Assignment Code: DA-AG-007

Statistics Advanced - 2 | Assignment

Total Marks: 100

## Question 1: What is hypothesis testing in statistics?

Answer:  
Hypothesis testing is a statistical method used to make inferences about a population parameter based on a sample of data. It involves:  
1. Making an assumption (the hypothesis).  
2. Using sample data to test whether the assumption is likely to be true.  
3. Making a decision — either rejecting or failing to reject the hypothesis.  
  
It helps determine whether an observed effect or relationship is statistically significant or due to random chance.

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

Answer:  
- Null Hypothesis (H₀): States there is no effect or no difference. It represents the status quo or default assumption.  
 Example: 'The average exam score = 50.'  
  
- Alternative Hypothesis (H₁ or Ha): States there is an effect or difference.  
 Example: 'The average exam score ≠ 50.'  
  
Difference: H₀ assumes no change; H₁ challenges that assumption and looks for evidence of change.

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

Answer:  
- The significance level (α) is the probability of rejecting the null hypothesis when it is actually true (Type I error).  
- Common values: 0.05 (5%), 0.01 (1%).  
- Role:  
 - If p-value ≤ α: Reject H₀ (evidence supports alternative).  
 - If p-value > α: Fail to reject H₀ (not enough evidence).  
  
It acts as a threshold for decision-making in hypothesis testing.

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

Answer:  
- Type I Error (False Positive): Rejecting H₀ when it is true.  
 Example: A medical test says a healthy person has a disease.  
  
- Type II Error (False Negative): Failing to reject H₀ when it is false.  
 Example: A medical test says a sick person is healthy.

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

Answer:  
- Z-test:  
 - Used when population standard deviation is known.  
 - Large sample size (n > 30).  
 - Based on standard normal distribution.  
  
- T-test:  
 - Used when population standard deviation is unknown.  
 - Small sample size (n ≤ 30).  
 - Based on Student’s t-distribution.  
  
Summary: Z-test = large samples, known σ; T-test = small samples, unknown σ.

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

Answer (Python Code):  
```python  
import numpy as np  
import matplotlib.pyplot as plt  
  
# Parameters  
n, p = 10, 0.5  
  
# Generate binomial data  
data = np.random.binomial(n, p, 1000)  
  
# Plot histogram  
plt.hist(data, bins=range(n+2), edgecolor='black', alpha=0.7)  
plt.title("Binomial Distribution (n=10, p=0.5)")  
plt.xlabel("Number of Successes")  
plt.ylabel("Frequency")  
plt.show()  
```

## Question 7: Implement hypothesis testing using Z-statistics for a sample dataset in Python.

Answer (Python Code):  
```python  
import numpy as np  
from scipy import stats  
  
# 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]  
  
# Parameters  
mu0 = 50 # Hypothesized mean  
x\_bar = np.mean(sample\_data)  
sigma = np.std(sample\_data, ddof=1) # sample std  
n = len(sample\_data)  
  
# Z-statistic  
z = (x\_bar - mu0) / (sigma/np.sqrt(n))  
  
# p-value (two-tailed)  
p\_value = 2 \* (1 - stats.norm.cdf(abs(z)))  
  
print("Sample Mean:", x\_bar)  
print("Z-statistic:", z)  
print("p-value:", p\_value)  
```  
Interpretation: Sample mean ≈ 50.1, Z-stat ≈ 0.73, p-value ≈ 0.46 (> 0.05). Fail to reject H₀. No significant difference from mean=50.

## 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.

Answer (Python Code):  
```python  
import numpy as np  
import matplotlib.pyplot as plt  
from scipy import stats  
  
# Generate normal data  
data = np.random.normal(loc=100, scale=15, size=200)  
  
# Sample mean and std  
mean = np.mean(data)  
std\_err = stats.sem(data)  
  
# 95% confidence interval  
ci = stats.t.interval(0.95, len(data)-1, loc=mean, scale=std\_err)  
  
print("Mean:", mean)  
print("95% CI:", ci)  
  
# Plot histogram  
plt.hist(data, bins=20, edgecolor='black', alpha=0.7)  
plt.axvline(mean, color='red', linestyle='dashed', linewidth=2, label=f"Mean={mean:.2f}")  
plt.title("Normal Distribution Sample Data")  
plt.legend()  
plt.show()  
```

## Question 9: Write a Python function to calculate the Z-scores from a dataset and visualize the standardized data using a histogram.

Answer (Python Code):  
```python  
import numpy as np  
import matplotlib.pyplot as plt  
from scipy import stats  
  
# Function to calculate Z-scores  
def calculate\_z\_scores(data):  
 return stats.zscore(data)  
  
# Generate dataset  
data = np.random.normal(50, 10, 100)  
  
# Compute Z-scores  
z\_scores = calculate\_z\_scores(data)  
  
print("First 10 Z-scores:", z\_scores[:10])  
  
# Plot histogram  
plt.hist(z\_scores, bins=20, edgecolor='black', alpha=0.7)  
plt.title("Histogram of Z-scores")  
plt.xlabel("Z-score")  
plt.ylabel("Frequency")  
plt.show()  
```  
  
Explanation: Z-scores show how many standard deviations each data point is from the mean. Standardized data has mean=0 and std=1.