Assignment Code: DA-AG-007

Statistics Advanced - 2| **Assignment**

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

Total Marks: 180

Question 1: What is hypothesis testing in statistics?

Answer:

Hypothesis testing is a statistical method used to make decisions or draw conclusions about a population based on sample data. It involves formulating a null hypothesis and an alternative hypothesis, selecting a significance level, calculating a test statistic, and deciding whether to reject or fail to reject the null hypothesis based on the evidence provided by the data.

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

Answer:

The null hypothesis (H_0) states that there is no effect, no difference, or no relationship in the population and represents the default assumption.
The alternative hypothesis (H_1) states that there is an effect, a difference, or a relationship. Hypothesis testing evaluates whether there is enough statistical evidence to reject the null hypothesis in favor of the alternative hypothesis.

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. Common values are 0.05 or 0.01. If the calculated p-value is less than or equal to α, the null hypothesis is rejected; otherwise, it is not reject

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

Answer:
A Type I error occurs when the null hypothesis is rejected even though it is true.
Example: **Concluding a drug is effective when it is not.**
A Type II error occurs when the null hypothesis is not rejected even though it is false.
Example: **Concluding a drug is ineffective when it actually works.**

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

Answer:
A Z-test is used when the population variance is known and the sample size is large ($n \geq 30$).
A T-test is used when the population variance is unknown and the sample size is small ($n < 30$).
Z-tests assume a normal distribution, while T-tests use the t-distribution to account for small sample sizes.

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 random
import matplotlib.pyplot as plt

n = 10
p = 0.5
trials = 1000

data = []
for _ in range(trials):
    successes = 0
    for _ in range(n):
        if random.random() < p:
            successes += 1
    data.append(successes)

plt.hist(data, bins=range(0, n+2), edgecolor='black')
plt.xlabel("Number of Successes")
plt.ylabel("Frequency")
plt.title("Binomial Distribution (n=10, p=0.5)")
plt.show()
```

Interpretation:
The histogram shows a symmetric distribution centered around 5, which matches the expected behavior of a binomial distribution when p = 0.5.

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 scipy.stats import norm

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]

mu = 50
sigma = 1
alpha = 0.05

sample_mean = np.mean(data)
n = len(data)

z = (sample_mean - mu) / (sigma / np.sqrt(n))
p_value = 2 * (1 - norm.cdf(abs(z)))

print("Sample Mean:", sample_mean)
print("Z-statistic:", z)
print("P-value:", p_value)
```

Interpretation:

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

np.random.seed(0)
data = np.random.normal(50, 5, 100)

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

z = stats.norm.ppf(0.975)
margin = z * (std / np.sqrt(n))

ci_lower = mean - margin
ci_upper = mean + margin

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

plt.hist(data, bins=15, edgecolor='black')
plt.axvline(mean, linestyle='dashed')
plt.title("Normal Distribution Data")
plt.xlabel("Values")
plt.ylabel("Frequency")
plt.show()
```

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 numpy as np
import matplotlib.pyplot as plt

data = [45, 48, 50, 52, 55, 47, 49, 51, 53, 54]

mean = np.mean(data)
std = np.std(data)

z_scores = [(x - mean) / std for x in data]

print("Z-scores:", z_scores)

plt.hist(z_scores, bins=8, edgecolor='black')
plt.title("Histogram of Z-scores")
plt.xlabel("Z-score")
plt.ylabel("Frequency")
plt.show()
```

A Z-score indicates how many standard deviations a data point is from the mean. Positive values indicate observations above the mean, negative values indicate observations below the mean, and zero indicates the mean itself.