statistics-advance-3

August 13, 2023

[ ]:
```
# Q1: What is Estimation Statistics? Explain point estimate and interval␣
↪estimate.
```

Estimation in statistics refers to the process of using sample data to make educated guesses about the population parameters.

Point Estimate: A point estimate is a single value that serves as the best guess for the population parameter. It is obtained by calculating a descriptive statistic from the sample data.

For example, if you want to estimate the average height of all students in a school, you could calculate the average height of a sample of students and use that as your point estimate for the population parameter.

Interval Estimate: An interval estimate, also known as a confidence interval, is a range of values within which the true population parameter is likely to fall. It provides a measure of uncertainty associated with the estimation process. Confidence intervals are typically expressed with a confidence level, which represents the probability that the interval contains the true population parameter.

Common confidence levels include 90%, 95%, and 99%.

For instance, if you calculate the average income of a sample of households and obtain a point estimate of $50,000 with a 95% confidence interval of $45,000 to $55,000, it means you are 95% confident that the true average income of all households falls within this range.

[4]:
```
# Q2. Write a Python function to estimate the population mean using a sample␣
↪mean and standard deviation


import math
import scipy.stats as stats

def estimate_population_mean(sample_mean, sample_stddev, sample_size):
    # Calculate the standard error of the mean (SEM)
    standard_error = sample_stddev / math.sqrt(sample_size)

    # Calculate the margin of error for a specific confidence level (e.g., 95%)
    confidence_level = 0.95
    z_score = abs(stats.norm.ppf((1 - confidence_level) / 2))  # For a␣
↪two-tailed test
    margin_of_error = z_score * standard_error
```

```
    # Calculate the lower and upper bounds of the confidence interval
    lower_bound = sample_mean - margin_of_error
    upper_bound = sample_mean + margin_of_error

    # Return the population mean estimate and the confidence interval
    return sample_mean, (lower_bound, upper_bound)

# Example usage
sample_mean = 150
sample_stddev = 20
sample_size = 50

population_mean_estimate, confidence_interval = estimate_population_mean(sample_mean, sample_stddev, sample_size)
print("Population Mean Estimate:", population_mean_estimate)
print("Confidence Interval:", confidence_interval)
```

```
Population Mean Estimate: 150
Confidence Interval: (144.4563847026013, 155.5436152973987)
```

[ ]:
```
# Q3: What is Hypothesis testing? Why is it used? State the importance of Hypothesis testing
```

Hypothesis testing is a fundamental concept in statistics used to make decisions and draw conclusions about a population based on a sample of data. It involves setting up two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (Ha), and then using sample data to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

Importance of Hypothesis Testing:

Informed Decision-Making: Hypothesis testing provides a structured framework for making decisions based on data. It helps us avoid making conclusions solely based on intuition or anecdotal evidence.

Scientific Inquiry: Hypothesis testing is crucial in scientific research as it allows researchers to test specific predictions and theories against empirical data. It helps validate or refine scientific theories.

Quality Control: In industries and manufacturing, hypothesis testing is used to ensure product quality and process stability. It helps identify deviations from desired standards.

[ ]:
```
# Q4. Create a hypothesis that states whether the average weight of male college students is greater than the average weight of female college students.
```

Null Hypothesis (H0): The average weight of male college students is equal to or less than the average weight of female college students.

Alternative Hypothesis (Ha): The average weight of male college students is greater than the

average weight of female college students.

In symbols:

H0: _male _female Ha: _male > _female

Null Hypothesis: The population mean weight of male college students is less than or equal to the population mean weight of female college students. Alternative Hypothesis: The population mean weight of male college students is greater than the population mean weight of female college students.

In hypothesis testing, you would collect data from a sample of male and female college students and use statistical methods to analyze whether there is enough evidence to support rejecting the null hypothesis in favor of the alternative hypothesis. This would involve calculating sample means, conducting a hypothesis test, and determining whether the observed difference in sample means is statistically significant.

[6]:
```
# Q5. Write a Python script to conduct a hypothesis test on the difference␣
 ↪between two population means, given a sample from each population.


import numpy as np
import scipy.stats as stats

# Sample data for two populations (replace with your actual data)
sample_male = np.array([70, 72, 75, 68, 73, 71, 74, 69, 70, 72])
sample_female = np.array([63, 65, 68, 61, 66, 64, 67, 62, 63, 65])

# Define significance level (alpha)
alpha = 0.05

# Calculate sample means and standard deviations
mean_male = np.mean(sample_male)
stddev_male = np.std(sample_male, ddof=1)
n_male = len(sample_male)

mean_female = np.mean(sample_female)
stddev_female = np.std(sample_female, ddof=1)
n_female = len(sample_female)

# Calculate pooled standard deviation (assuming equal variances)
pooled_stddev = np.sqrt(((n_male - 1) * stddev_male**2 + (n_female - 1) *␣
 ↪stddev_female**2) / (n_male + n_female - 2))

# Calculate t-statistic
t_statistic = (mean_male - mean_female) / (pooled_stddev * np.sqrt(1/n_male + 1/
 ↪n_female))

# Calculate degrees of freedom
```

```
degrees_of_freedom = n_male + n_female - 2

# Calculate critical t-value for a two-tailed test
critical_t_value = stats.t.ppf(1 - alpha/2, degrees_of_freedom)

# Calculate p-value
p_value = 2 * (1 - stats.t.cdf(abs(t_statistic), degrees_of_freedom))

# Conduct hypothesis test
if abs(t_statistic) > critical_t_value:
    conclusion = "Reject the null hypothesis"
else:
    conclusion = "Fail to reject the null hypothesis"

# Print results
print("Sample Mean (Male):", mean_male)
print("Sample Mean (Female):", mean_female)
print("t-statistic:", t_statistic)
print("Degrees of Freedom:", degrees_of_freedom)
print("Critical t-value:", critical_t_value)
print("p-value:", p_value)
print("Conclusion:", conclusion)
```

```
Sample Mean (Male): 71.4
Sample Mean (Female): 64.4
t-statistic: 7.047138579747256
Degrees of Freedom: 18
Critical t-value: 2.10092204024096
p-value: 1.4197141531280266e-06
Conclusion: Reject the null hypothesis
```

[ ]:
```
# Q6: What is a null and alternative hypothesis? Give some examples.
```

In hypothesis testing, the null hypothesis (H0) and the alternative hypothesis (Ha) are two competing statements that are used to make decisions about population parameters based on sample data.

Null Hypothesis (H0):

The null hypothesis is the default assumption or statement of no effect, no difference, or no change.

Alternative Hypothesis (Ha):

The alternative hypothesis is the statement that contradicts the null hypothesis.

Example: Coin Tossing

Null Hypothesis (H0): “The probability of getting heads in a fair coin toss is 0.5.” Alternative Hypothesis (Ha): “The probability of getting heads in a fair coin toss is not 0.5.”

[ ]:
```
# Q7: Write down the steps involved in hypothesis testing.
```

1. Formulate Hypotheses:

Null Hypothesis (H0): State the default assumption or statement of no effect, no difference, or no change. Alternative Hypothesis (Ha): State the contradictory statement that represents the effect, difference, or change you are trying to find evidence for.

2. Select Significance Level (Alpha):

Choose a significance level (alpha) that determines how strong the evidence must be to reject the null hypothesis. Common values for alpha include 0.05 (5%) or 0.01 (1%).

3. Collect and Prepare Data:

Collect data from a sample or samples that are relevant to the hypotheses being tested. Ensure the data meets the assumptions of the chosen statistical test.

4. Choose a Test Statistic:

Select an appropriate test statistic based on the nature of the data, the research question, and the hypotheses. Common test statistics include t-statistics, z-statistics, F-statistics, and chi-squared statistics.

5. Calculate the Test Statistic:

Compute the test statistic using the sample data and the chosen formula or calculation method.

6. Determine the Critical Region:

Determine the critical values or critical region(s) associated with the chosen significance level and test statistic. The critical region defines the range of values for which the null hypothesis will be rejected.

7. Calculate the P-Value:

Calculate the p-value, which is the probability of obtaining a test statistic as extreme as or more extreme than the observed value, assuming the null hypothesis is true. The p-value helps assess the strength of the evidence against the null hypothesis.

8. Make a Decision:

If the p-value is less than the chosen significance level (alpha), reject the null hypothesis in favor of the alternative hypothesis. If the p-value is greater than or equal to alpha, fail to reject the null hypothesis.

9. Draw a Conclusion:

Based on the decision made in the previous step, draw a conclusion about the hypotheses. Interpret the results in the context of the research question.

10. Communicate Results:

Present the findings, including the test statistic, p-value, conclusion, and any implications, in a clear and understandable manner. Discuss the practical significance of the results and their implications for the broader context.

[ ]:
```
# Q8. Define p-value and explain its significance in hypothesis testing
```

The p-value (short for “probability value”) is a crucial concept in hypothesis testing. It quantifies the strength of the evidence against the null hypothesis.

The smaller the p-value, the stronger the evidence against the null hypothesis.

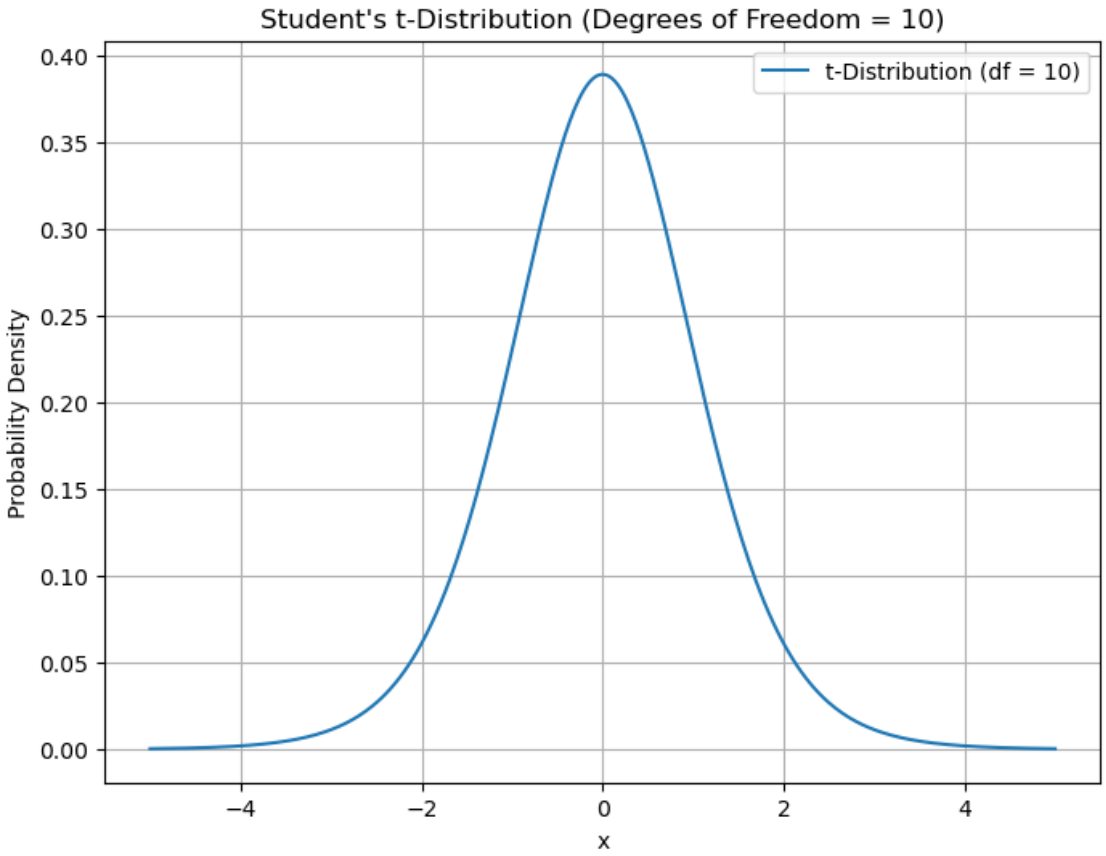
Significance Of P Value :-

Quantifies Evidence: The p-value quantifies the strength of evidence against the null hypothesis. It provides a numerical measure of how likely it is to observe the data or more extreme results under the assumption that the null hypothesis is true. A low p-value indicates stronger evidence against the null hypothesis.

Basis for Decision-Making: The p-value serves as a critical factor in making decisions about hypotheses. It provides an objective criterion for deciding whether to reject the null hypothesis in favor of the alternative hypothesis, based on a predetermined significance level (alpha).

[7]:
```
# Q9. Generate a Student's t-distribution plot using Python's matplotlib␣
 ↪library, with the degrees of freedom parameter set to 10.


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

# Degrees of freedom
degrees_of_freedom = 10

# Generate a range of x values
x = np.linspace(-5, 5, 400)

# Calculate the probability density function (PDF) values for the t-distribution
pdf_values = stats.t.pdf(x, df=degrees_of_freedom)

# Create the plot
plt.figure(figsize=(8, 6))
plt.plot(x, pdf_values, label=f't-Distribution (df = {degrees_of_freedom})')

# Add labels and title
plt.xlabel('x')
plt.ylabel('Probability Density')
plt.title(f'Student\'s t-Distribution (Degrees of Freedom =␣
 ↪{degrees_of_freedom})')

# Add a legend
plt.legend()

# Show the plot
```

```
plt.grid()
plt.show()
```



[8]:
```
# Q10. Write a Python program to calculate the two-sample t-test for␣
↪independent samples, given two random samples of equal size and a null␣
↪hypothesis that the population means are equal.


import numpy as np
import scipy.stats as stats

# Generate two random samples of equal size
np.random.seed(42)
sample1 = np.random.normal(50, 10, 30)
sample2 = np.random.normal(55, 10, 30)

# Define significance level (alpha)
alpha = 0.05

# Calculate sample statistics
```

```
mean1 = np.mean(sample1)
stddev1 = np.std(sample1, ddof=1)
n1 = len(sample1)

mean2 = np.mean(sample2)
stddev2 = np.std(sample2, ddof=1)
n2 = len(sample2)

# Calculate pooled standard deviation
pooled_stddev = np.sqrt(((n1 - 1) * stddev1**2 + (n2 - 1) * stddev2**2) / (n1 +␣
 ↪n2 - 2))

# Calculate the t-statistic
t_statistic = (mean1 - mean2) / (pooled_stddev * np.sqrt(1/n1 + 1/n2))

# Calculate degrees of freedom
degrees_of_freedom = n1 + n2 - 2

# Calculate critical t-value for a two-tailed test
critical_t_value = stats.t.ppf(1 - alpha/2, degrees_of_freedom)

# Calculate p-value
p_value = 2 * (1 - stats.t.cdf(abs(t_statistic), degrees_of_freedom))

# Conduct hypothesis test
if abs(t_statistic) > critical_t_value:
    conclusion = "Reject the null hypothesis"
else:
    conclusion = "Fail to reject the null hypothesis"

# Print results
print("Sample 1 Mean:", mean1)
print("Sample 2 Mean:", mean2)
print("t-statistic:", t_statistic)
print("Degrees of Freedom:", degrees_of_freedom)
print("Critical t-value:", critical_t_value)
print("p-value:", p_value)
print("Conclusion:", conclusion)
```

```
Sample 1 Mean: 48.118531041489625
Sample 2 Mean: 53.788375297100565
t-statistic: -2.398115152010242
Degrees of Freedom: 58
Critical t-value: 2.0017174830120923
p-value: 0.01971794186575826
Conclusion: Reject the null hypothesis
```

[ ]:
```
# Q11: What is Student's t distribution? When to use the t-Distribution.
```

The Student's t-distribution, often referred to as the t-distribution, is a probability distribution that is used in statistical hypothesis testing when the sample size is small and the population standard deviation is unknown.

The t-distribution is characterized by its degrees of freedom (df), which determine the shape of the distribution. As the degrees of freedom increase, the t-distribution approaches the standard normal distribution. The t-distribution is symmetrical and bell-shaped, like the normal distribution, but it has wider tails.

When to Use the t-Distribution:

The t-distribution is used in scenarios where certain assumptions are met and the sample size is relatively small:

Small Sample Size: The t-distribution is particularly useful when the sample size is small (typically less than 30) and doesn't meet the requirement for a large sample size assumed by the normal distribution.

Population Standard Deviation Unknown: When the population standard deviation is unknown, the t-distribution is used to estimate the standard error of the sample mean. This occurs frequently in practice since population standard deviations are often unavailable.

Hypothesis Testing: The t-distribution is commonly used in hypothesis testing when dealing with small sample sizes. It helps calculate t-values for various statistical tests, such as one-sample t-test, two-sample t-test, and paired t-test.

Confidence Intervals: When constructing confidence intervals for population parameters (e.g., population mean), the t-distribution is used when the population standard deviation is unknown and the sample size is small.

Comparing Means: In cases where you want to compare the means of two groups (independent samples) or the means of two related variables (paired samples), and the sample sizes are small, the t-distribution is appropriate.

Errors in Measurement: When dealing with situations where there might be errors or variability in measurement, the t-distribution provides a more accurate representation of uncertainty compared to the normal distribution.

Statistical Estimation: When estimating population parameters from small samples, the t-distribution provides a better approximation of the sampling distribution of the statistic.

[ ]:
```
# Q12: What is t-statistic? State the formula for t-statistic.
```

The t-statistic is a numerical value that measures the difference between the sample mean and a hypothesized population mean in terms of standard error. It quantifies how many standard errors the sample mean is away from the hypothesized population mean.

One-Sample t-Test: Used to test whether the mean of a single sample is significantly different from a hypothesized population mean. Formula:

t = Sample Mean − Hypothesized Mean / Standard Error of the Sample Mean

Two-Sample t-Test for Independent Samples: Used to compare the means of two independent samples to determine if there is a statistically significant difference between them. Formula: t = Difference in Sample Means / Pooled Standard Error of the Difference

Paired t-Test:

Used to compare the means of paired samples (e.g., before and after measurements) to determine if there is a significant difference.

Formula: t= Standard Error of the Differences / Sample Mean of Differences

[ ]:
```
# Q13. A coffee shop owner wants to estimate the average daily revenue for␣
 ↪their shop.
# They take a random sample of 50 days and find the sample mean revenue to be␣
 ↪$500 with a standard deviation of $50.
# Estimate the population mean revenue with a 95% confidence interval
```

To estimate the population mean revenue with a 95% confidence interval, we can use the formula for the confidence interval for a population mean based on a normal distribution. The formula is:

Confidence Interval = Sample Mean ± Margin of Error

Given the information: Sample Mean (x) = $500 Standard Deviation ( ) = $50 Sample Size (n) = 50 Confidence Level = 95%

[9]:
```
import math

# Given values
sample_mean = 500
standard_deviation = 50
sample_size = 50
confidence_level = 0.95
critical_value = 1.96  # For a 95% confidence level

# Calculate standard error of the mean
standard_error = standard_deviation / math.sqrt(sample_size)

# Calculate margin of error
margin_of_error = critical_value * standard_error

# Calculate confidence interval
lower_bound = sample_mean - margin_of_error
upper_bound = sample_mean + margin_of_error

# Print the confidence interval
print("95% Confidence Interval:")
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)
```

```
95% Confidence Interval:
Lower Bound: 486.1407070887437
```

```
Upper Bound: 513.8592929112564
```

[10]:
```python
# Q14. A researcher hypothesizes that a new drug will decrease blood pressure␣
 ↪by 10 mmHg.
# They conduct a clinical trial with 100 patients and find that the sample mean␣
 ↪decrease in blood pressure is 8 mmHg with a standard deviation of 3 mmHg.
# Test the hypothesis with a significance level of 0.05.



import math
import scipy.stats as stats

# Given values
sample_mean = 8
standard_deviation = 3
sample_size = 100
hypothesized_mean = 10
significance_level = 0.05

# Calculate the t-statistic
t_statistic = (sample_mean - hypothesized_mean) / (standard_deviation / math.
 ↪sqrt(sample_size))

# Calculate degrees of freedom
degrees_of_freedom = sample_size - 1

# Calculate critical t-value for a two-tailed test
critical_t_value = stats.t.ppf(1 - significance_level/2, degrees_of_freedom)

# Perform the hypothesis test
if abs(t_statistic) > critical_t_value:
    conclusion = "Reject the null hypothesis"
else:
    conclusion = "Fail to reject the null hypothesis"

# Print results
print("Sample Mean:", sample_mean)
print("t-statistic:", t_statistic)
print("Degrees of Freedom:", degrees_of_freedom)
print("Critical t-value:", critical_t_value)
print("Conclusion:", conclusion)
```

```
Sample Mean: 8
t-statistic: -6.666666666666667
Degrees of Freedom: 99
Critical t-value: 1.9842169515086827
```

Conclusion: Reject the null hypothesis

[11]:
```
# Q15. An electronics company produces a certain type of product with a mean weight of 5 pounds and a standard deviation of 0.5 pounds.
# A random sample of 25 products is taken, and the sample mean weight is found to be 4.8 pounds.
# Test the hypothesis that the true mean weight of the products is less than 5 pounds with a significance level of 0.01.


import math
import scipy.stats as stats

# Given values
sample_mean = 4.8
standard_deviation = 0.5
sample_size = 25
hypothesized_mean = 5
significance_level = 0.01

# Calculate the t-statistic
t_statistic = (sample_mean - hypothesized_mean) / (standard_deviation / math.sqrt(sample_size))

# Calculate degrees of freedom
degrees_of_freedom = sample_size - 1

# Calculate critical t-value for a lower-tailed test
critical_t_value = stats.t.ppf(significance_level, degrees_of_freedom)

# Perform the hypothesis test
if t_statistic < critical_t_value:
    conclusion = "Reject the null hypothesis"
else:
    conclusion = "Fail to reject the null hypothesis"

# Print results
print("Sample Mean:", sample_mean)
print("t-statistic:", t_statistic)
print("Degrees of Freedom:", degrees_of_freedom)
print("Critical t-value:", critical_t_value)
print("Conclusion:", conclusion)
```

```
Sample Mean: 4.8
t-statistic: -2.0000000000000018
Degrees of Freedom: 24
Critical t-value: -2.4921594731575762
```

Conclusion: Fail to reject the null hypothesis

[12]:
```
# Q16. Two groups of students are given different study materials to prepare␣
 ↪for a test.
# The first group (n1 = 30) has a mean score of 80 with a standard deviation of␣
 ↪10, and the second group (n2 = 40) has a mean score of 75 with a standard␣
 ↪deviation of 8.
# Test the hypothesis that the population means for the two groups are equal␣
 ↪with a significance level of 0.01.



import math
import scipy.stats as stats

# Given values for Group 1
sample_mean_1 = 80
sample_stddev_1 = 10
sample_size_1 = 30

# Given values for Group 2
sample_mean_2 = 75
sample_stddev_2 = 8
sample_size_2 = 40

# Significance level
significance_level = 0.01

# Calculate pooled standard error of the difference in means
pooled_std_error = math.sqrt(
    (sample_stddev_1 ** 2 / sample_size_1) + (sample_stddev_2 ** 2 /␣
 ↪sample_size_2)
)

# Calculate t-statistic
t_statistic = (sample_mean_1 - sample_mean_2) / pooled_std_error

# Calculate degrees of freedom
degrees_of_freedom = sample_size_1 + sample_size_2 - 2

# Calculate critical t-value for a two-tailed test
critical_t_value = stats.t.ppf(1 - significance_level / 2, degrees_of_freedom)

# Perform the hypothesis test
if abs(t_statistic) > critical_t_value:
    conclusion = "Reject the null hypothesis"
else:
```

```
    conclusion = "Fail to reject the null hypothesis"

# Print results
print("t-statistic:", t_statistic)
print("Degrees of Freedom:", degrees_of_freedom)
print("Critical t-value:", critical_t_value)
print("Conclusion:", conclusion)
```

```
t-statistic: 2.2511258444537408
Degrees of Freedom: 68
Critical t-value: 2.6500812928169553
Conclusion: Fail to reject the null hypothesis
```

[13]:
```
# Q17. A marketing company wants to estimate the average number of ads watched␣
 ↪by viewers during a TV program.
# They take a random sample of 50 viewers and find that the sample mean is 4␣
 ↪with a standard deviation of 1.5.
# Estimate the population mean with a 99% confidence interval.


import math

# Given values
sample_mean = 4
standard_deviation = 1.5
sample_size = 50
confidence_level = 0.99
critical_value = 2.576  # For a 99% confidence level

# Calculate standard error of the mean
standard_error = standard_deviation / math.sqrt(sample_size)

# Calculate margin of error
margin_of_error = critical_value * standard_error

# Calculate confidence interval
lower_bound = sample_mean - margin_of_error
upper_bound = sample_mean + margin_of_error

# Print the confidence interval
print("99% Confidence Interval:")
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)
```

```
99% Confidence Interval:
Lower Bound: 3.453547879499036
Upper Bound: 4.546452120500964
```