INFERENTIAL STATISTICS

```
#Inferential Statistics
```

INTRODUCTION

Inferential statistics is a branch of statistics that allows us to make conclusions or inferences about a population based on data from a sample of that population. It involves the use of various techniques and methods to analyze sample data, make predictions, and test hypotheses. Here are the key components and concepts of inferential statistics:

1. Population vs. Sample Population: The entire group of individuals or instances about whom we hope to learn. Sample: A subset of the population, selected for measurement and analysis.
2. Parameters and Statistics Parameter: A numerical summary of a population (e.g., population mean μ μ, population proportion P P). Statistic: A numerical summary of a sample (e.g., sample mean x ¯ x ¯ , sample proportion P ^ P ^).
3. Sampling Methods Random Sampling: Each member of the population has an equal chance of being selected. Stratified Sampling: The population is divided into strata, and random samples are taken from each stratum. Cluster Sampling: The population is divided into clusters, and entire clusters are randomly selected. Systematic Sampling: Every k t h k th member of the population is selected.
4. Hypothesis Testing Null Hypothesis (H 0 H 0): A statement of no effect or no difference, which we test against. Alternative Hypothesis (H 1 H 1): A statement that indicates the presence of an effect or a difference. P-value: The probability of observing the sample data, or something more extreme, assuming the null hypothesis is true. Significance Level (α α): The threshold at which we reject the null hypothesis (commonly 0.05).
5. Confidence Intervals A range of values, derived from the sample statistics, that is likely to contain the population parameter. Formula: x ¯ ± z s n x ¯ ±z n

s, where x ¯ x ¯ is the sample mean, z z is the z-score corresponding to the desired confidence level, s s is the sample standard deviation, and n n is the sample size.

6. Common Tests and Techniques Z-Test: Used for hypothesis testing when the population variance is known and the sample size is large. T-Test: Used when the population variance is unknown and the sample size is small. One-sample T-test: Tests whether the sample mean is different from a known value. Two-sample T-test: Tests whether the means of two independent samples are different. Paired T-test: Tests whether the means of two related samples are different. ANOVA (Analysis of Variance): Used to compare the means of three or more samples. Chi-Square Test: Used for categorical data to assess how likely it is that an observed distribution is due to chance. Regression Analysis: Used to examine the relationship between variables.

7. Assumptions in Inferential Statistics Random Sampling: The sample should be randomly selected from the population. Independence: Observations must be independent of each other. Normality: Data should be approximately normally distributed (especially for small sample sizes). Homogeneity of Variance: Variances within different groups should be roughly equal.
8. Errors in Hypothesis Testing Type I Error (α α): Rejecting the null hypothesis when it is true (false positive). Type II Error (β β): Failing to reject the null hypothesis when it is false (false negative). Applications of Inferential Statistics Inferential statistics is widely used in various fields such as:

Medicine: For clinical trials and health studies to determine the effectiveness of treatments.

Economics:

To forecast economic trends and make policy decisions. Psychology:

To understand behavior patterns and psychological phenomena. Marketing: To analyze consumer behavior and improve marketing strategies.

Conclusion

Inferential statistics provides the tools to make informed decisions based on sample data. By understanding the principles and methods of inferential statistics, researchers can draw meaningful conclusions about a population, despite only having a subset of data.

Inferential Statistics with Python

1. Introduction to Parameters and Statistics Parameter: A measure (mean, median, variance, etc.) for population data. Statistic: A measure (mean, median, variance, etc.) for sample data.

In practice, calculating parameters for entire populations is often infeasible due to size and cost constraints, so we use sample statistics to estimate population parameters.

Example: Estimating the average life expectancy in India using a sample.

2. Sampling Techniques a) Simple Random Sampling

In simple random sampling, every member of the population has an equal chance of being selected.

Example: Selecting 12 tomato plants from a farm of 98 for a growth study.

```
import numpy as np

# Total population
population = np.arange(1, 99)
# Sample size
sample_size = 12

# Simple random sampling without replacement
sample_without_replacement = np.random.choice(population, sample_size, replace=False)
print("Sample without replacement:", sample_without_replacement)

# Simple random sampling with replacement
sample_with_replacement = np.random.choice(population, sample_size, replace=True)
print("Sample with replacement:", sample_with_replacement)
```

```
Sample without replacement: [ 5 84 93 89 79  3 11 87 45 40 17 22]
Sample with replacement: [46 19 82 41 10  1 89  2 83 93  4 61]
```

b) Stratified Sampling

In stratified sampling, the population is divided into strata, and random samples are taken from each stratum.

Example: Selecting 2 roses of each color from a nursery.

```
import pandas as pd

# Example data
roses = ['White', 'Pink', 'White', 'Red', 'Yellow', 'Orange', 'Orange', 'Red', 'Yellow', 'White', 'Pink', 'White', 'Red', 'Orange']
roses_df = pd.DataFrame({'Color': roses})

# Grouping by color
grouped = roses_df.groupby('Color').apply(lambda x: x.sample(2)).reset_index(drop=True)
print(grouped)
```

```
    Color
0  Orange
1  Orange
2    Pink
3    Pink
4     Red
5     Red
6   White
7   White
8  Yellow
9  Yellow
```

c) Systematic Sampling

In systematic sampling, the first data point is selected randomly, and subsequent points are selected at regular intervals.

Example: Ann selecting 4 marbles starting from the 2nd.

```
# Total population
marbles = np.arange(1, 21)
# Start point
start_point = 2
# Interval
interval = 5
# Sample size
sample_size = 4

# Systematic sampling
sample_systematic = marbles[start_point-1:start_point-1+sample_size*interval:interval]
print("Systematic sample:", sample_systematic)
```

```
Systematic sample: [ 2  7 12 17]
```

3. Central Limit Theorem (CLT)

For a sufficiently large sample size (typically n > 30), the distribution of sample means approximates a normal distribution, regardless of the population's distribution.

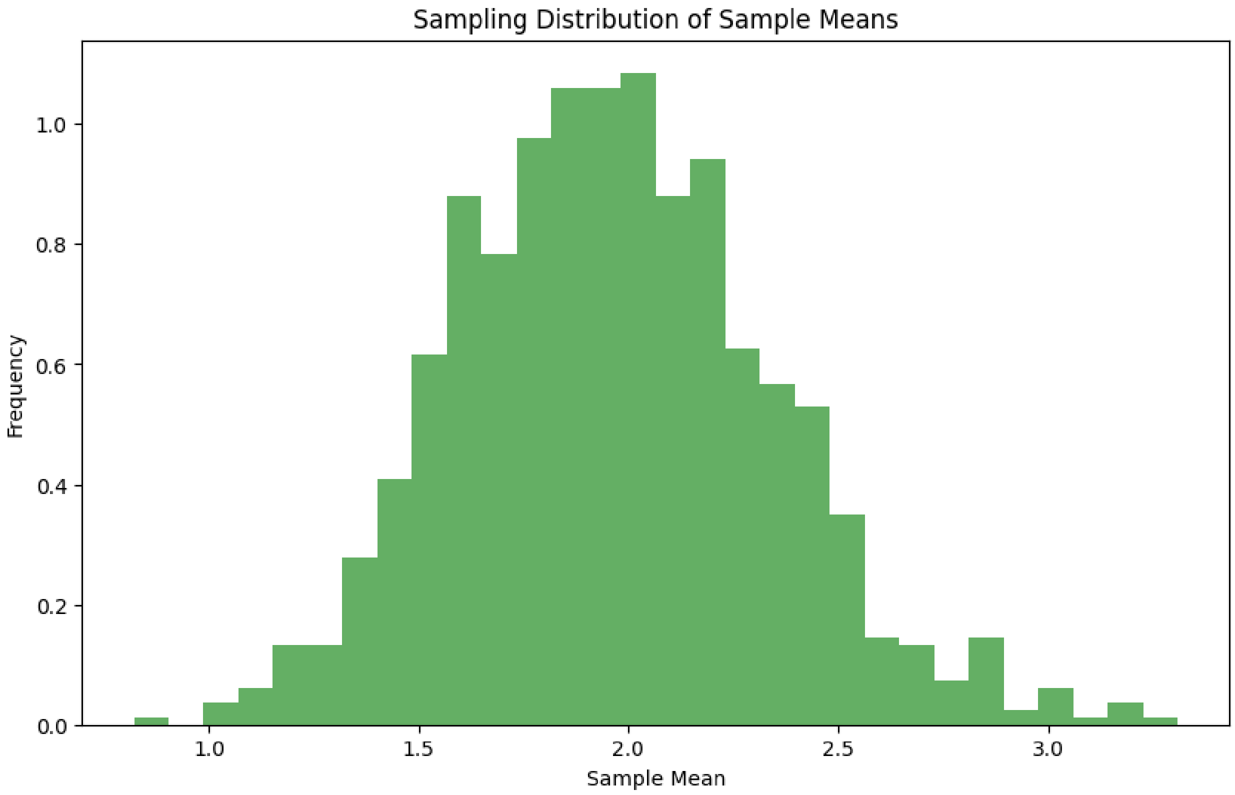
Example: Demonstrating CLT with Python.

```
import matplotlib.pyplot as plt

# Simulate population data
population = np.random.exponential(scale=2, size=1000)

# Draw samples and compute means
sample_means = [np.mean(np.random.choice(population, 30)) for _ in range(1000)]

# Plotting
plt.figure(figsize=(10, 6))
plt.hist(sample_means, bins=30, density=True, alpha=0.6, color='g')
plt.title('Sampling Distribution of Sample Means')
plt.xlabel('Sample Mean')
plt.ylabel('Frequency')
plt.show()
```



4. Estimation

a) Point Estimation

Using sample statistics to estimate population parameters.

Example: Estimating the mean length of steel rods.

```
# Sample data
length_rod = [25.2, 26.3, 28, 21.9, 23.4, 24, 27.2, 23, 29.2, 28.7, 23.1, 23.5, 26.4, 22.8, 24.7]
sample_mean = np.mean(length_rod)
print("Sample mean:", sample_mean)
```

```
Sample mean: 25.16
```

b) Interval Estimation (Confidence Intervals)

Estimating a range within which the population parameter likely falls.

Example: Calculating a 90% confidence interval for average sugar content in baby food.

```
import scipy.stats as stats

# Given data
sample_mean = 24
std_dev = 8
sample_size = 37
confidence_level = 0.90

# Z-score for 90% confidence level
z_score = stats.norm.isf((1 - confidence_level) / 2)
margin_of_error = z_score * (std_dev / np.sqrt(sample_size))

# Confidence interval
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)
print("90% Confidence Interval:", confidence_interval)
```

```
90% Confidence Interval: (21.83670183570907, 26.16329816429093)
```

5. Hypothesis **Testing**

a) Null Hypothesis (H0): There is no effect or difference. b) Alternative Hypothesis (Ha): There is an effect or difference.

Example: Testing the weight of green tea bags.

```
# Given data
sample_mean = 3.28
population_mean = 3.5
std_dev = 0.6
sample_size = 50
alpha = 0.10

# Test statistic (Z-score)
z_score = (sample_mean - population_mean) / (std_dev / np.sqrt(sample_size))

# P-value
p_value = stats.norm.sf(z_score)
print("P-value:", p_value)

# Conclusion
if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

```
P-value: 0.9952390544079506
Fail to reject the null hypothesis
```

Summary

Sampling Techniques: Simple random, stratified, and systematic sampling methods help in selecting representative samples. Central Limit Theorem: Demonstrates that the distribution of sample means is approximately normal for large samples. Estimation: Point estimation and confidence intervals are used to infer population parameters. Hypothesis Testing: Involves testing claims about population parameters using sample data. These concepts are fundamental to inferential statistics and form the basis for more advanced statistical techniques and data analysis in various fields.

Graphs and Visualization

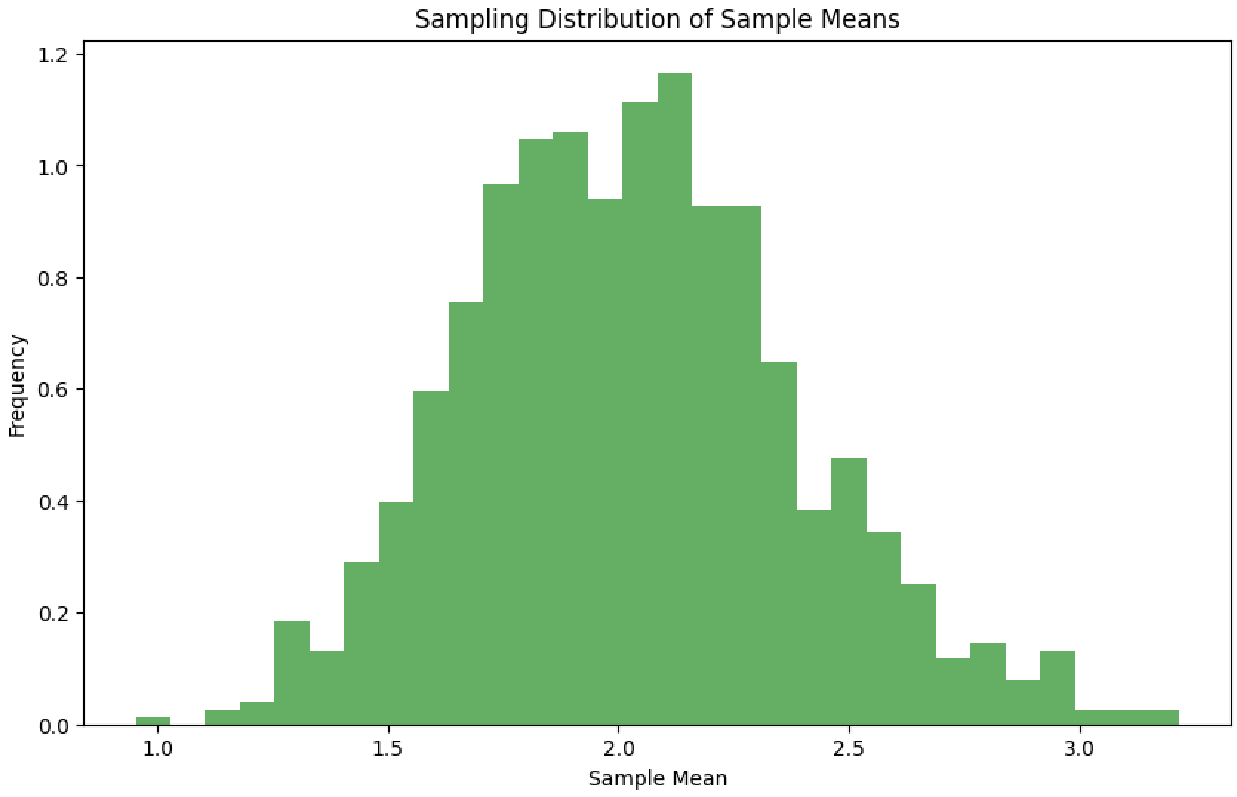
Graphs and visualizations play a crucial role in understanding and presenting data. Here are some example visualizations for the concepts discussed:

1. Sampling Distribution (Central Limit Theorem)

```
# Simulate population data
population = np.random.exponential(scale=2, size=1000)

# Draw samples and compute means
sample_means = [np.mean(np.random.choice(population, 30)) for _ in range(1000)]

# Plotting
plt.figure(figsize=(10, 6))
plt.hist(sample_means, bins=30, density=True, alpha=0.6, color='g')
plt.title('Sampling Distribution of Sample Means')
plt.xlabel('Sample Mean')
plt.ylabel('Frequency')
plt.show()
```

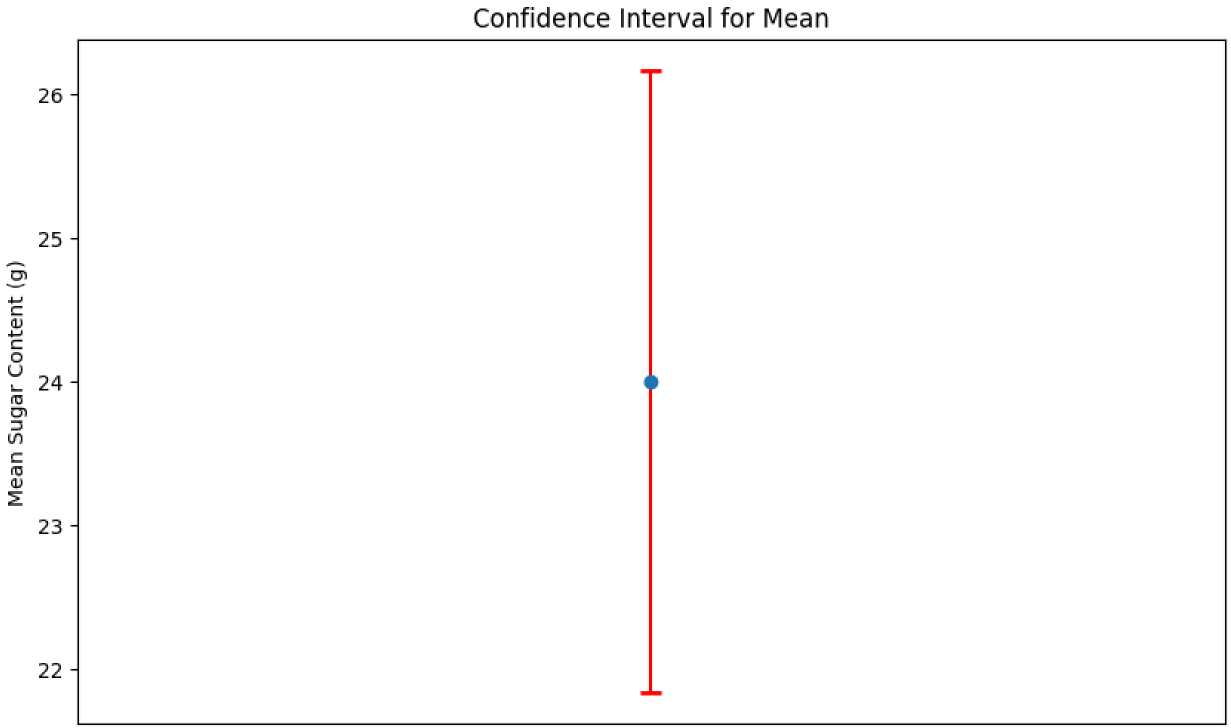


2. Confidence Interval Visualization

```
# Data for confidence interval
sample_mean = 24
std_dev = 8
sample_size = 37
confidence_level = 0.90

# Calculate margin of error and confidence interval
z_score = stats.norm.isf((1 - confidence_level) / 2)
margin_of_error = z_score * (std_dev / np.sqrt(sample_size))
confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)

# Plotting
plt.figure(figsize=(10, 6))
plt.errorbar(x=0, y=sample_mean, yerr=margin_of_error, fmt='o', capsize=5, capthick=2, ecolor='red')
plt.title('Confidence Interval for Mean')
plt.ylabel('Mean Sugar Content (g)')
plt.xticks([])
plt.show()
```



Below are examples of bell curve graphs (normal distribution curves) illustrating the rejection regions for the null hypothesis H 0 H 0in hypothesis testing. The rejection regions are based on the significance level α α.

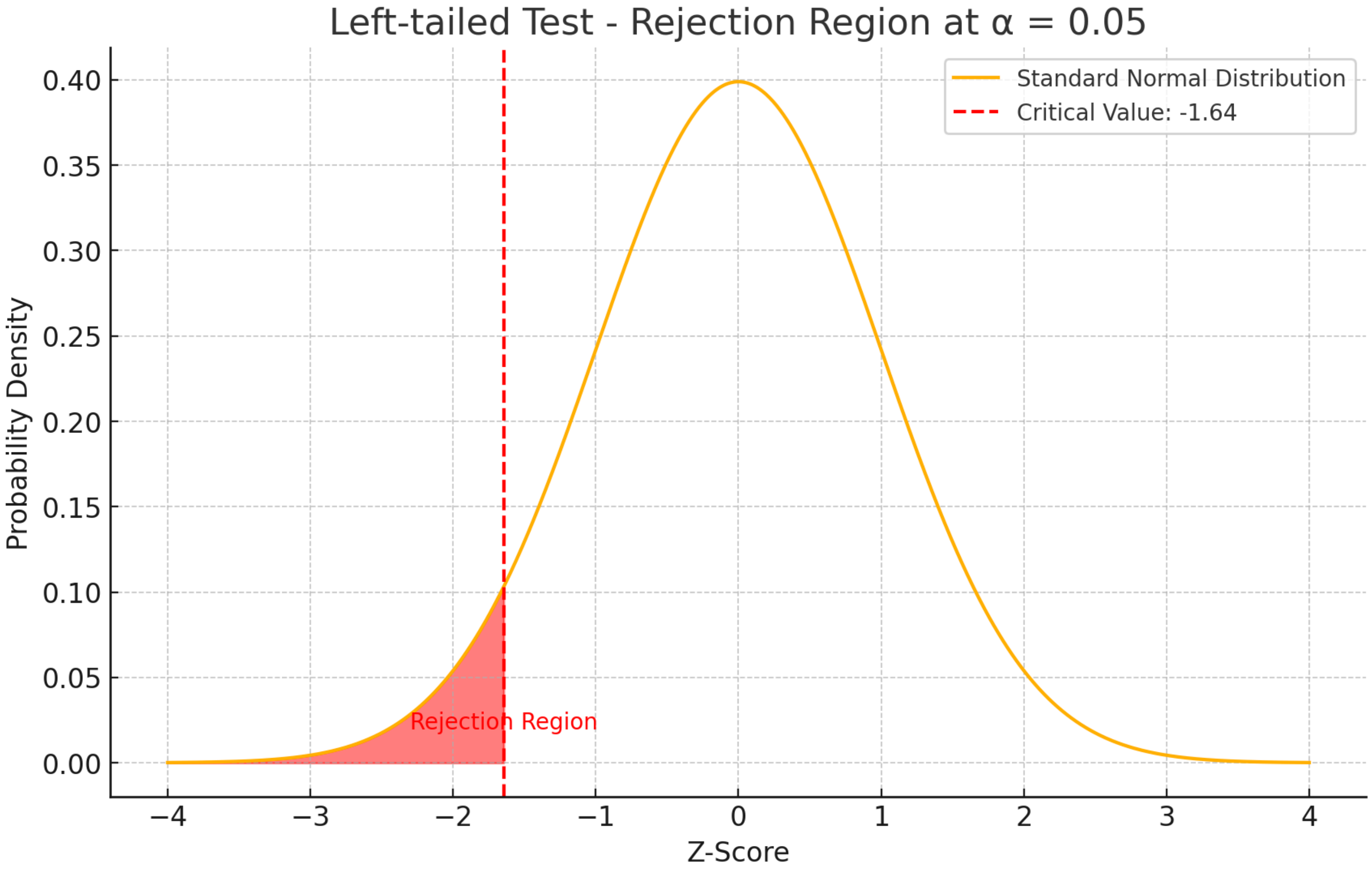
I'll provide visual representations for:

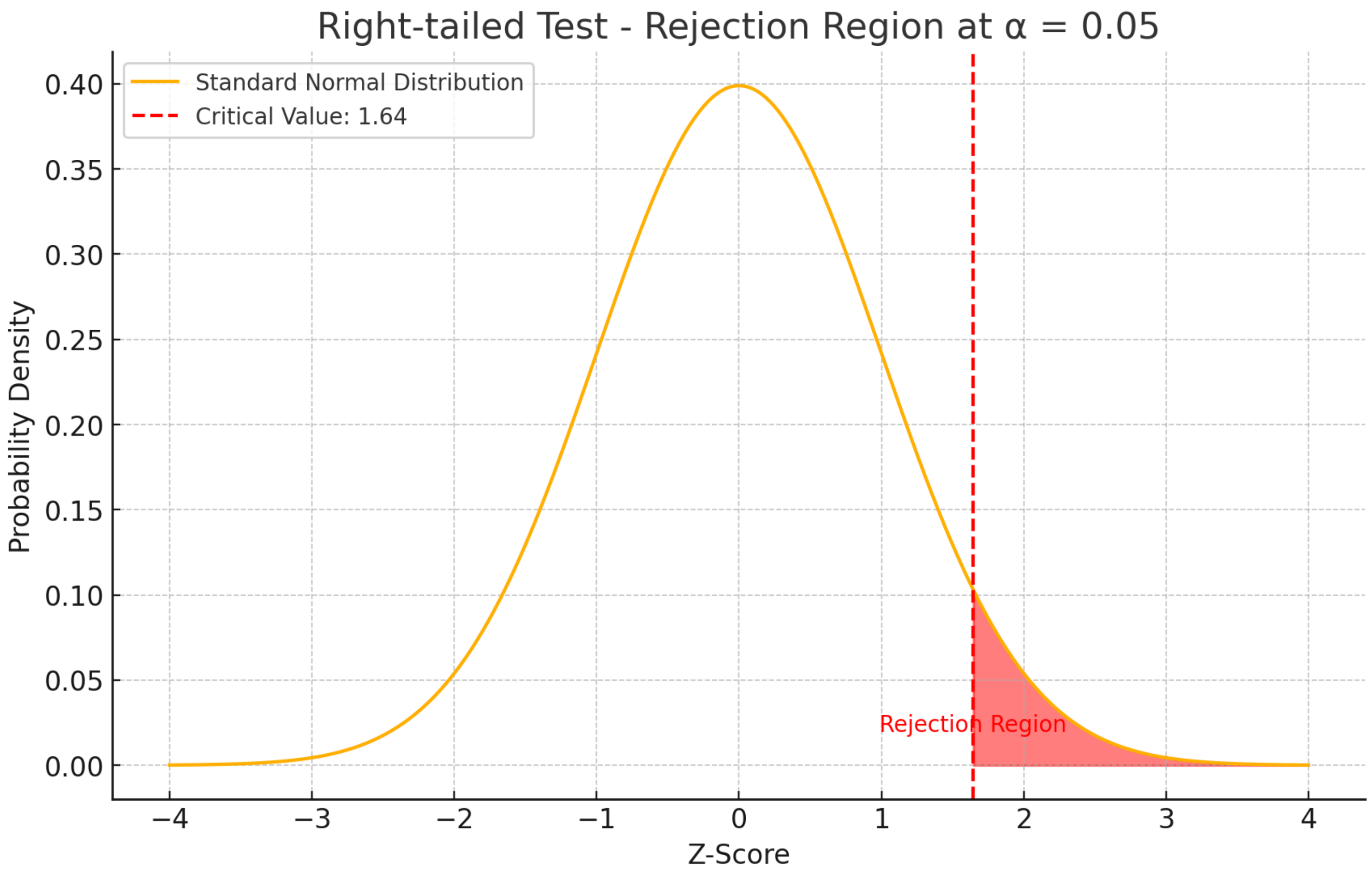
Left-Tailed Test Right-Tailed Test Two-Tailed Test Let's generate these graphs using Python.

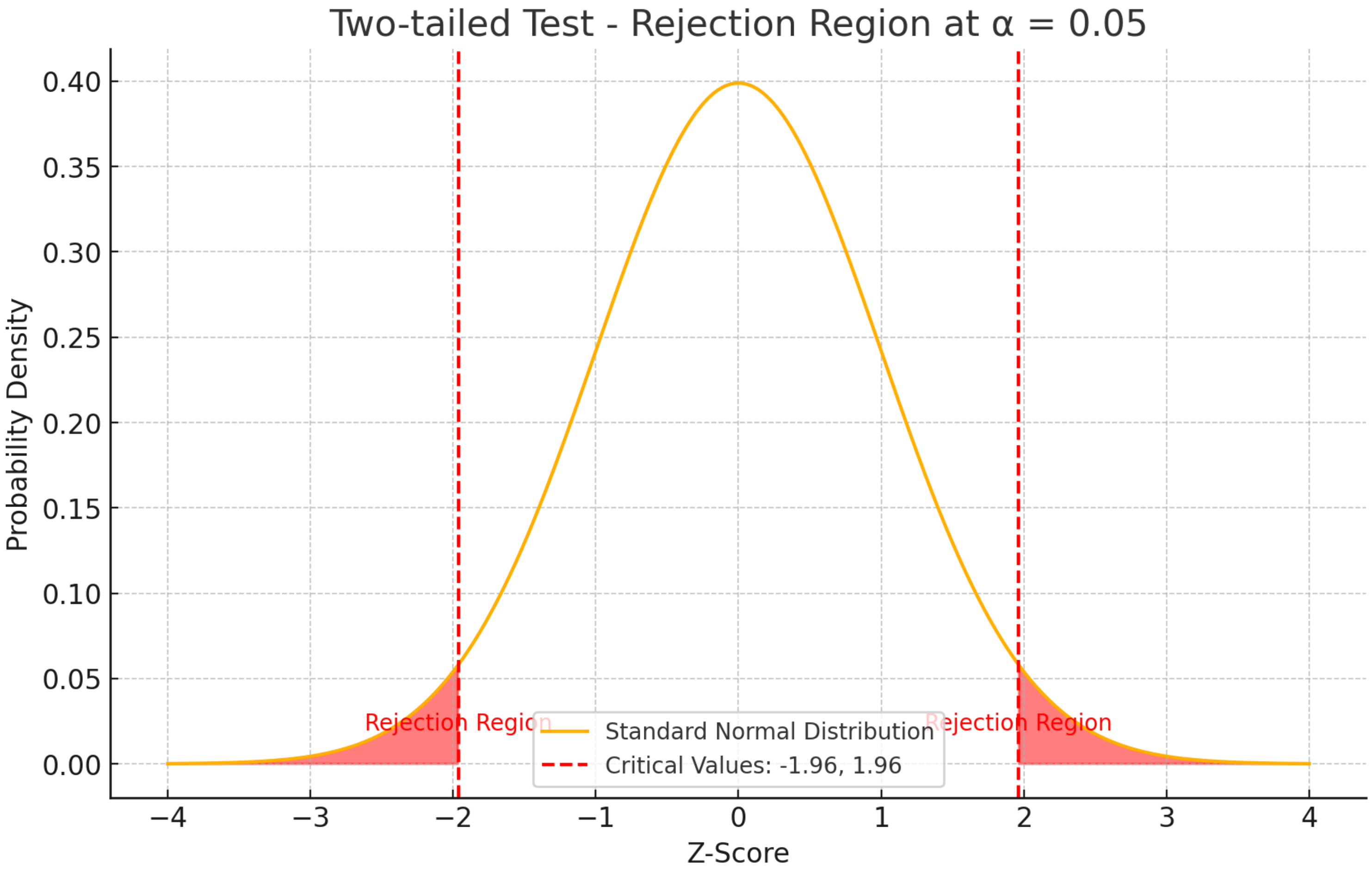
Left-Tailed Test In a left-tailed test, the rejection region is on the left side of the mean. We reject H 0 H 0if the test statistic falls in this region.

Right-Tailed Test In a right-tailed test, the rejection region is on the right side of the mean. We reject H 0 H 0if the test statistic falls in this region.

Two-Tailed Test In a two-tailed test, there are two rejection regions: one on the left side and one on the right side of the mean. We reject H 0 H 0 if the test statistic falls in either of these regions.


Left-tailed Test - Rejection Region at α = 0.05
Standard Normal Distribution
Critical Value: -1.64
Rejection Region
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
Z-Score


Right-tailed Test - Rejection Region at α = 0.05
Standard Normal Distribution
Critical Value: 1.64
Rejection Region
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Z-Score
−4
−3
−2
−1
0
1
2
3
4



Here are the bell curve graphs illustrating the rejection regions for hypothesis testing:

Left-Tailed Test:

The rejection region is on the left side of the mean. If the test statistic falls in this region, we reject the null hypothesis H 0 H 0. Right-Tailed Test:

The rejection region is on the right side of the mean. If the test statistic falls in this region, we reject the null hypothesis H 0 H 0. Two-Tailed Test:

There are two rejection regions, one on the left and one on the right side of the mean. If the test statistic falls in either of these regions, we reject the null hypothesis H 0 H 0. These graphs visually demonstrate how the significance level α α determines the critical values and rejection regions for different types of hypothesis tests.

Summary of Inferential Statistics in Python

Inferential statistics involves making predictions or inferences about a population based on a sample of data drawn from that population. This involves several key concepts and methods:

Parameters and Statistics:

Parameter: A measure for the entire population (e.g., population mean μ μ). Statistic: A measure from a sample of the population (e.g., sample mean $\bar{x}$ x¯). Sampling Techniques: