**Assignment Code: DA-AG-007 Statistics Advanced - 2 I 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 (Ho):** Assumes no effect or no difference (status quo).

**Alternative Hypothesis (H1):** Assumes there is an effect or difference. Example:

**Ho:** The average salary of employees = 50,000.

**Hl:** The average salary of employees *-:t* 50,000.

**Set the significance level (a):**

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.-test, t-test, chi-square test, ANOVA).

**Calculate the p-value:**

Probability of getting the observed result if Ho is true.

**Make a decision:**

If p-value a *-7* Reject Ho (evidence supports **H,).**

If p-value > a *-7* Fail to reject Ho (not enough evidence against Ho).

**Simple Example:**

Suppose a company claims the average life of its battery is 100 hours.

Ho: µ = 100

**H1:** µ -:t 100

You test with a sample of 30 batteries. If the p-value is 0.02,and a = 0.05 *-7* Reject Ho.

Conclusion: The average battery life is not 100 hours.

**Types of Hypothesis Tests**

1. **Z-Test** *-7* For large samples (n > 30) when population variance is known.
2. **T-Test** *-7* For small samples (n < 30) when variance is unknown.
3. **Chi-Square Test** *-7* For categorical data (goodness of fit, independence).

4. **ANOVA (Analysis of Variance)** *-7* 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 (Ho):**
   * 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.

• Ho: µ = 50,000 (average salary is 50,000).

1. **Alternative Hypothesis (H1 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 Ho in favor of **H,.**

Example:

* **H,:** µ *-:t* 50,000 (average salary is not 50,000).

1. **Key Differences between Null and Alternative Hypothesis**

**Aspect Null Hypothesis (Ho) Alternative Hypothesis (H1)**

Assumes there is an effect or

**Meaning**

**Nature**

**Testing Objective**

**Symbol Example**

Assumes no effect or no difference.

Represents status quo or existing condition.

To be rejected or retained after testing.

Ho

µ = 50,000

difference.

Represents a new claim or research statement.

To be accepted if Ho is rejected.

**H,** (or Ha)

µ *':t* 50,000

1. **Conclusion**

In hypothesis testing, the **null hypothesis (Ho)** assumes no change, while the **alternative hypothesis (H1)** proposes a possible change or difference. The goal of statistical testing is to analyze data and decide whether to reject Ho in favor of **H,.**

**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 a (alpha), is the probability of rejecting the null hypothesis (Ho) when it is actually true.

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

1. Common Values of Significance Level
   * a = 0.05 (5%) *-7* Most commonly used.
   * a = 0.01 (1%) *-7* Stricter, used in highly sensitive experiments.
   * a = 0.10 (10%) *-7* Sometimes used in exploratory studies.
2. Role in Hypothesis Testing
   * The significance level acts as a threshold to decide whether the observed data provides enough evidence to reject Ho.
   * If the p-value S a,we reject Ho *-7* evidence supports the alternative hypothesis.
   * If the p-value > a,we fail to reject Ho *-7* not enough evidence against Ho.
3. Example

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

* + Null Hypothesis (Ho): Cure rate = 90%
  + Alternative Hypothesis (H,): Cure rate *-:t.* 90%
  + Significance level: a = 0.05

If the test gives a p-value = 0.03, since 0.03 < 0.05, we reject Ho. Conclusion: The cure rate is not 90%.

1. 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.
2. Conclusion

The significance level (a) 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 a, researchers can decide whether to accept or reject the null hypothesis with confidence.

Question 4: What are Type Iand Type II errors? Give examples of each. Answer Type Iand 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 Ierror and Type II error.

1. Type I Error (False Positive)
   * A Type Ierror occurs when the null hypothesis (Ho) 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 = a (significance level). Example:

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

* + Reality: Person is healthy (Ho true).
  + Decision: Test says person is sick (Ho rejected).

1. Type II Error (False Negative)
   * A Type II error occurs when the null hypothesis (Ho) 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 = 13 (beta). Example:

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

* + Reality: Person is sick (Ho false).
  + Decision: Test says person is healthy (Ho not rejected).

1. Comparison Table Aspect Type I Error

Definition Rejecting Ho when it is true Also called False Positive

Probability a (significance level)

Type II Error

Failing to reject Ho when it is false False Negative

(power of the test = 1- )

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

1. Conclusion

Both Type Iand Type II errors represent risks in hypothesis testing.

* + Type I error *-7* Concluding something is true when it is not.
  + Type II error *-7* 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.

1. Z-Test
   * The Z-test is used when:

1. The sample size is large (n > 30).

2. The population variance (a2) 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.

1. **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.

1. **Key Differences Between Z-Test and T-Test Aspect Z-Test**

**Sample Size** Large sample (n > 30)

**Population**

**T-Test**

Small sample (n < 30)

**Variance**

**Distribution Used**

Known Unknown

Standard Normal Distribution (Z) Student's t-Distribution

**Curve Shape Use Case**

Narrow,symmetric

Comparing sample mean with population mean when *a* is known

Wider, heavier tails Comparing sample mean with

population mean when *a* is unknown

1. **When to Use Each Test**
   * Use **Z-Test** *-7* when sample size is large **and** population variance is known.
   * Use **T-Test** *-7* when sample size is small **and** population variance is unknown.
2. **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=lO 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·(l-p)n-kP(X = k) = \binom{n}{k} \cdot pAk \cdot (1-p) A{n-k}P(X=k)=(kn

)·pk·(l-p)n-k

In this question:

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

1. **Python Program**

import numpy as np

import matplotlib.pyplot as pit

# 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=lO, p=O.S)") plt.xlabel("Number of Successes")

plt.ylabel("Frequency") pit.show()

1. Explanation of Code
   * np.random.binomial(n, p, size): Generates random samples from a binomial distribution.
   * pit.hist(): Plots a histogram of the generated data.
   * bins=range(n+2): Ensures bins cover all possible outcomes (O to 10).
   * The graph shows how many times each number of successes (0-10) occurs out of 1000 experiments.
2. Expected Output
   * A histogram with outcomes ranging from 0 to 10 successes.
   * The highest frequencies will be around 5 successes, since p=O.S 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 (Ho): µ = SO (population mean is SO).
  + Alternative Hypothesis (H1): µ *-:t* SO (population mean is not SO).
  + Significance level (a) = O.OS.

1. **Python Code**

import numpy as np from scipy import stats

# Sample dataset

sample\_data = [49.1,S0.2, Sl.O, 48.7, SO.S, 49.8, S0.3, S0.7, S0.2, 49.6,

S0.1,49.9, S0.8, S0.4, 48.9, S0.6, so.o, 49.7, S0.2, 49.5,

S0.1,S0.3, S0.4, so.s, so.o, S0.7, 49.3, 49.8, S0.2, S0.9, S0.3, S0.4, so.o, 49.7, so.s, 49.9]

# Convert to numpy array data = np.array(sample\_data)

# Hypothesized population mean mu = SO

# Sample statistics sample\_mean = np.mean(data)

sample\_std = np.std(data, ddof=l) # sample standard deviation n = len(data)

# Z-statistic calculation

z\_stat = (sample\_mean - mu) *I* (sample\_std *I* 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 = O.OS

if p\_value < alpha:

print("Reject Null Hypothesis: There is significant difference from SO.") else:

print("Fail to Reject Null Hypothesis: No significant difference from SO.")

1. **Expected Output (Values may slightly vary)**

Sample Mean: S0.03

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 SO.

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

**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 (Cl)** gives a range of values within which the true population mean is likely to fall.
   * A **95% Cl** means we are 9S% confident that the interval contains the true mean. Formula for Cl of the mean:

Cl=x-±Za/2xsnCI = \bar{x} \pm Z\_{\alpha/2} \times \frac{s}{\sqrt{n}}Cl=x-±Za/2xns where:

* + x-\bar{x}x- = sample mean
  + sss = sample standard deviation
  + nnn = sample size

• Za/2Z\_{\alpha/2}Za/2 = 1.96 for 9S% Cl

1. **Python Script**

import numpy as np

import matplotlib.pyplot as pit from scipy import stats

# Step 1: Simulate data from a normal distribution np.random.seed(42) # for reproducibility

data = np.random.normal(loc=SO, scale=S, size=lOOO) # mean=SO, std=S, n=lOOO

# Step 2: Calculate sample statistics sample\_mean = np.mean(data) sample\_std = np.std(data, ddof=l) n = len(data)

# Step 3: Calculate 9S% confidence interval alpha = O.OS

z\_value = stats.norm.ppf(l - alpha/2) # 1.96 for 9S% Cl margin\_of\_error = z\_value \* (sample\_std *I* 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=SO, std=S, n=lOOO)") plt.xlabel("Value")

plt.ylabel(''Density'')

# Add vertical lines for Cl

plt.axvline(ci\_lower,color='red', linestyle='dashed', linewidth=2, label="95% Cl Lower") plt.axvline(ci\_upper,color='red', linestyle='dashed', linewidth=2, label="95% Cl Upper") plt.axvline(sample\_mean, color='blue', linestyle='solid', linewidth=2, label="Sample Mean")

pit.legend() pit.show()

1. **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.

1. **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 Cl is approximately **(49.47, 50.07),** which includes the true mean (SO).

**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-µcrZ = \frac{X - \mu}{\sigma}Z=crX-µ where:

* XXX = individual data value
* µ\muµ = mean of dataset
* cr\sigmacr = standard deviation of dataset

Interpretation:

* **Z = 0** *-7* data point is exactly at the mean.
* **Z = +1** *-7* data point is 1standard deviation above the mean.
* **Z = -2** *-7* data point is 2 standard deviations below the mean.

1. **Python Function & Visualization**

import numpy as np

import matplotlib.pyplot as pit

# Function to calculate Z-scores def calculate\_z\_scores(data):

mean = np.mean(data)

std = np.std(data, ddof=l) # sample std z\_scores = (data - mean) *I* std

return z\_scores

# Generate sample dataset np.random.seed(42)

data = np.random.normal(loc=l OO, scale=lS, size=SOO) # mean=lOO, std=lS

# Calculate Z-scores

z\_scores = calculate\_z\_scores(data)

# Print first few Z-scores

print("First 10 Z-scores:", z\_scores[:l O])

# Plot histogram of Z-scores

plt.hist(z\_scores, bins=30, edgecolor='black', alpha=0.7, density=True) pit.title(" Histogram of Z-Scores (Standardized Data)")

plt.xlabel("Z-Score") plt.ylabel(''Density'')

# Add reference lines

plt.axvline(O, color='blue', linestyle='solid', linewidth=2, label="Mean (Z=O)") plt.axvline(l,color='red', linestyle='dashed', linewidth=2, label="Z=+l") plt.axvline(-1,color='red', linestyle='dashed', linewidth=2, label="Z=-1") pit.legend()

pit.show()

1. **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=O (mean).
  + Red dashed lines at Z=±l (one standard deviation from the mean).

1. **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).