testing

1. **State the hypotheses**

    - **Null Hypothesis (H0H_0H0)**: The default assumption (no effect, no difference).

    - **Alternative Hypothesis (H1H_1H1 or HaH_aHa)**: What we want to test/prove (there is an effect or difference).

2. **Choose a significance level (α\alphaα)**

    - Common choices: 0.05, 0.01

- It represents the probability of rejecting H0H_0H0 when it is actually true (Type I error).

3. **Select a test statistic**
    - Depends on data type and distribution (z-test, t-test, chi-square, etc.).
4. **Compute the p-value**
    - The probability of obtaining results as extreme as (or more extreme than) the observed data, assuming H0H_0H0 is true.
5. **Make a decision**
    - If **p-value ≤ α** → Reject H0H_0H0.
    - If **p-value > α** → Fail to reject H0H_0H0.

Example

Suppose a factory claims that the average weight of a sugar packet is **1 kg**.

- H0H_0H0: μ = 1 kg
- H1H_1H1: μ ≠ 1 kg

- Take a sample, compute test statistic and p-value.
- If p-value < 0.05 → evidence suggests the average weight is not 1 kg.

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

Ans. **Null Hypothesis (H0H_0H0)**

- The **default assumption** or starting point in hypothesis testing.
- States that there is **no effect, no difference, or no relationship** in the population.
- It is the claim we try to **disprove or reject**.

✅ Example:
A company claims the average weight of chips in a packet is 100 g.

- H0:μ=100H_0: \mu = 100H0:μ=100

Alternative Hypothesis (H1H_1H1 or HaH_aHa)

- Represents what we want to test or find evidence for.
- States that there **is an effect, difference, or relationship**.
- If evidence against H0H_0H0 is strong enough, we **accept HaH_aHa as more plausible**.

✅ Example:

- H1:μ≠100H_1: \mu \neq 100H1:μ=100 (two-tailed test, mean is not 100).
- Or H1:μ>100H_1: \mu > 100H1:μ>100 (one-tailed test, mean is greater than 100).

Key Differences

Feature	Null Hypothesis (H0H_0H0)	Alternative Hypothesis (H1H_1H1)
Meaning	Default claim, "no effect"	Competing claim, "effect exists"
Symbol	H0H_0H0	H1H_1H1 or HaH_aHa
Goal	Try to **reject** it	Supported if H0H_0H0 is rejected
Example	μ=100\mu = 100μ=100	μ≠100\mu \neq 100μ=100, μ>100\mu > 100μ>100, or μ<100\mu < 100μ<100

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

Ans. **Significance Level (α\alphaα) in Hypothesis Testing**

- The **significance level** is the probability of rejecting the null hypothesis (H0H_0H0) when it is actually true.

- In other words, it is the **tolerated risk of making a Type I error** (false positive).

It is usually set **before** conducting the test.

Common Values of α\alphaα:

- α=0.05\alpha = 0.05α=0.05 (5%) → most common
- α=0.01\alpha = 0.01α=0.01 (1%) → stricter
- α=0.10\alpha = 0.10α=0.10 (10%) → more lenient

Role in Deciding the Outcome

1. **Compare p-value with α\alphaα:**

- If **p-value ≤ α\alphaα** → Reject H0H_0H0 (evidence supports H1H_1H1).
- If **p-value > α\alphaα** → Fail to reject H0H_0H0 (not enough evidence).

2. **Defines the Rejection Region:**

- For a z-test with α=0.05\alpha = 0.05α=0.05, the critical values are approximately ±1.96.
- If the test statistic falls beyond this range, reject H0H_0H0.

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

Ans. **Type I Error (α\alphaα)**

- Occurs when we **reject the null hypothesis (H0H_0H0) even though it is true**.
- It's a **false positive**.
- Probability of making a Type I error = **significance level (α\alphaα)**.

✅ **Example:**
A medical test for a disease:

- H0H_0H0: The patient does **not** have the disease.
- H1H_1H1: The patient has the disease.
- Type I error → The test says the patient has the disease when they actually don't.

Type II Error (β\betaβ)

- Occurs when we **fail to reject the null hypothesis (H0H_0H0) even though it is false**.
- It's a **false negative**.

- Probability of making a Type II error = β\betaβ.
- The power of a test = 1−β1 - \beta1−β.

✅ **Example:**
Same medical test:

- H0H_0H0: The patient does not have the disease.
- H1H_1H1: The patient has the disease.
- Type II error → The test says the patient is healthy when they actually have the disease.

Comparison Table

Error Type	What Happens?	Real-life Meaning	Probability
Type I	Reject H0H_0H0 when it is true	False alarm (detecting something that isn't there)	α\alpha α
Type II	Fail to reject H0H_0H0 when it is false	Missed detection (not detecting something that is there)	β\betaβ

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

Ans. **1. Z-Test**

- A hypothesis test based on the **standard normal distribution (Z-distribution)**.
- **Used when:**
    - Population variance (σ2\sigma^2σ2) is **known**, or
    - Sample size is **large** (n≥30n \geq 30n≥30) (by Central Limit Theorem).

✅ **Examples:**

- Testing if the mean height of students = 160 cm, when population standard deviation is known.
- Quality control in manufacturing (large samples).

---

2. T-Test

- A hypothesis test based on the **Student's t-distribution**.
- **Used when:**
    - Population variance (σ2\sigma^2σ2) is **unknown**, and

    - Sample size is **small** (n<30n < 30n<30).

- As nnn increases, the t-distribution approaches the normal distribution.

✅ **Examples:**

- Testing if the average exam score of 15 students = 70 when population variance is unknown.

- Comparing means of two small groups (independent or paired t-test).

Key Differences

Feature	Z-Test	T-Test
Distribution used	Standard Normal (Z)	Student's t
Population variance (σ2\sigma^2σ2)	Known	Unknown

Sample size	Large (n≥30n \geq 30n≥30)	Small (n<30n < 30n<30)
Shape	Fixed bell curve	Flatter & wider (more variability), depends on degrees of freedom
Common uses	Large-sample mean tests, proportion tests	Small-sample mean tests

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

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 samples

Generate binomial distribution data

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

Plot histogramplt.hist(data, bins=np.arange(-0.5, n+1.5, 1), edgecolor='black', alpha=0.7)

```
plt.title(f"Binomial Distribution (n={n}, p={p})")

plt.xlabel("Number of Successes")

plt.ylabel("Frequency")

plt.xticks(range(n+1))  # set x-axis ticks from 0 to n

plt.show()
```

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]

Ans.

```
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: Known population parameters

mu0 = 50        # hypothesized population mean
```

```
sigma = 1       # assumed known population std
alpha = 0.05    # significance level

# Step 2: Sample statistics
sample_mean = np.mean(sample_data)
n = len(sample_data)

# Step 3: Compute Z-statistic
z_stat = (sample_mean - mu0) / (sigma / np.sqrt(n))

# Step 4: Critical Z-value for two-tailed test
z_critical = norm.ppf(1 - alpha/2)

# Step 5: Compute p-value
p_value = 2 * (1 - norm.cdf(abs(z_stat)))

# Output results
print("Sample Mean:", sample_mean)
print("Z-statistic:", z_stat)
print("Critical Z-value:", z_critical)
```

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


# Interpretation
if abs(z_stat) > z_critical:
    print("Reject the null hypothesis (H0).")
else:
    print("Fail to reject the null hypothesis (H0).")
```

Question 8: What is the concept of expected value in a probability distribution?

Ans. **Expected Value (EV) in Probability**

The **expected value** of a random variable is the **long-run average** or **mean outcome** of a probability distribution if the experiment were repeated many times.

It represents the "center" or "balance point" of the distribution.

---

1. For a Discrete Random Variable

If a random variable XXX takes values x1,x2,…,xnx_1, x_2, \dots, x_nx1,x2,…,xn with probabilities p1,p2,…,pnp_1, p_2, \dots, p_np1,p2,…,pn, then

E[X]=∑i=1nxi⋅piE[X] = \sum_{i=1}^n x_i \cdot p_iE[X]=i=1∑nxi⋅pi

✅ **Example:** Rolling a fair die (X=1,2,3,4,5,6X = 1,2,3,4,5,6X=1,2,3,4,5,6):

$E[X]=1+2+3+4+5+66=3.5E[X] = \frac{1+2+3+4+5+6}{6} = 3.5E[X]=61+2+3+4+5+6=3.5$

Meaning: on average, a die roll is 3.5 in the long run.

---

2. For a Continuous Random Variable

If XXX has probability density function f(x)f(x)f(x), then

$E[X]=\int_{-\infty}^{\infty} x \cdot f(x) \, dxE[X] = \int_{-\infty}^{\infty} x \cdot f(x) \, dxE[X]=\int_{-\infty}^{\infty}x \cdot f(x)dx$

✅ **Example:** If $X \sim U(0,1)X \sim U(0,1)X~U(0,1)$ (uniform distribution),

$E[X]=\int_0^1 x \, dx = 0.5E[X] = \int_0^1 x \, dx = 0.5E[X]=\int_0^1xdx=0.5$

Why is Expected Value Important?

1. **Measure of central tendency** – tells us the "average" value.
2. **Decision-making** – used in economics, gambling, insurance, and finance to predict long-term outcomes.
3. **Basis for variance & standard deviation** – both use expected value in their formulas.

Question 9: Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution.

Ans. import numpy as np

import matplotlib.pyplot as plt

```
# Generate 1000 random numbers from Normal(μ=50, σ=5)
data = np.random.normal(loc=50, scale=5, size=1000)

# Compute sample mean and standard deviation
mean_val = np.mean(data)
std_val = np.std(data)

print("Sample Mean:", mean_val)
print("Sample Standard Deviation:", std_val)

# Plot histogram
plt.hist(data, bins=30, edgecolor='black', alpha=0.7)
plt.title("Histogram of Normal Distribution (μ=50, σ=5)")
plt.xlabel("Value")
plt.ylabel("Frequency")

# Draw a vertical line at the computed mean
plt.axvline(mean_val, color='red', linestyle='dashed', linewidth=2,
label=f"Mean = {mean_val:.2f}")
plt.legend()
```

```
plt.show()
```
