Statistics Advanced - 2| Assignment

Q1 : What is hypothesis testing in statistics?

Ans:

Hypothesis testing is a formal statistical process used to make inferences or draw conclusions about a population based on sample data. It involves making an initial assumption (the hypothesis) and then using evidence from the data to determine whether that assumption is statistically plausible. It helps researchers decide if observed effects are likely due to chance or if they represent a real phenomenon.

Q2 : What is the null hypothesis, and how does it differ from the alternative hypothesis?

Ans:

The null hypothesis (H_0) is a statement that assumes no effect, no difference, or no relationship exists in the population. It represents the default assumption that any observed difference is due to random variation.

The alternative hypothesis (H_1 or H_a) contradicts the null hypothesis and represents the presence of an effect, difference, or relationship.

Difference:

- The null hypothesis assumes *no change or effect*.
- The alternative hypothesis assumes *a significant change or effect* exists.

Q3 : Explain the significance level in hypothesis testing and its role in deciding the outcome of a test.

Ans:

The significance level (α) is the probability of rejecting the null hypothesis when it is actually true. It represents the threshold for determining statistical significance and is commonly set at 0.05 (5%).

Role in decision-making:

- If the p-value ≤ α, the null hypothesis is rejected.
- If the p-value > α, the null hypothesis is not rejected.

Thus, the significance level controls the risk of making a Type I error.

Q4 : What are Type I and Type II errors? Give examples of each.

Ans:

Type I Error (False Positive): Occurs when we reject a null hypothesis that is actually true.

- *Example:* A medical test indicates a patient has a disease when they are actually healthy.

Type II Error (False Negative): Occurs when we fail to reject a null hypothesis that is actually false.

- *Example:* A medical test indicates a patient is healthy when they actually have the disease.

Q5 : What is the difference between a Z-test and a T-test? Explain when to use each.

Ans:

Z-test: A Z-test is a statistical test used to determine whether there is a significant difference between a sample mean and a population mean when the population standard deviation is known and the sample size is large (n ≥ 30). It is based on the standard normal distribution (Z distribution). Z-tests are commonly used in large-sample situations where the Central Limit Theorem applies.

T-test : A T-test is a statistical test used to determine whether there is a significant difference between a sample mean and a population mean when the population standard deviation is unknown and the sample size is small (n < 30). It is based on the Student's t-distribution, which accounts for additional uncertainty caused by estimating the population standard deviation from the sample.

When to Use Each Test-

- Use a Z-test when the population standard deviation is known and the sample size is large.
- Use a T-test when the population standard deviation is unknown and the sample size is small.

Q6 : Write a Python program to generate a binomial distribution with n=10 and p=0.5, then plot its histogram.

Hint: Generate random number using random function.

Ans:

```
import random
import matplotlib.pyplot as plt

# Generate binomial distribution
n = 10
p = 0.5
data = [sum(1 for _ in range(n) if random.random() < p) for _ in range(1000)]

# Plot histogram
plt.hist(data, bins=range(12), edgecolor='black')
plt.xlabel("Number of Successes")
plt.ylabel("Frequency")
plt.title("Binomial Distribution (n=10, p=0.5)")
plt.show()
```

Output:

The histogram shows the frequency of successes in 10 trials. The distribution is symmetric around 5, which is the expected mean (n × p).

Q7 : 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 import stats

        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]

        population_mean = 50
        population_std = 1
        alpha = 0.05

        sample_mean = np.mean(sample_data)
        z_score = (sample_mean - population_mean) / (population_std /
        np.sqrt(len(sample_data)))
        p_value = 2 * (1 - stats.norm.cdf(abs(z_score)))

        print("Z-score:", z_score)
print("P-value:", p_value)

        -- Since the p-value is greater than 0.05, we fail to reject the null hypothesis. There
        is no statistically significant difference between the sample mean and the
        population mean.
```

Q8 : 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.

Ans:

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

        data = np.random.normal(loc=50, scale=5, size=100)

        mean = np.mean(data)
        std = np.std(data, ddof=1)

        confidence_interval = stats.norm.interval(0.95, loc=mean,
        scale=std/np.sqrt(len(data)))

        print("95% Confidence Interval:", confidence_interval)

        plt.hist(data, bins=20, edgecolor='black')
        plt.title("Normal Distribution Simulation")
        plt.xlabel("Values")
        plt.ylabel("Frequency")
plt.show()
```

Output:

The confidence interval represents the range within which the true population mean lies with 95% confidence.

Q9 : 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.

Ans:

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

def calculate_z_scores(data):
    mean = np.mean(data)
    std = np.std(data)
    z_scores = [(x - mean) / std for x in data]
    return z_scores

# Sample dataset
data = [10, 12, 14, 15, 18, 20, 22, 25]

# Calculate Z-scores
z_scores = calculate_z_scores(data)

# Plot histogram of Z-scores
plt.hist(z_scores, bins=10, edgecolor='black')
plt.xlabel("Z-scores")
plt.ylabel("Frequency")
plt.title("Histogram of Standardized Data (Z-scores)")
plt.show()

print("Z-scores:", z_scores)
```

Explanation:

Z-scores represent how many standard deviations a data point is away from the mean of the dataset.

They are calculated using the formula:

$$Z = \frac{(x - \mu)}{\sigma}$$

Where:

- x = data point
- mu μ = mean of the dataset
- sigma σ = standard deviation

Interpretation:

- A Z-score of 0 means the value is exactly at the mean.
- A positive Z-score indicates the value is above the mean.
- A negative Z-score indicates the value is below the mean.
- Larger absolute Z-scores indicate values farther from the mean.

The histogram of Z-scores shows the standardized distribution, making it easier to compare data points on a common scale and identify outliers.