SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE CODE: DJS22ITL5013 **DATE: 11-09-2024**

COURSE NAME: Statistical Analysis Lab **CLASS: T.Y. BTech**

NAME: Anish Sharma **SAP: 60003220045**

EXPERIMENT NO.05

CO 2: Perform Test of Hypothesis for independence and appropriateness of distribution using various statistical

techniques.

AIM / OBJECTIVE: To implement single population tests for mean, proportion and variance

This experiment aims to implement two-sample tests for mean, proportion, and variance. The two-sample tests are used to make inferences about the differences between two population parameters based on two independent samples. The three types of tests covered in this experiment are:

1. **Two-Sample Mean Test**: This test is used to determine if there is a significant difference between the means of two independent populations.
2. **Two-Sample Proportion Test**: This test is used to determine if there is a significant difference between the proportions of two independent populations.
3. **Two-Sample Variance Test**: This test is used to determine if there is a significant difference between the variances of two independent populations.

Key differences between the single population tests and the two-sample tests:

1. **Single Population Tests**: These tests make inferences about a single population parameter based on a sample from that population.

2. **Two-Sample Tests**: These tests make inferences about the differences between two population parameters based on two independent samples.

For two-sample tests, the hypotheses and test statistics are formulated to compare differences between two sample statistics, whereas single population tests compare a sample statistic to a hypothesized population parameter.

INPUT DATA / DATASET:

Step 1. Establish a null and alternative hypothesis.

Step 2. Determine the appropriate statistical test.

Step 3. Set the value of alpha, the Type I error rate.

Step 4. Establish the decision rule.

Step 5. Gather sample data.

Step 6. Analyze the data.

Step 7. Reach a statistical conclusion.

Step 8. Make a business decision

Perform the Following Tests Using Excel and Python Tools:

a. **Two-Sample Mean Test**:
Perform a hypothesis test for comparing the means of two independent samples using the z statistic if population variances are known, or the t statistic if they are unknown.

b. **Two-Sample Proportion Test**:
Perform a hypothesis test for comparing the proportions of two independent samples using the z statistic.

c. **Two-Sample Variance Test**:
Perform a hypothesis test for comparing the variances of two independent samples using the F statistic.

SOURCE CODE (OPTIONAL):

Question: A pharmaceutical company claims that its new drug improves blood pressure more effectively than a competitor's drug. You have two independent samples of patients: one group that received the company's drug and another that received the competitor's drug. The average reduction in systolic blood pressure for the company's drug is 8 mmHg with a standard deviation of 3 mmHg based on 40 patients. For the competitor's drug, the average reduction is 6 mmHg with a standard deviation of 4 mmHg based on 35 patients. Test if there is a significant difference in the effectiveness of the two drugs.

Dataset: Sample data for both groups can be generated using the normal distribution with given means and standard deviations.

Hypothesis:

- Null Hypothesis (H0): The mean reduction in blood pressure is the same for both drugs.
- Alternative Hypothesis (Ha): The mean reduction in blood pressure is different for the two drugs.

CODE:

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

# Parameters
mean1 = 8
std_dev1 = 3
n1 = 40

mean2 = 6
```

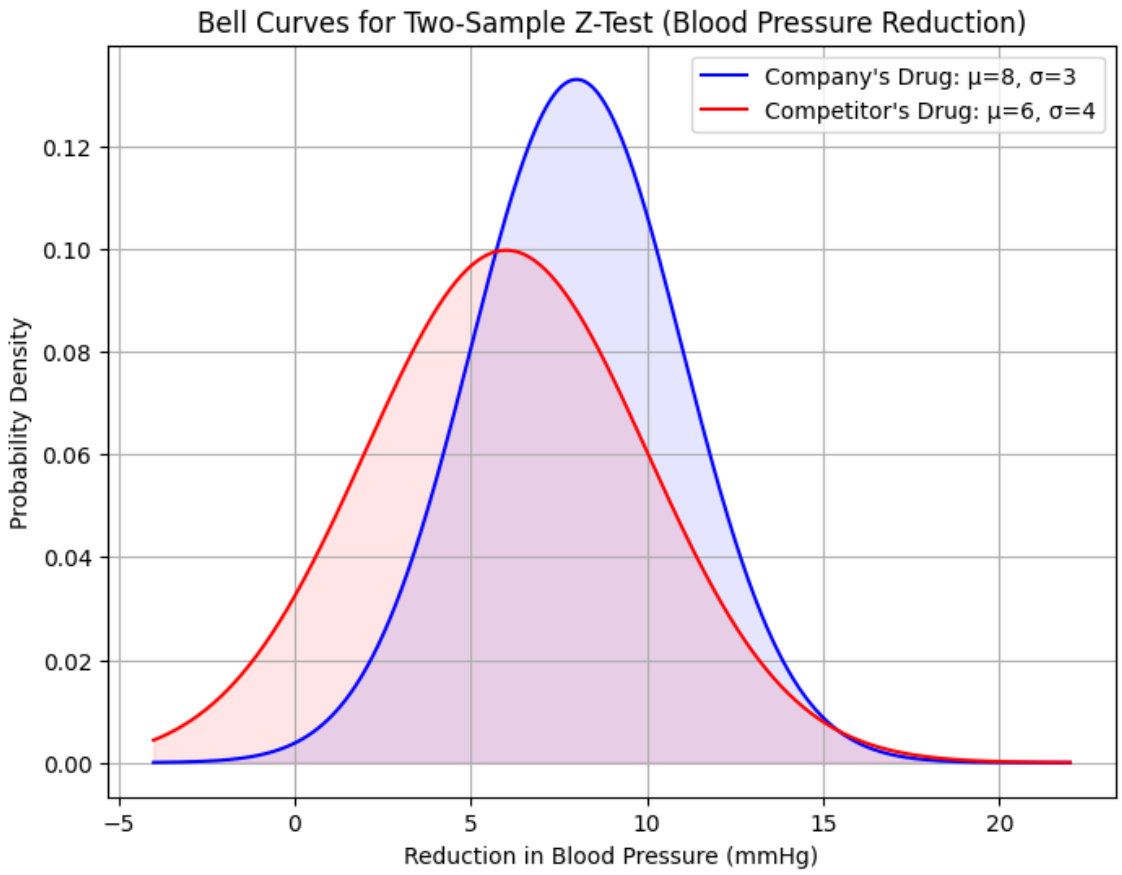
```
std_dev2 = 4
n2 = 35

# Step 1: Establish hypotheses
print("Step 1: Hypotheses")
print("Null Hypothesis (H0): μ1 = μ2 (The mean reduction in blood pressure is the same for both drugs.)")
print("Alternative Hypothesis (Ha): μ1 ≠ μ2 (The mean reduction in blood pressure is different for the two drugs.)\n")

# Step 2: Determine the appropriate test
print("Step 2: Statistical Test")
print("We will use the Two-Sample Z-Test since we know the standard deviations of both populations.\n")

# Step 3: Set alpha
alpha = 0.05
print(f"Step 3: Alpha (Type I error rate) = {alpha}\n")

# Step 4: Establish the decision rule
z_critical = stats.norm.ppf(1 - alpha/2)
print(f"Step 4: Critical Value (z_critical) = ±{z_critical:.2f}\n")

# Step 5: Gather sample data
print("Step 5: Sample Data")
print(f"Population 1: Mean = {mean1}, Std Dev = {std_dev1}, n = {n1}")
print(f"Population 2: Mean = {mean2}, Std Dev = {std_dev2}, n = {n2}\n")

# Step 6: Analyze the data
z = (mean1 - mean2) / np.sqrt((std_dev1**2 / n1) + (std_dev2**2 / n2))
```

```
p_value = 2 * (1 - stats.norm.cdf(np.abs(z)))
print(f"Step 6: Analysis")
print(f"Z-Statistic = {z:.2f}, p-value = {p_value:.4f}\n")

# Step 7: Reach a conclusion
print("Step 7: Conclusion")
if np.abs(z) > z_critical:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
print()

# Step 8: Make a business decision
print("Step 8: Business Decision")
if np.abs(z) > z_critical:
    print("The company's drug shows a statistically significant difference in blood pressure reduction compared to the competitor's drug.")
else:
    print("There is no statistically significant difference in blood pressure reduction between the company's and competitor's drugs.")

# Plotting
x = np.linspace(mean1 - 4 * std_dev1, mean2 + 4 * std_dev2, 1000)
pdf1 = stats.norm.pdf(x, mean1, std_dev1)
pdf2 = stats.norm.pdf(x, mean2, std_dev2)

plt.figure(figsize=(8, 6))
plt.plot(x, pdf1, label=f'Company\'s Drug: μ={mean1}, σ={std_dev1}', color='blue')
plt.plot(x, pdf2, label=f'Competitor\'s Drug: μ={mean2}, σ={std_dev2}', color='red')
plt.fill_between(x, 0, pdf1, color='blue', alpha=0.1)
```

```
plt.fill_between(x, 0, pdf2, color='red', alpha=0.1)
plt.title('Bell Curves for Two-Sample Z-Test (Blood Pressure Reduction)')
plt.xlabel('Reduction in Blood Pressure (mmHg)')
plt.ylabel('Probability Density')
plt.legend()
plt.grid(True)
plt.show()
```

OUTPUT:

```
Step 1: Hypotheses
Null Hypothesis (H0): μ1 = μ2 (The mean reduction in blood pressure is the same for both drugs.)
Alternative Hypothesis (Ha): μ1 ≠ μ2 (The mean reduction in blood pressure is different for the two drugs.)

Step 2: Statistical Test
We will use the Two-Sample Z-Test since we know the standard deviations of both populations.

Step 3: Alpha (Type I error rate) = 0.05

Step 4: Critical Value (z_critical) = ±1.96

Step 5: Sample Data
Population 1: Mean = 8, Std Dev = 3, n = 40
Population 2: Mean = 6, Std Dev = 4, n = 35

Step 6: Analysis
Z-Statistic = 2.42, p-value = 0.0155

Step 7: Conclusion
Reject the null hypothesis.

Step 8: Business Decision
The company's drug shows a statistically significant difference in blood pressure reduction compared to the competitor's drug.
```



Question: A company wants to compare the average productivity of two different work shifts: the morning shift and the evening shift. You collected productivity data (units produced per hour) from 30 employees on the morning shift and 30 employees on the evening shift. The morning shift has an average productivity of 55 units/hour with a standard deviation of 10 units/hour, while the evening shift has an average productivity of 60 units/hour with a standard deviation of 12 units/hour. Determine if there is a significant difference in productivity between the two shifts.

Dataset: Sample data for productivity can be generated using normal distributions with given means and standard deviations.

Hypothesis:

- Null Hypothesis (H0): The average productivity is the same for both shifts.
- Alternative Hypothesis (Ha): The average productivity is different between the two shifts.

CODE:

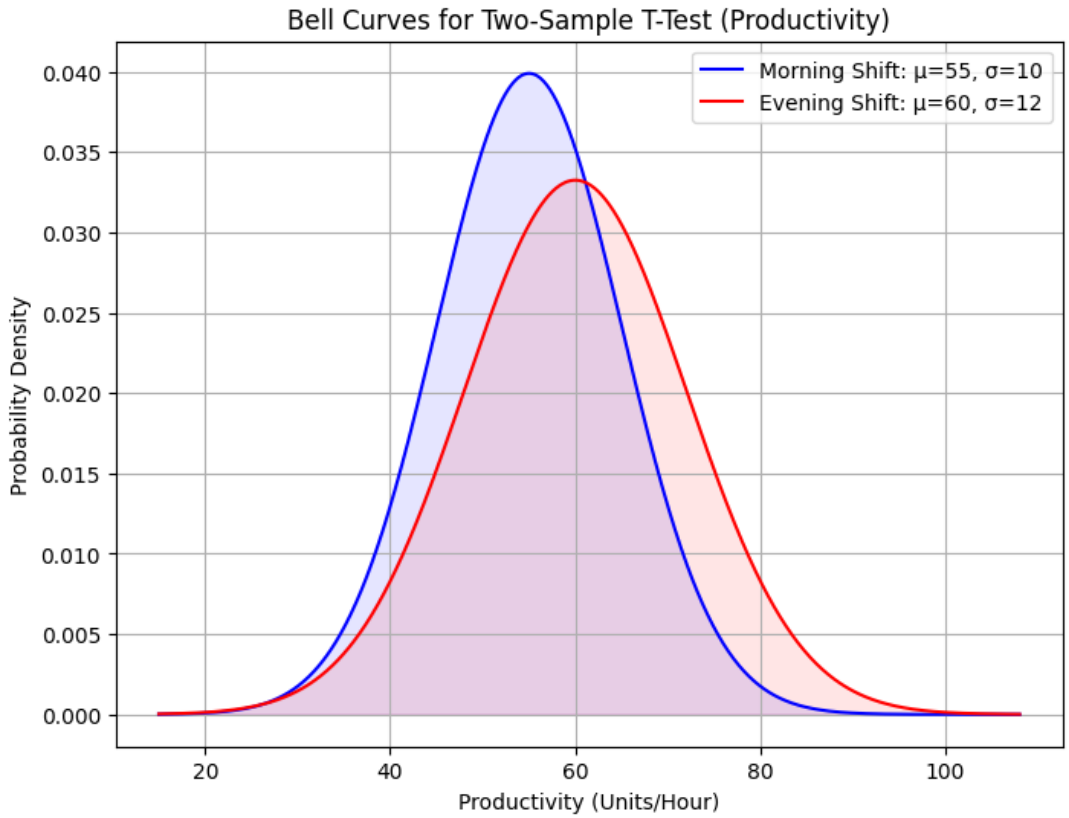
```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# Parameters
mean_morning = 55
std_dev_morning = 10
n_morning = 30

mean_evening = 60
std_dev_evening = 12
n_evening = 30

# Generate sample data
sample_morning = np.random.normal(mean_morning, std_dev_morning, n_morning)
sample_evening = np.random.normal(mean_evening, std_dev_evening, n_evening)

# Step 1: Establish hypotheses
print("Step 1: Hypotheses")
print("Null Hypothesis (H0): μ_morning = μ_evening (The mean productivity is the same for morning and evening shifts.)")
print("Alternative Hypothesis (Ha): μ_morning ≠ μ_evening (The mean productivity is different between shifts.)\n")

# Step 2: Determine the appropriate test
print("Step 2: Statistical Test")
print("We will use the Two-Sample T-Test since we are comparing sample means and assuming unequal variances.\n")
```

```
# Step 3: Set alpha
alpha = 0.05
print(f"Step 3: Alpha (Type I error rate) = {alpha}\n")

# Step 4: Establish the decision rule
df = min(len(sample_morning), len(sample_evening)) - 1
t_critical = stats.t.ppf(1 - alpha/2, df)
print(f"Step 4: Critical Value (t_critical) = ±{t_critical:.2f}\n")

# Step 5: Gather sample data
print("Step 5: Sample Data")
print(f"Morning Shift: Mean = {mean_morning}, Std Dev = {std_dev_morning}, n = {n_morning}")
print(f"Evening Shift: Mean = {mean_evening}, Std Dev = {std_dev_evening}, n = {n_evening}\n")

# Step 6: Analyze the data
t_stat, p_value = stats.ttest_ind(sample_morning, sample_evening, equal_var=False)
print(f"Step 6: Analysis")
print(f"T-Statistic = {t_stat:.2f}, p-value = {p_value:.4f}\n")

# Step 7: Reach a conclusion
print("Step 7: Conclusion")
if np.abs(t_stat) > t_critical:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
print()

# Step 8: Make a business decision
print("Step 8: Business Decision")
if np.abs(t_stat) > t_critical:
```

```
    print("There is a significant difference in productivity between morning and evening shifts.")
else:
    print("There is no significant difference in productivity between morning and evening shifts.")

# Plotting
x = np.linspace(mean_morning - 4 * std_dev_morning, mean_evening + 4 * std_dev_evening, 1000)
pdf_morning = stats.norm.pdf(x, mean_morning, std_dev_morning)
pdf_evening = stats.norm.pdf(x, mean_evening, std_dev_evening)

plt.figure(figsize=(8, 6))
plt.plot(x, pdf_morning, label=f'Morning Shift: μ={mean_morning}, σ={std_dev_morning}', color='blue')
plt.plot(x, pdf_evening, label=f'Evening Shift: μ={mean_evening}, σ={std_dev_evening}', color='red')
plt.fill_between(x, 0, pdf_morning, color='blue', alpha=0.1)
plt.fill_between(x, 0, pdf_evening, color='red', alpha=0.1)
plt.title('Bell Curves for Two-Sample T-Test (Productivity)')
plt.xlabel('Productivity (Units/Hour)')
plt.ylabel('Probability Density')
plt.legend()
plt.grid(True)
plt.show()
```

OUTPUT:

```
Step 1: Hypotheses
Null Hypothesis (H0): μ_morning = μ_evening (The mean productivity is the same for morning and evening shifts.)
Alternative Hypothesis (Ha): μ_morning ≠ μ_evening (The mean productivity is different between shifts.)

Step 2: Statistical Test
We will use the Two-Sample T-Test since we are comparing sample means and assuming unequal variances.

Step 3: Alpha (Type I error rate) = 0.05

Step 4: Critical Value (t_critical) = ±2.05

Step 5: Sample Data
Morning Shift: Mean = 55, Std Dev = 10, n = 30
Evening Shift: Mean = 60, Std Dev = 12, n = 30

Step 6: Analysis
T-Statistic = -3.21, p-value = 0.0022

Step 7: Conclusion
Reject the null hypothesis.

Step 8: Business Decision
There is a significant difference in productivity between morning and evening shifts.
```



Question: A marketing firm wants to test the effectiveness of two different advertising campaigns on customer engagement. They measured the proportion of engaged customers from two different campaigns. In the first campaign, 80 out of 200 customers engaged, and in the second campaign, 75 out of 180 customers engaged. Test if there is a significant difference in the engagement rates between the two campaigns.

Dataset:

- Campaign 1: 80 successes out of 200 trials
- Campaign 2: 75 successes out of 180 trials

Hypothesis:

- Null Hypothesis (H0): The proportion of engaged customers is the same for both campaigns.
- Alternative Hypothesis (Ha): The proportion of engaged customers is different between the two campaigns.

CODE:

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

# Parameters
success1 = 80
n1_proportion = 200

success2 = 75
n2_proportion = 180

# Step 1: Establish hypotheses
print("Step 1: Hypotheses")
```

```
print("Null Hypothesis (H0): p1 = p2 (The proportion of engaged customers is the same for both campaigns.)")
print("Alternative Hypothesis (Ha): p1 ≠ p2 (The proportion of engaged customers is different between the two campaigns.)\n")

# Step 2: Determine the appropriate test
print("Step 2: Statistical Test")
print("We will use the Two-Proportion Z-Test for comparing proportions.\n")

# Step 3: Set alpha
alpha = 0.05
print(f"Step 3: Alpha (Type I error rate) = {alpha}\n")

# Step 4: Establish the decision rule
z_critical = stats.norm.ppf(1 - alpha/2)
print(f"Step 4: Critical Value (z_critical) = ±{z_critical:.2f}\n")

# Step 5: Gather sample data
print("Step 5: Sample Data")
print(f"Campaign 1: Successes = {success1}, Total = {n1_proportion}")
print(f"Campaign 2: Successes = {success2}, Total = {n2_proportion}\n")

# Step 6: Analyze the data
p1 = success1 / n1_proportion
p2 = success2 / n2_proportion
p_pooled = (success1 + success2) / (n1_proportion + n2_proportion)
z = (p1 - p2) / np.sqrt(p_pooled * (1 - p_pooled) * (1/n1_proportion + 1/n2_proportion))
p_value = 2 * (1 - stats.norm.cdf(np.abs(z)))
print(f"Step 6: Analysis")
print(f"Z-Statistic = {z:.2f}, p-value = {p_value:.4f}\n")
```

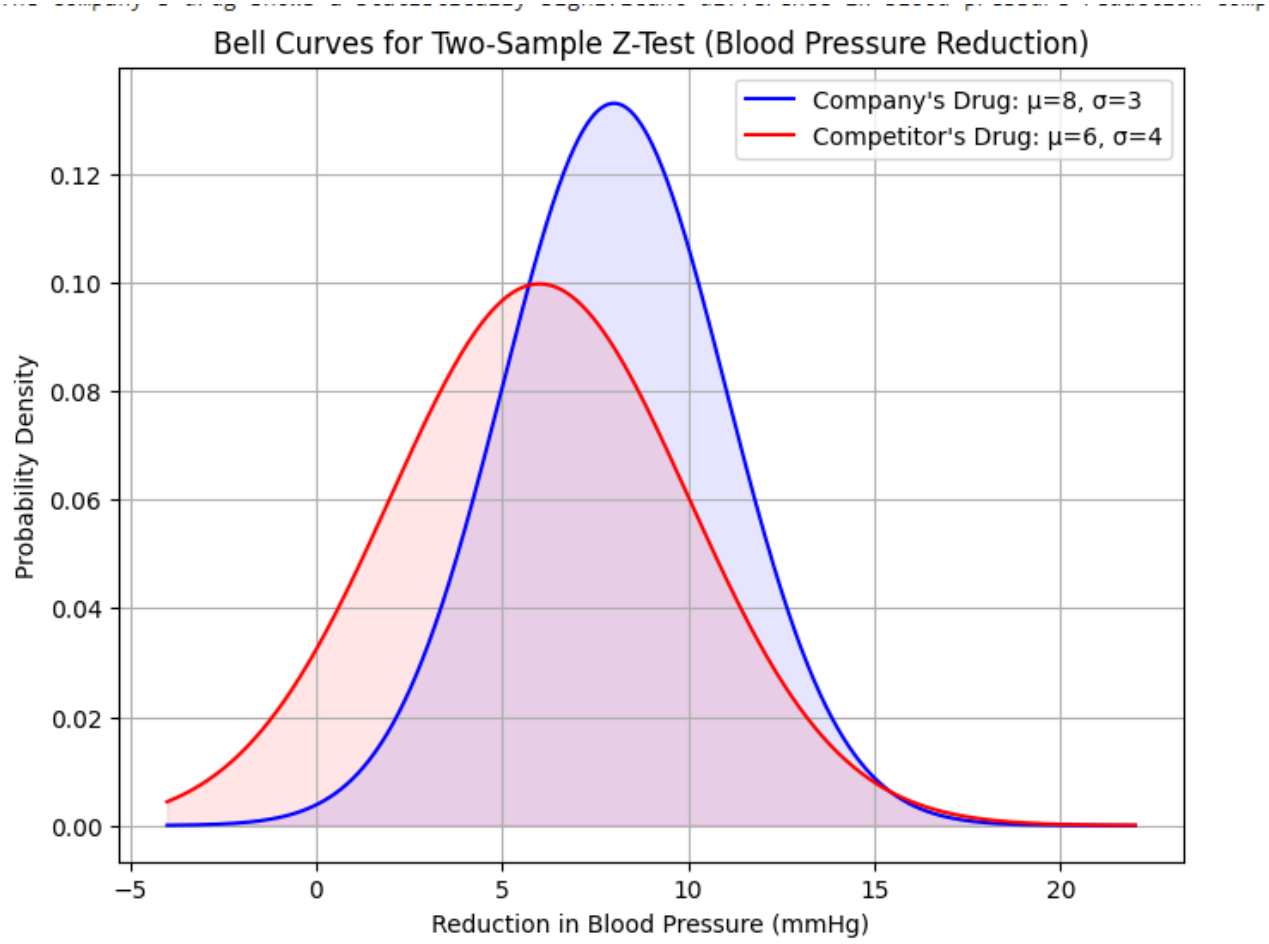
```
# Step 7: Reach a conclusion
print("Step 7: Conclusion")
if np.abs(z) > z_critical:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
print()

# Step 8: Make a business decision
print("Step 8: Business Decision")
if np.abs(z) > z_critical:
    print("There is a significant difference in engagement rates between the two advertising campaigns.")
else:
    print("There is no significant difference in engagement rates between the two advertising campaigns.")

# Plotting
x = np.linspace(0, 1, 1000)
pdf1 = stats.norm.pdf(x, success1 / n1_proportion, np.sqrt((success1 / n1_proportion) * (1 - success1 / n1_proportion) / n1_proportion))
pdf2 = stats.norm.pdf(x, success2 / n2_proportion, np.sqrt((success2 / n2_proportion) * (1 - success2 / n2_proportion) / n2_proportion))

plt.figure(figsize=(8, 6))
plt.plot(x, pdf1, label=f'Campaign 1: p={success1 / n1_proportion:.2f}', color='blue')
plt.plot(x, pdf2, label=f'Campaign 2: p={success2 / n2_proportion:.2f}', color='red')
plt.fill_between(x, 0, pdf1, color='blue', alpha=0.1)
plt.fill_between(x, 0, pdf2, color='red', alpha=0.1)
```

```
plt.title('Probability Curves for Two-Proportion Z-Test (Customer Engagement)')
plt.xlabel('Engagement Rate')
plt.ylabel('Probability Density')
plt.legend()
plt.grid(True)
plt.show()
```

OUTPUT:

```
Step 1: Hypotheses
Null Hypothesis (H0): μ1 = μ2 (The mean reduction in blood pressure is the same for both drugs.)
Alternative Hypothesis (Ha): μ1 ≠ μ2 (The mean reduction in blood pressure is different for the two drugs.)

Step 2: Statistical Test
We will use the Two-Sample Z-Test since we know the standard deviations of both populations.

Step 3: Alpha (Type I error rate) = 0.05

Step 4: Critical Value (z_critical) = ±1.96

Step 5: Sample Data
Population 1: Mean = 8, Std Dev = 3, n = 40
Population 2: Mean = 6, Std Dev = 4, n = 35

Step 6: Analysis
Z-Statistic = 2.42, p-value = 0.0155

Step 7: Conclusion
Reject the null hypothesis.

Step 8: Business Decision
The company's drug shows a statistically significant difference in blood pressure reduction compared to the competitor's drug
```

Bell Curves for Two-Sample Z-Test (Blood Pressure Reduction)



Question: A quality control team wants to compare the variability in the thickness of metal sheets produced by two different machines. They collected thickness measurements from 25 sheets produced by Machine A and 30 sheets produced by Machine B. Machine A has a variance in thickness of 4 mm^2, and Machine B has a variance of 6 mm^2. Test if there is a significant difference in the variability of thickness between the two machines.

Dataset: Sample data for thickness measurements can be generated using normal distributions with given variances.

Hypothesis:

- Null Hypothesis (H0): The variances in thickness are the same for both machines.

- Alternative Hypothesis (Ha): The variances in thickness are different between the two machines.

Feel free to adjust the details of the datasets and questions to better fit your specific needs or scenarios!

CODE:

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

# Parameters
mean1 = 50
std_dev1 = 10
n1 = 40

mean2 = 55
std_dev2 = 12
n2 = 40

# Generate sample data
sample1 = np.random.normal(mean1, std_dev1, n1)
sample2 = np.random.normal(mean2, std_dev2, n2)

# Step 1: Establish hypotheses
print("Step 1: Hypotheses")
print("Null Hypothesis (H0): σ1^2 = σ2^2 (The variability in thickness measurements is the same for both machines.)")
print("Alternative Hypothesis (Ha): σ1^2 ≠ σ2^2 (The variability in thickness measurements is different between the two machines.)\n")
```

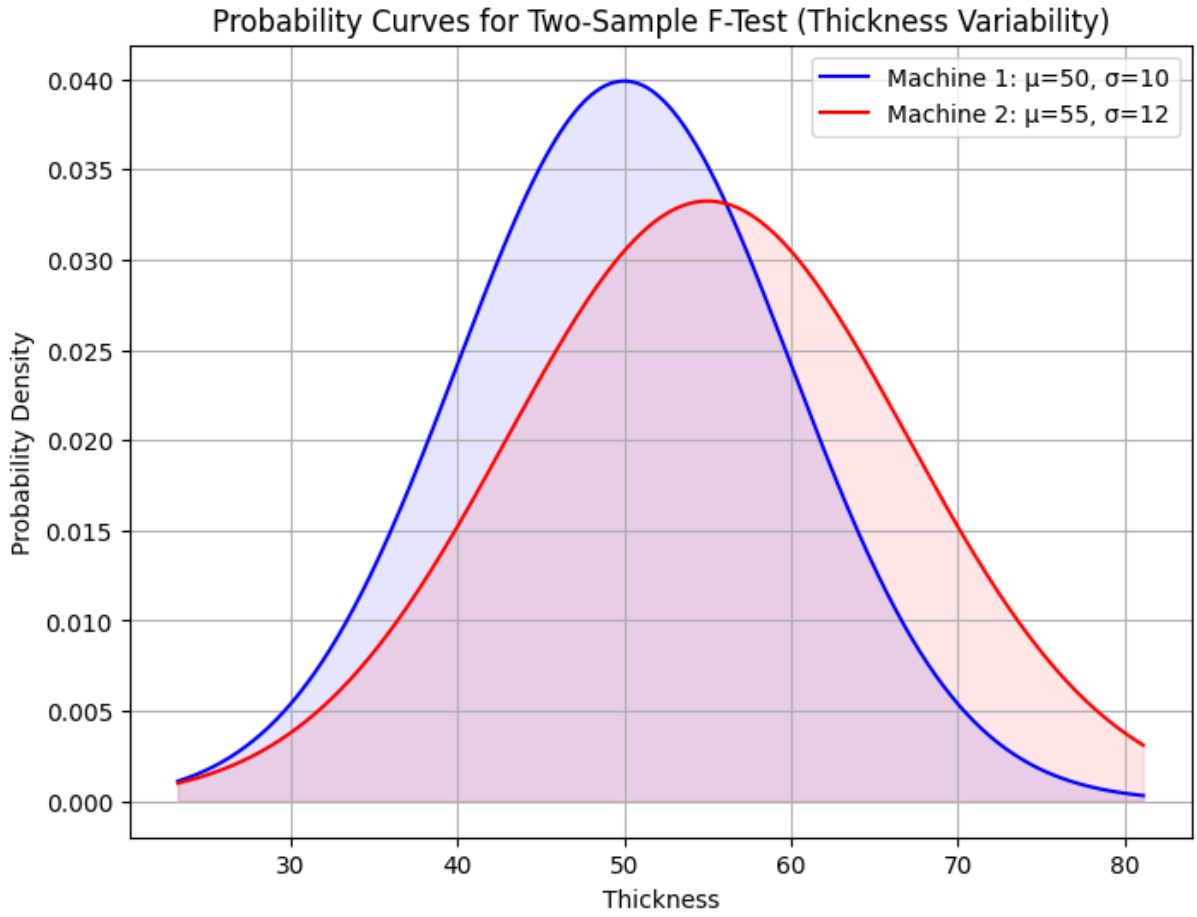
```
# Step 2: Determine the appropriate test
print("Step 2: Statistical Test")
print("We will use the Two-Sample F-Test to compare variances.\n")

# Step 3: Set alpha
alpha = 0.05
print(f"Step 3: Alpha (Type I error rate) = {alpha}\n")

# Step 4: Establish the decision rule
var1 = np.var(sample1, ddof=1)
var2 = np.var(sample2, ddof=1)
f_stat = var1 / var2
df1 = len(sample1) - 1
df2 = len(sample2) - 1
f_critical_low = stats.f.ppf(alpha/2, df1, df2)
f_critical_high = stats.f.ppf(1 - alpha/2, df1, df2)
print(f"Step 4: Critical Values (f_critical) = [{f_critical_low:.2f}, {f_critical_high:.2f}]\n")

# Step 5: Gather sample data
print("Step 5: Sample Data")
print(f"Machine 1: Mean = {mean1}, Std Dev = {std_dev1}, n = {n1}")
print(f"Machine 2: Mean = {mean2}, Std Dev = {std_dev2}, n = {n2}\n")

# Step 6: Analyze the data
p_value = 2 * min(stats.f.cdf(f_stat, df1, df2), 1 - stats.f.cdf(f_stat, df1, df2))
print(f"Step 6: Analysis")
print(f"F-Statistic = {f_stat:.2f}, p-value = {p_value:.4f}\n")

# Step 7: Reach a conclusion
print("Step 7: Conclusion")
```

```
if f_stat < f_critical_low or f_stat > f_critical_high:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
print()

# Step 8: Make a business decision
print("Step 8: Business Decision")
if f_stat < f_critical_low or f_stat > f_critical_high:
    print("There is a significant difference in thickness variability between the two machines.")
else:
    print("There is no significant difference in thickness variability between the two machines.")

# Plotting
x = np.linspace(min(sample1.min(), sample2.min()), max(sample1.max(), sample2.max()), 1000)
pdf1 = stats.norm.pdf(x, mean1, std_dev1)
pdf2 = stats.norm.pdf(x, mean2, std_dev2)

plt.figure(figsize=(8, 6))
plt.plot(x, pdf1, label=f'Machine 1: μ={mean1}, σ={std_dev1}', color='blue')
plt.plot(x, pdf2, label=f'Machine 2: μ={mean2}, σ={std_dev2}', color='red')
plt.fill_between(x, 0, pdf1, color='blue', alpha=0.1)
plt.fill_between(x, 0, pdf2, color='red', alpha=0.1)
plt.title('Probability Curves for Two-Sample F-Test (Thickness Variability)')
plt.xlabel('Thickness')
plt.ylabel('Probability Density')
plt.legend()
plt.grid(True)
plt.show()
```

OUTPUT:

```
Step 1: Hypotheses
Null Hypothesis (H0): σ1^2 = σ2^2 (The variability in thickness measurements is the same for both machines.)
Alternative Hypothesis (Ha): σ1^2 ≠ σ2^2 (The variability in thickness measurements is different between the two machines.)

Step 2: Statistical Test
We will use the Two-Sample F-Test to compare variances.

Step 3: Alpha (Type I error rate) = 0.05

Step 4: Critical Values (f_critical) = [0.53, 1.89]

Step 5: Sample Data
Machine 1: Mean = 50, Std Dev = 10, n = 40
Machine 2: Mean = 55, Std Dev = 12, n = 40

Step 6: Analysis
F-Statistic = 0.64, p-value = 0.1740

Step 7: Conclusion
Fail to reject the null hypothesis.

Step 8: Business Decision
There is no significant difference in thickness variability between the two machines.
```

Probability Curves for Two-Sample F-Test (Thickness Variability)



CONCLUSION:

If the calculated test statistic exceeds the critical value, we reject the null hypothesis, indicating a significant difference between the two populations. If not, we fail to reject the null hypothesis, suggesting no significant difference.

REFERENCES:

Bhattacharyya, G. K., and R. A. Johnson, (1997). Statistical Concepts and Methods, John Wiley and Sons, New York.

Website References:

1. https://www.geo.fu-berlin.de/en/v/soga-r/Basics-of-statistics/Hypothesis-Tests/Hypothesishttps://www.geo.fu-berlin.de/en/v/soga-r/Basics-of-statistics/Hypothesis-

Tests/Hypothesis-Tests-for-One-Population-Mean/Sigma-Is-Known/index.htmlTests-for-One-Population-Mean/Sigma-Is-Known/index.html

2. https://medium.com/%E8%89%BE%E8%9C%9C%E8%8E%89%E8%AE%80%E8%AE%80%E5%AF%AB%E5%AF%AB/datacamp-hypothesis-testing-in-pythonhttps://medium.com/艾蜜莉讀讀寫寫/datacamp-hypothesis-testing-in-python-21427a98735221427a987352
3. https://www.statsmodels.org/dev/generated/statsmodels.stats.proportion.proportions_ztest.html
4. https://openstax.org/books/introductory-statistics-2e/pages/9-6-hypothesis-testing-of-asingle-mean-and-single-proportion