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 test a claim or statement about a population parameter. It involves making an assumption or claim about a population (e.g., the average age of the Indian population is 45 years) and then using a sample to verify this hypothesis.
The process of hypothesis testing involves:

1. **Framing the hypotheses**: This includes formulating a null hypothesis (H_o) and an alternative hypothesis (H_a).
    - **Null Hypothesis (H_o)**: The initial or default assumption, often stating no effect or no difference (e.g., a person is not guilty of a crime, or the average age is exactly 45 years). It always includes an equality sign.
    - **Alternate Hypothesis (H_a)**: The opposite of the null hypothesis (e.g., a person is guilty, or the average age is not 45 years, or is greater than 45 years).
2. **Statistical analysis**: This involves using statistical methods like p-value and significance level.
3. **Conclusion**: Deciding whether to reject the null hypothesis or fail to reject the null hypothesis.

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

Answer:

The **null hypothesis (H_0)** is the default assumption in hypothesis testing. It usually states that there is **no effect**, **no difference**, or **no relationship** between variables. It **always includes an equality sign**, like =, ≥, or ≤. For example: *H_0: The average age is 45 years.*
The **alternative hypothesis (H_a)** is what we aim to **prove**. It suggests that there **is** an effect, difference, or relationship. It is **opposite** of the null and uses ≠, <, or >. For example: *H_a: The average age is not 45 years.*
It's like a courtroom:

- H_0 = "Not guilty" (assumed true until evidence proves otherwise)
- H_a = "Guilty" (what you try to prove with evidence)

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

Answer:

The **significance level**, denoted as **α (alpha)**, is the **threshold** we set **before testing** a hypothesis. It tells us the **maximum probability of making a Type I error** — that is, **rejecting the null hypothesis (H_0) when it is actually true**.

Common α values:

- **0.05** (5%) → Most common
- **0.01** (1%) → Very strict
- **0.10** (10%) → Less strict

Role in Decision-Making:

- After performing a test, we get a **p-value**.
- We compare **p-value with α**:
  - If **$p \leq \alpha$ → Reject H_0** (result is **statistically significant**)
  - If **$p > \alpha$ → Do not reject H_0**

Example:

Testing a new drug:

- H_0: Drug has no effect
- H_a: Drug has an effect
- $\alpha = 0.05$
- If **p-value = 0.03**, then $0.03 < 0.05$ → **Reject H_0** → Drug likely works

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

Answer:

Type I and Type II Errors in Hypothesis Testing

In hypothesis testing, we make decisions based on sample data — but since we never know the full population truth, **errors** can happen.
There are **two types of errors** we can make:

● **1. Type I Error (False Positive)**

Definition:

Type I error occurs when we **reject the null hypothesis (H_0)** even though it is **actually true**.
In other words, we say there **is an effect or difference**, when in reality, **there isn't**.

Example:

Let's say a company is testing a new drug.

- **H_0 (Null Hypothesis):** The new drug has no effect.
- **H_a (Alternative Hypothesis):** The new drug has a positive effect.

Now suppose the test result leads the researcher to **reject H_0**, and they conclude the drug works. But in reality, the drug **does not work**.
This is a **Type I error** — we made a **false claim** that the drug works when it actually doesn't.

Risk Level:

The probability of making a Type I error is equal to the **significance level (α)**.
So, if **α = 0.05**, we are allowing a **5% chance** of wrongly rejecting a true null hypothesis.

2. Type II Error (False Negative)
Definition:
Type II error happens when we **fail to reject the null hypothesis (H_0)** even though it is **actually false**.
In other words, we **miss detecting a real effect or difference**.
Example:
Using the same drug test:

- **H_0:** The drug has no effect.
- **H_a:** The drug has a positive effect.

Now suppose the test result leads the researcher to **accept H_0**, meaning they say the drug has no effect.
But in reality, the drug **does work**.
This is a **Type II error** — we **missed a real effect**.
Risk Level:
The probability of making a Type II error is denoted by **β (beta)**.
Power of a test = 1 − β (higher power means less chance of Type II error).

Aspect	Type I Error (α)	Type II Error (β)
What happens?	Reject H_0 when it is actually true	Fail to reject H_0 when it is false
Also called	False Positive	False Negative
Risk controlled by	Significance Level (α)	Power (1 – β)
Real-world analogy	Convicting an innocent person	Letting a guilty person go free

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

Answer:

Z-test vs. T-test: A Comprehensive Comparison

Both Z-tests and T-tests are inferential statistical tests used to compare means. They help determine if the difference observed between sample means is statistically significant or due to random chance. While their core purpose is similar, the conditions under which they are applied, and consequently their underlying assumptions, differ significantly.

1. Z-test (Z-statistic)

The Z-test is a hypothesis test that follows a standard normal distribution. It is typically used when you have a large sample size or when the population standard deviation is known.

Key Characteristics and When to Use:

- Known Population Standard Deviation (σ): This is the most crucial criterion. If you know the standard deviation of the entire population from which your sample is drawn, a Z-test is appropriate. This often happens in quality control settings where historical data provides this information.
- Large Sample Size (n > 30): Even if the population standard deviation is unknown, the Central Limit Theorem states that for sufficiently large sample sizes (generally n > 30), the sampling distribution of the mean will approximate a normal distribution. In such cases, the sample standard deviation (s) can be used as a good estimate for the population standard deviation (σ), and a Z-test can be applied. This is why the document you provided emphasizes the n > 30 rule.

- **Normally Distributed Population:** Ideally, the population from which the sample is drawn should be normally distributed. However, due to the Central Limit Theorem, this assumption becomes less critical with larger sample sizes.
- **Comparing Sample Mean to Population Mean:** A common use is to compare a sample mean to a known population mean to see if the sample significantly differs from the population.
- **Hypothesis Testing for Proportions:** Z-tests can also be used for hypothesis testing involving population proportions.

Formula (for a single sample mean):

$Z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$

Where:

- $\bar{x}$ = sample mean
- μ = population mean
- σ = population standard deviation
- n = sample size

Advantages:

- More powerful than a t-test when population standard deviation is known.
- Simpler to calculate if all parameters are met.

Disadvantages:

- Rarely applicable in real-world scenarios where the population standard deviation is unknown.

Examples of Use:

- A shoe manufacturer knows the standard deviation of all shoes produced in the past and wants to test if a new batch of shoes has a significantly different average size.
- A polling company wants to determine if the average age of voters in a sample is significantly different from the known average age of the entire voting population.

2. T-test (T-statistic)

The T-test is a hypothesis test that follows a Student's t-distribution. It is used when the population standard deviation is unknown and the sample size is small.

Key Characteristics and When to Use:

- **Unknown Population Standard Deviation (σ):** This is the defining characteristic. In most real-world research, the population standard deviation is not known. In such cases, we rely on the sample standard deviation (s) to estimate it.
- **Small Sample Size (n < 30):** When the sample size is small, the sample standard deviation may not be a reliable estimate of the population standard deviation. The t-distribution accounts for this increased uncertainty by having "fatter tails" than the normal distribution, leading to larger critical values and wider confidence intervals.
- **Degrees of Freedom:** The t-distribution is characterized by its "degrees of freedom" (df), which is typically n-1 for a single sample t-test. The degrees of freedom influence the shape of the t-distribution; as df increases, the t-distribution approaches the normal distribution.
- **Normally Distributed Population (or approximately):** For small sample sizes, the assumption of a normally distributed population is more important for the validity of the t-test.

Types of T-tests:

- **One-Sample T-test:** Compares the mean of a single sample to a known or hypothesized population mean.
- **Independent Samples T-test (Two-Sample T-test):** Compares the means of two independent groups to determine if there's a significant difference between them.

- **Paired Samples T-test:** Compares the means of two related (dependent) groups, such as before-and-after measurements on the same subjects.

Formula (for a single sample mean):

$t = (\bar{x} - \mu) / (s / \sqrt{n})$

Where:

- $\bar{x}$ = sample mean
- μ = hypothesized population mean
- s = sample standard deviation
- n = sample size

Advantages:

- Applicable in a much wider range of real-world scenarios as population standard deviation is rarely known.
- Accounts for the increased variability associated with small sample sizes.

Disadvantages:

- Less powerful than a z-test if the population standard deviation were actually known.

Examples of Use:

- A new teaching method is introduced to a class of 20 students. A t-test would be used to compare their average test scores to the known average score of students taught by the old method.
- A medical researcher wants to compare the effectiveness of two different drugs on a small group of patients (e.g., 15 patients per group).
- A psychologist wants to see if there's a significant difference in test scores before and after a specific intervention for the same group of individuals.

Summary Table: Z-test vs. T-test

Feature	Z-test	T-test
Population σ	Known	Unknown (estimated by sample s)
Sample Size (n)	Large (n > 30) or any size if σ known	Small (n < 30), or any size if σ unknown
Distribution Used	Standard Normal Distribution	Student's t-distribution
Assumptions	Normal population (less critical for n>30)	Normal population (more critical for small n)
Purpose	Test population mean when σ is known	Test population mean when σ is unknown
Degrees of Freedom	Not applicable (fixed normal distribution)	Applicable (n-1 for one sample)
Conservatism	Less conservative (narrower confidence)	More conservative (wider confidence due to uncertainty)

In essence, the choice between a Z-test and a T-test boils down to the information you have about the population standard deviation and the size of your sample. The T-test is a more robust and widely applicable tool for most practical research scenarios due to the common lack of knowledge regarding population parameters. The Z-test, while theoretically ideal under specific conditions, is less frequently encountered outside of specific quality control or standardized testing environments where population parameters are well-established.

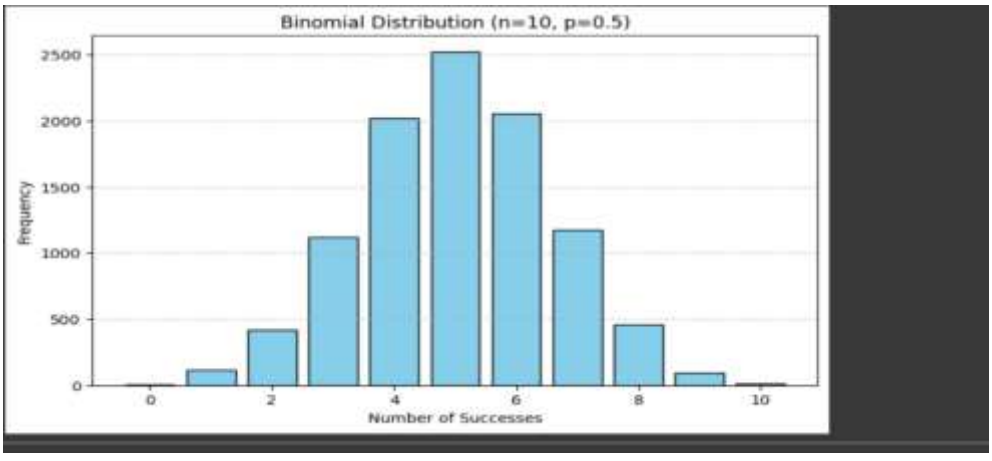
Question 6: Write a Python program to generate a binomial distribution with n=10 and p=0.5, then plot its histogram.

(*Include your Python code and output in the code box below.*)

Hint: Generate random number using random function.

Answer:

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

# Parameters
n = 10      # number of trials
p = 0.5     # probability of success
size = 10000  # number of samples
data = np.random.binomial(n, p, size)

 Ploting t histogram
plt.figure(figsize=(8,5))
plt.hist(data, bins=range(n+2), align='left', rwidth=0.8, color='skyblue', edgecolor='black')
plt.title(f'Binomial Distribution (n={n}, p={p})')
plt.xlabel('Number of Successes')
plt.ylabel('Frequency')
plt.grid(axis='y', linestyle='--', alpha=0.7)
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]

(*Include your Python code and output in the code box below.*)

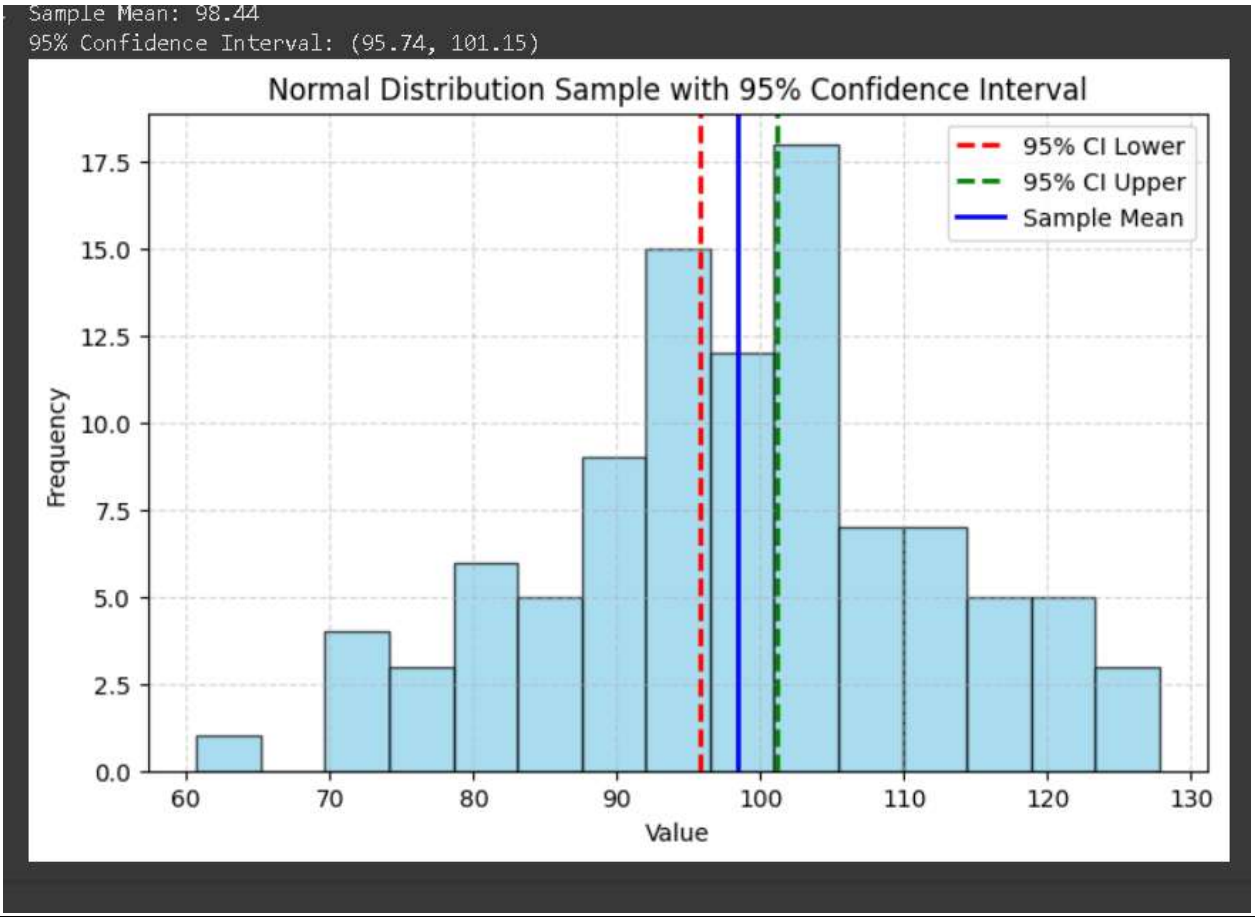
Answer:

```
 import numpy as np
from scipy.stats import norm

# Updated sample data (36 values)
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]

# Null hypothesis mean
mu = 50

# Sample statistics
sample_mean = np.mean(sample_data)
sample_std = np.std(sample_data, ddof=1)  # Sample standard deviation
n = len(sample_data)

# Z-statistic
z = (sample_mean - mu) / (sample_std / np.sqrt(n))

# Two-tailed p-value
p_value = 2 * (1 - norm.cdf(abs(z)))

# Results
print(f"Sample Size: {n}")
print(f"Sample Mean: {sample_mean:.4f}")
print(f"Sample Standard Deviation: {sample_std:.4f}")
print(f"Z-statistic: {z:.4f}")
print(f"P-value: {p_value:.4f}")

# Hypothesis decision
alpha = 0.05
if p_value < alpha:
    print("Conclusion: Reject the null hypothesis (H0). Significant difference exists.")
else:
    print("Conclusion: Fail to reject the null hypothesis (H0). No significant difference.")

output
Sample Size: 36
Sample Mean: 50.0889
Sample Standard Deviation: 0.5365
Z-statistic: 0.9940
P-value: 0.3202
Conclusion: Fail to reject the null hypothesis (Ho). No significant difference.
```

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

#  Simulate data from a normal distribution
np.random.seed(42)  # for reproducibility
mean = 100
std_dev = 15
sample_size = 100

data = np.random.normal(loc=mean, scale=std_dev, size=sample_size)

# Calculate sample mean and standard error
sample_mean = np.mean(data)
standard_error = stats.sem(data)  # SEM = std_dev / sqrt(n)

# Calculate 95% confidence interval
confidence_level = 0.95
confidence_interval = stats.t.interval(confidence_level, df=len(data)-1,
                                       loc=sample_mean, scale=standard_error)

# Print results
print(f"Sample Mean: {sample_mean:.2f}")
print(f"95% Confidence Interval: ({confidence_interval[0]:.2f}, {confidence_interval[1]:.2f})")

#lets we are Plot the histogram of the data
plt.figure(figsize=(8, 5))
plt.hist(data, bins=15, color='skyblue', edgecolor='black', alpha=0.7)
plt.axvline(confidence_interval[0], color='red', linestyle='dashed', linewidth=2, label='95% CI
Lower')
plt.axvline(confidence_interval[1], color='green', linestyle='dashed', linewidth=2, label='95% CI
Upper')
plt.axvline(sample_mean, color='blue', linestyle='solid', linewidth=2, label='Sample Mean')
plt.title('Normal Distribution Sample with 95% Confidence Interval')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.legend()
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```

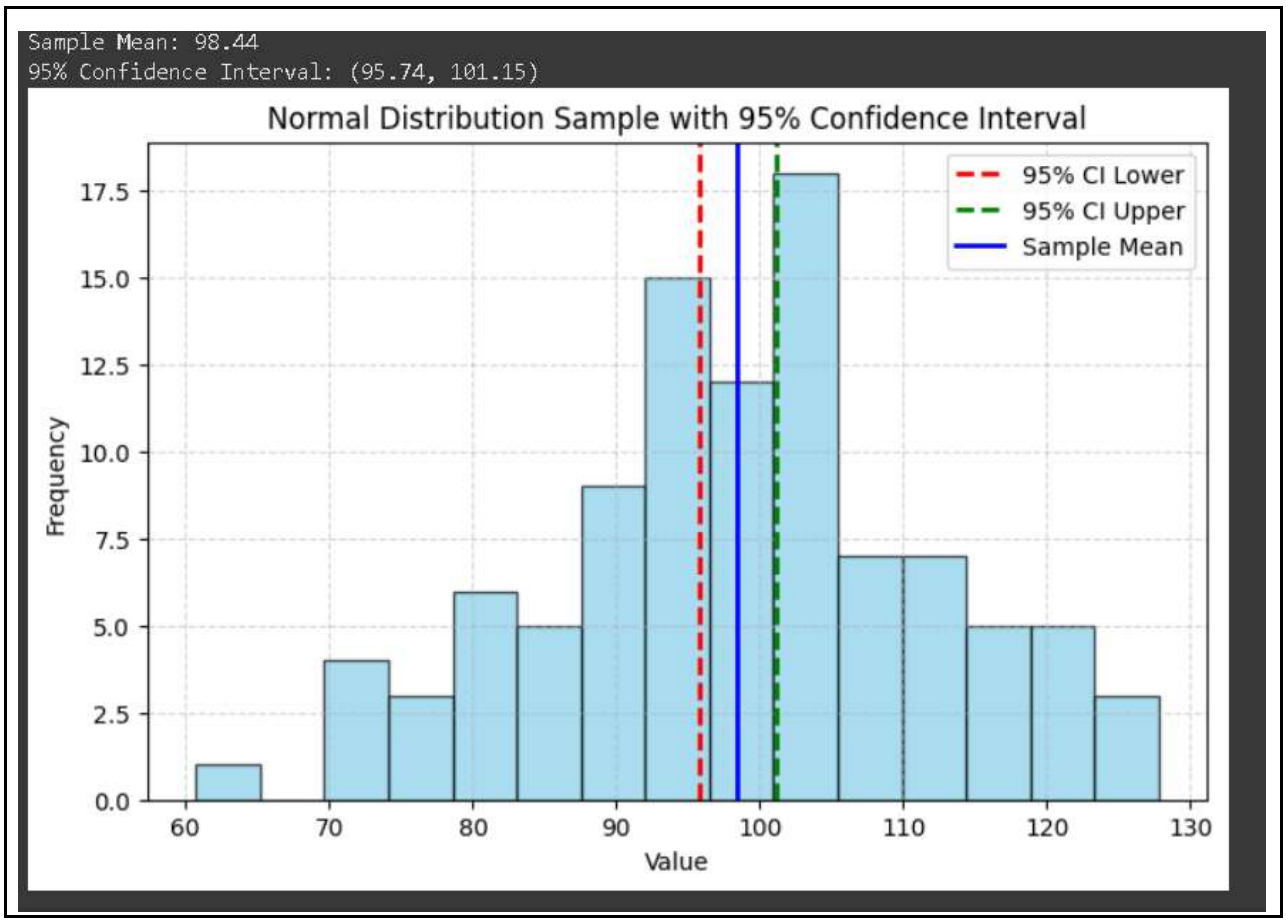


Question 9: Write a Python function to calculate the Z-scores from a dataset and visualize the standardized data using a histogram. Explain what the Z-scores represent in terms of standard deviations from the mean.

(*Include your Python code and output in the code box below.*) **Answer:**

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

#  Simulate data from a normal distribution
np.random.seed(42)  # for reproducibility
mean = 100
std_dev = 15
sample_size = 100

data = np.random.normal(loc=mean, scale=std_dev, size=sample_size)
```

```
#  Calculate sample mean and standard error
sample_mean = np.mean(data)
standard_error = stats.sem(data)  # SEM = std_dev / sqrt(n)

#  Calculate 95% confidence interval
confidence_level = 0.95
confidence_interval = stats.t.interval(confidence_level, df=len(data)-1,
                                       loc=sample_mean, scale=standard_error)

# : Print results
print(f"Sample Mean: {sample_mean:.2f}")
print(f"95% Confidence Interval: ({confidence_interval[0]:.2f}, {confidence_interval[1]:.2f})")

# Plot the histogram of the data
plt.figure(figsize=(8, 5))
plt.hist(data, bins=15, color='skyblue', edgecolor='black', alpha=0.7)
plt.axvline(confidence_interval[0], color='red', linestyle='dashed', linewidth=2, label='95% CI
Lower')
plt.axvline(confidence_interval[1], color='green', linestyle='dashed', linewidth=2, label='95% CI
Upper')
plt.axvline(sample_mean, color='blue', linestyle='solid', linewidth=2, label='Sample Mean')
plt.title('Normal Distribution Sample with 95% Confidence Interval')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.legend()
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```


Sample Mean: 98.44
95% Confidence Interval: (95.74, 101.15)
Normal Distribution Sample with 95% Confidence Interval
95% CI Lower
95% CI Upper
Sample Mean
Frequency
Value
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
60
70
80
90
100
110
120
130