Statistics Advanced - 2| Assignment

Question 1: *What is hypothesis testing in statistics?*
Answer: Hypothesis testing is a fundamental statistical inference method used to determine if data from a sample provides sufficient evidence to reject a specific claim, known as the null hypothesis (H0), about a population parameter. It involves comparing this default assumption against an alternative hypothesis (Ha) to make data-driven decisions.

Question 2: *What is the null hypothesis, and how does it differ from the alternative hypothesis?*
Answer: The null hypothesis (H0) is a statistical statement assuming no significant difference, effect, or relationship exists between variables, representing the default "no effect" position. It differs from the alternative hypothesis (Ha), which asserts that a significant effect or difference does exist. H0 usually includes equality (=,<=, >=), while Ha asserts inequality . The null hypothesis is the position to be tested and potentially rejected. The alternative hypothesis is the researcher's proposed explanation or what they are trying to prove.

Question 3: *Explain the significance level in hypothesis testing and its role in deciding the outcome of a test.*
Answer: Significance levels represent the probability of rejecting a true null hypothesis in a statistical test. In simpler terms, they indicate the maximum acceptable risk of concluding that an effect exists when it actually doesn't.

P-value ≤ α	Statistically Significant	The result is unlikely to have happened by chance. You reject the Null Hypothesis.
P-value > α	Not Statistically Significant	The result could easily be a fluke. You fail to reject the Null Hypothesis.

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

1. Type I Error: A Type I error occurs when you reject the null hypothesis (H0) when it is actually true. You conclude that an effect or relationship exists, but in reality, it doesn't.

Example: A medical test tells a patient they have a disease when they are actually perfectly healthy. In a courtroom, this would be convicting an innocent person.

2. Type II Error: A Type II error occurs when you fail to reject the null hypothesis (H0) when it is actually false. You conclude that there is no effect or relationship, but you actually missed a real one.

Example: A medical test tells a patient they are healthy when they actually have an underlying disease. In a courtroom, this would be acquitting a guilty person

Question 5: *What is the difference between a Z-test and a T-test? Explain when to use each.* **Answer:** The main difference between a Z-test and a T-test comes down to how much you actually know about your population and how much data you've collected. While both tests compare means to see if a difference is statistically significant, they use different distributions to get there.

1. The Z-test

The Z-test is the "ideal world" test. It assumes you have a solid grasp of the entire population's parameters.You know the population standard deviation and your sample size is large enough (usually $n \geq 30$) for the Central Limit Theorem to kick in, ensuring a normal distribution.

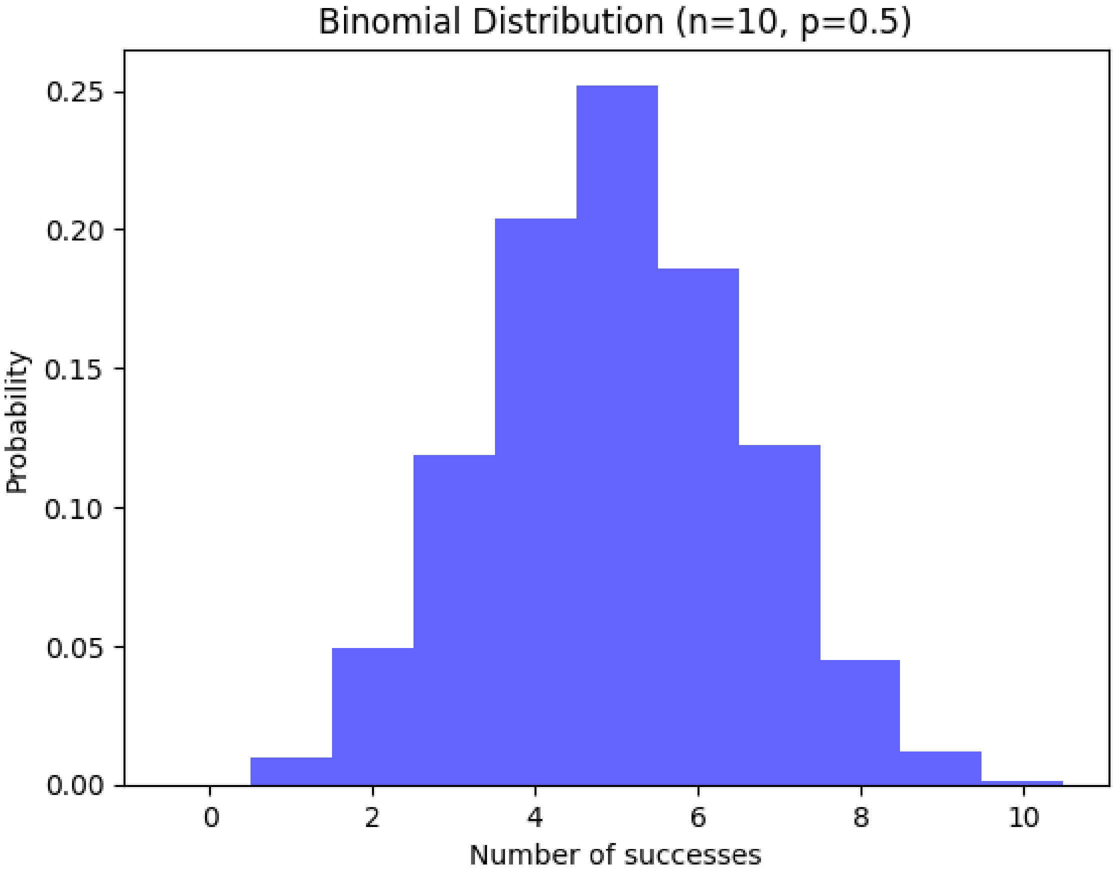
2. The T-test

The T-test is the "real world" workhorse. It is flatter and has "heavier tails" than a normal distribution, which accounts for the extra uncertainty that comes with smaller samples.You do not know the population standard deviation (which is most of the time) and must estimate it using your sample data. It is also mandatory for small sample sizes ($n < 30$).

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

Answer:

```python
import numpy as np
import matplotlib.pyplot as plt
import random
n = 10
p = 0.5
data = np.random.binomial(n, p, 1000)
plt.hist(data, bins=np.arange(-0.5, n+1.5, 1), density=True, alpha=0.6,
color='b')
plt.title('Binomial Distribution (n=10, p=0.5)')
plt.xlabel('Number of successes')
plt.ylabel('Probability')
plt.show()
```



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:

```
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]
mean_sample = np.mean(sample_data)
std_sample = np.std(sample_data, ddof=1)
n = len(sample_data)
```

```
z_statistic = (mean_sample - 50) / (std_sample / np.sqrt(n))
p_value = 2 * (1 - stats.norm.cdf(abs(z_statistic)))
print(f"Z-statistic: {z_statistic:.4f}")
print(f"P-value: {p_value:.4f}")
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis (H0). There is evidence to suggest
that the mean is not equal to 50.")
else:
    print("Fail to reject the null hypothesis (H0). There is not enough
evidence to suggest that the mean is different from 50.")
```

OUTPUT

```
Z-statistic: 0.9940
P-value: 0.3202
Fail to reject the null hypothesis (H0). There is not enough evidence to
suggest that the mean is different from 50.
```

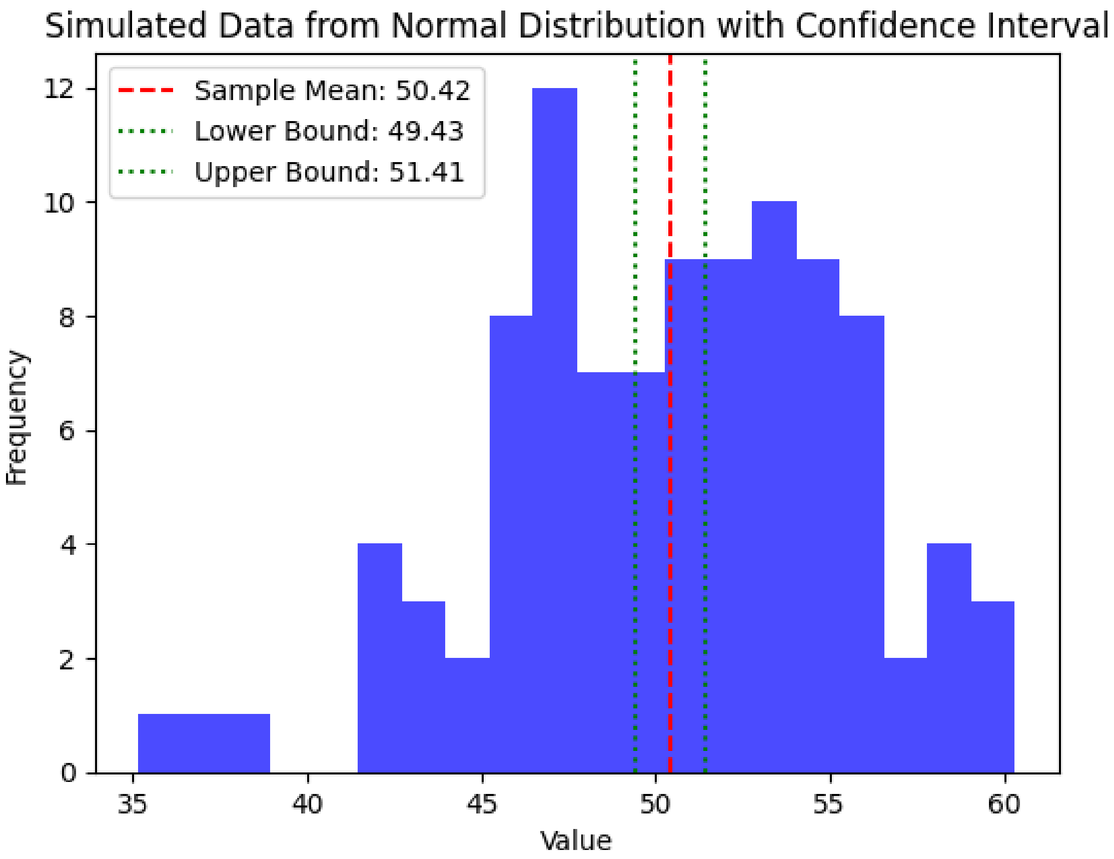
Question 8: *Write a Python script to simulate data from a normal distribution and calculate the 95% confidence interval for its mean. Plot the data using Matplotlib.*

Answer:

```
import numpy as np
import matplotlib.pyplot as plt
# Simulate data from a normal distribution
mean = 50
std_dev = 5
n_samples = 100
simulated_data = np.random.normal(mean, std_dev, n_samples)
# Calculate the 95% confidence interval for the mean
sample_mean = np.mean(simulated_data)
sample_std = np.std(simulated_data, ddof=1)
n = len(simulated_data)
margin_error = 1.96 * (sample_std / np.sqrt(n))  # 1.96 is the Z-score for
95% confidence
lower_bound = sample_mean - margin_error
upper_bound = sample_mean + margin_error
print(f"Sample Mean: {sample_mean:.2f}")
print(f"95% Confidence Interval: [{lower_bound:.2f}, {upper_bound:.2f}]")
# Plot the data
plt.hist(simulated_data, bins=20, color='blue', alpha=0.7)
plt.axvline(sample_mean, color='red', linestyle='--', label=f'Sample Mean:
{sample_mean:.2f}')
plt.axvline(lower_bound, color='green', linestyle=':', label=f'Lower
Bound: {lower_bound:.2f}')
plt.axvline(upper_bound, color='green', linestyle=':', label=f'Upper
Bound: {upper_bound:.2f}')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.title('Simulated Data from Normal Distribution with Confidence
Interval')
plt.legend()
plt.show()
```

OUTPUT

```
Sample Mean: 50.42
95% Confidence Interval: [49.43, 51.41]
```



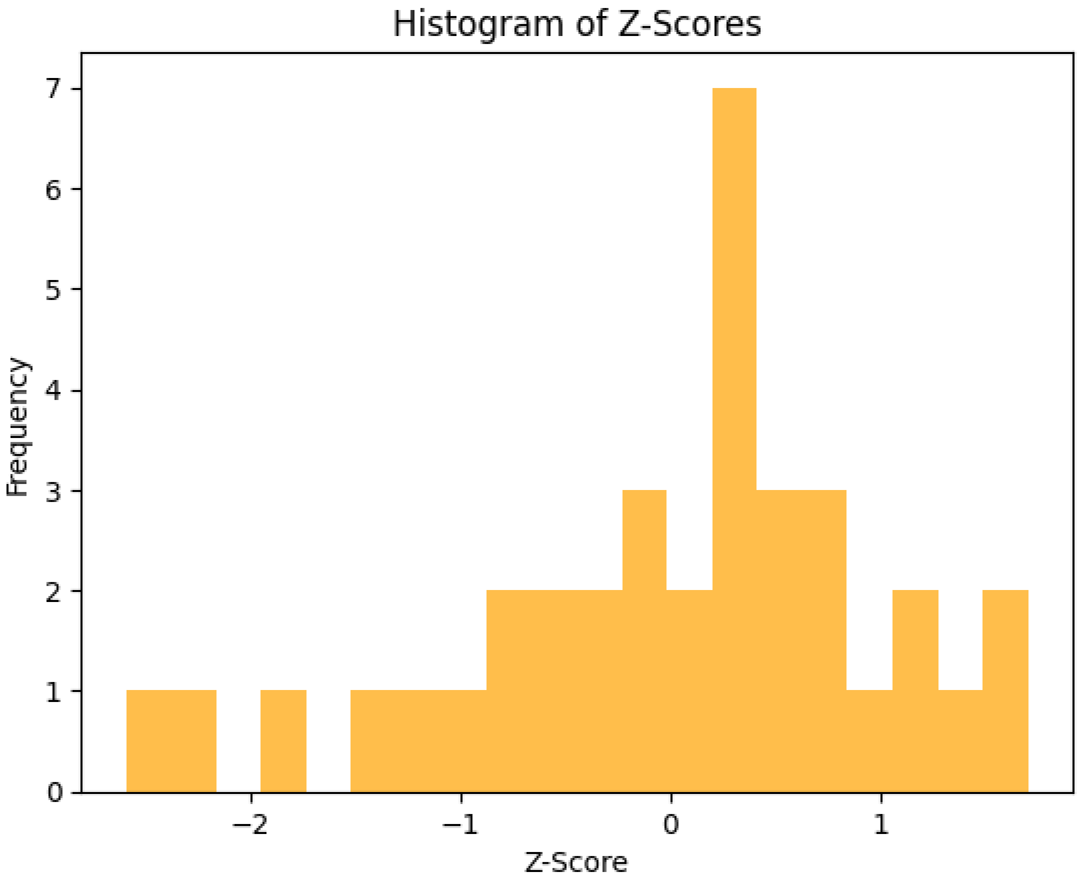
Question 9: *Write a Python function to calculate the Z-scores from a dataset and visualize the standardized data using a histogram. Explain what the Z-scores represent in terms of standard deviations from the mean.*

Answer:

```
import numpy as np
import matplotlib.pyplot as plt
def calculate_z_scores(data):
    mean = np.mean(data)
    std_dev = np.std(data, ddof=1)
    z_scores = (data - mean) / std_dev
    return z_scores

dataset = [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]
z_scores = calculate_z_scores(dataset)
plt.hist(z_scores, bins=20, color='orange', alpha=0.7)
plt.xlabel('Z-Score')
plt.ylabel('Frequency')
plt.title('Histogram of Z-Scores')
plt.show()
```

OUTPUT



A Z-score is a numerical value that tells you exactly how many standard deviations a specific data point is from the mean of its data set.