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 structured procedure to decide, using sample data, whether there is enough statistical evidence to support a claim about a population parameter.
- We begin with a claim (the null hypothesis H_0) and an alternative claim (H_1 or H_a).
- We compute a test statistic from the sample, translate it to a p-value (or compare it with a critical value), and then reject or fail to reject H_0 at a chosen significance level α.
- It controls long-run error rates when making decisions under uncertainty.

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

Answer:

- **Null hypothesis (H_0)**: the default/benchmark claim—typically "no effect," "no difference," or "parameter equals a specific value." Example: H_0: $\mu=50$.
- **Alternative hypothesis (H_1 or H_a)**: the competing claim you want to find evidence for—an effect, a difference, or a deviation from the benchmark. Examples:
    - Two-sided: H_a: $\mu\neq50$
    - Right-tailed: H_a: $\mu>50$
    - Left-tailed: H_a: $\mu<50$
- They differ in direction and purpose: H_0 is tested and potentially rejected; H_a is supported only if evidence against H_0 is strong enough.

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 maximum probability you are willing to tolerate for incorrectly rejecting H_0 when it is true (a Type I error).
- Common choices are 0.10, 0.05, or 0.01.
- We reject H_0 if the p-value $\leq \alpha$.
- Thus, α sets the decision threshold and controls false-alarm risk.

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

Answer:

- **Type I error (α)**: Rejecting a true H_0 (false positive). Example: Concluding a new supplement improves blood pressure when it actually doesn't.
- **Type II error (β)**: Failing to reject a false H_0 (false negative). Example: Concluding the supplement has no effect when it actually does.
- There's a trade-off: lowering α typically increases β unless sample size is increased. The power of a test is 1−β1.

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

Answer:

- **Z-test**: Used for tests on a population mean when the population standard deviation **σ** is known and data are approximately normal. Test statistic $Z=\frac{\bar{X}-\mu 0}{\sigma/\sqrt{n}}$ follows the standard normal under H_0.
- **T-test**: Used when **σ** is unknown and estimated by sample standard deviation s. Test statistic $T=\frac{\bar{X}-\mu 0}{s/\sqrt{n}}$ follows a t-distribution with n−1 df under H_0.
  Rule of thumb: If σ known → Z-test; else → T-test. For small samples, normality matters more for both.

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:

Explanation:
I simulated 10,000 observations and plotted the histogram. The sample mean (~4.986) and variance (~2.513) are close to the theoretical values np=5 and np(1−p)=2.5.

Code:

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

# Parameters
n = 10
p = 0.5
size = 10000
data = np.random.binomial(n, p, size)
print("Sample Mean:", np.mean(data))
print("Sample Variance:", np.var(data))
plt.hist(data, bins=range(n + 2), align='left', edgecolor='black')
plt.title("Binomial Distribution (n=10, p=0.5)")
plt.xlabel("Number of Successes")
plt.ylabel("Frequency")
plt.show()
```

Output:

```
Sample Mean: 5.0035
Sample Variance: 2.5306877500000002
```

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:

Code:

```
import numpy as np
import math
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]

data = np.array(sample_data)
mu0 = 50
sigma = 1
n = len(data)
xbar = np.mean(data)
z_stat = (xbar - mu0) / (sigma / math.sqrt(n))
Phi = lambda z: 0.5 * (1 + math.erf(z / math.sqrt(2)))
p_value = 2 * (1 - Phi(abs(z_stat)))

print("Sample size (n):", n)
print("Sample mean:", round(xbar, 4))
print("Hypothesized mean (H0):", mu0)
print("Z-statistic:", round(z_stat, 4))
print("p-value:", round(p_value, 6))
alpha = 0.05
if p_value < alpha:
    print("Decision: Reject H0 (significant evidence against null hypothesis).")
else:
    print("Decision: Fail to reject H0 (no significant evidence against null hypothesis).")
```

Output:

```
Sample size (n): 36
Sample mean: 50.0889
Hypothesized mean (H0): 50
Z-statistic: 0.5333
p-value: 0.593803
Decision: Fail to reject H0 (no significant evidence against null
hypothesis).
```

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:

Code:

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

# Parameters for the normal distribution
mu = 100       # true mean
sigma = 15     # true standard deviation
n = 500        # sample size
data = np.random.normal(mu, sigma, n)
xbar = np.mean(data)
s = np.std(data, ddof=1)
z_crit = 1.96
margin_error = z_crit * (s / math.sqrt(n))
ci_low = xbar - margin_error
ci_high = xbar + margin_error
print("Sample Mean:", round(xbar, 4))
print("Sample Std Dev:", round(s, 4))
print("95% Confidence Interval: (", round(ci_low, 4), ",", round(ci_high, 4), ")")

# Plot histogram
plt.hist(data, bins=30, edgecolor='black')
plt.title("Simulated Normal Distribution Data")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.show()
```

Output:

```
Sample Mean: 100.4988
Sample Std Dev: 15.3306
95% Confidence Interval: ( 99.155 , 101.8426 )
```

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:

Code:

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

# Function to calculate Z-scores
def calculate_zscores(data):
    data = np.array(data, dtype=float)
    mean = np.mean(data)
    std = np.std(data, ddof=0)
    z_scores = (data - mean) / std
    return z_scores
np.random.seed(42)
data = np.random.normal(loc=100, scale=15, size=500)
z_scores = calculate_zscores(data)
print("Mean of Z-scores:", round(np.mean(z_scores), 4))
print("Std of Z-scores:", round(np.std(z_scores, ddof=0), 4))
plt.hist(z_scores, bins=30, edgecolor='black')
plt.title("Histogram of Z-scores (Standardized Data)")
plt.xlabel("Z-score")
plt.ylabel("Frequency")
plt.show()
```

Output:

```
Mean of Z-scores: -0.0
Std of Z-scores: 1.0
```