

**Data Science**  
**PRACTICAL NO. 4**

**Aim: Hypothesis Testing -**

- a) Formulate null and alternative hypotheses for a given problem.
- b) Conduct a hypothesis test using appropriate statistical tests (e.g., t-test, chi-square test).
- c) Interpret the results and draw conclusions based on the test outcomes.

**CODE:**

➤ **Importing libraries**

```
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
df = pd.read_csv("tip.csv")
df.head()
```

➤ **1. One-Sample t-test (Mean Tip Amount)**

```
data = df['tip'].values
print("Data:", data)
H0 = "Average tip = 3 dollars"
H1 = "Average tip ≠ 3 dollars"
t_stat, p_value = stats.ttest_1samp(data, 3)
print("Test Statistic:", t_stat)
print("P-value:", p_value)
if p_value < 0.05:
    print("Reject H0 → Mean tip is significantly different from 3.")
else:
    print("Fail to Reject H0 → No evidence tip is different from 3.")

plt.figure(figsize=(6,5))
plt.hist(data, bins=20)
plt.axvline(3, color='red', linestyle='--', label='Hypothesized Mean = 3')
plt.title("One-Sample t-test: Distribution of Tips")
plt.xlabel("Tip Amount")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

➤ **2. Independent Two-Sample t-test (Male vs Female Tip Amount)**

```
from scipy.stats import ttest_ind
male = df[df['sex']=="Male"]['tip']
female = df[df['sex']=="Female"]['tip']
print("Male mean:", np.mean(male))
print("Female mean:", np.mean(female))
ttest, pval = ttest_ind(male, female)
print("p-value:", pval)
if pval < 0.05:
    print("Reject H0 → Males and Females give different tips.")
else:
    print("Accept H0 → No difference in tip amounts by gender.")

plt.figure(figsize=(6,5))
plt.boxplot([male, female], labels=['Male', 'Female'])
```

**Data Science**  
**PRACTICAL NO. 4**

```
plt.title("Independent t-test: Tips by Gender")
plt.ylabel("Tip Amount")
plt.show()
```

➤ **3. Paired Sample t-test**

```
from scipy import stats
ttest, pval = stats.ttest_rel(df['total_bill'], df['tip'])
print("p-value:", pval)
if pval < 0.05:
    print("Reject H0 → Significant difference between total bill & tip.")
else:
    print("Accept H0 → No significant difference.")
```

```
plt.figure(figsize=(6,5))
plt.scatter(df['total_bill'], df['tip'])
plt.title("Paired t-test: Total Bill vs Tip")
plt.xlabel("Total Bill")
plt.ylabel("Tip")
plt.show()
```

➤ **4. One-Sample Z-test (Using Total Bill)**

```
from statsmodels.stats import weightstats as stests
ztest, pval = stests.ztest(df['total_bill'], value=20)
print("p-value:", float(pval))
if pval < 0.05:
    print("Reject H0 → Mean total bill ≠ 20.")
else:
    print("Accept H0 → Evidence supports mean = 20.")
```

```
plt.figure(figsize=(6,5))
plt.hist(df['total_bill'], bins=20)
plt.axvline(20, color='red', linestyle='--', label='Hypothesized Mean = 20')
plt.title("Z-Test: Distribution of Total Bill")
plt.xlabel("Total Bill")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

➤ **5. Two-Sample Z-test (Smoker vs Non-Smoker Total Bill)**

```
smoker = df[df['smoker']=="Yes"]['total_bill']
non_smoker = df[df['smoker']=="No"]['total_bill']
ztest, pval = stests.ztest(smoker, x2=non_smoker, value=0)
print("p-value:", float(pval))
if pval < 0.05:
    print("Reject H0 → Smokers & Non-smokers differ in spending.")
else:
    print("Accept H0 → No difference in spending.")
```

```
plt.figure(figsize=(6,5))
plt.boxplot([smoker, non_smoker], labels=['Smoker', 'Non-Smoker'])
plt.title("Two-Sample Z-Test: Total Bill Comparison")
plt.ylabel("Total Bill")
plt.show()
```

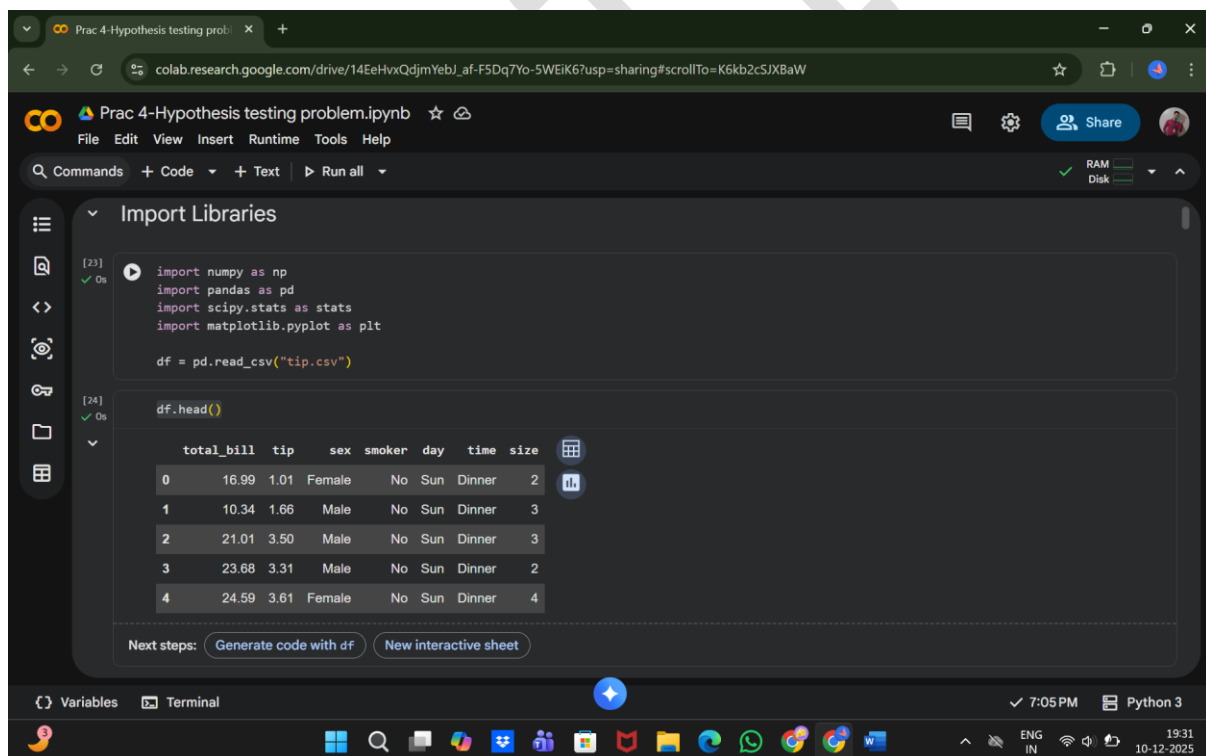
Data Science  
PRACTICAL NO. 4

➤ 6. Chi-Square Test (Gender vs Smoker Status)

```
from scipy.stats import chi2_contingency
table = pd.crosstab(df['sex'], df['smoker'])
print("Contingency Table:\n", table)
chi2, pval, dof, expected = chi2_contingency(table)
print("Chi-square:", chi2)
print("p-value:", pval)
print("Degrees of freedom:", dof)
print("Expected frequencies:\n", expected)
if pval < 0.05:
    print("Reject H0 → Gender and Smoking are dependent.")
else:
    print("Accept H0 → Gender and Smoking are independent.")
```

```
plt.figure(figsize=(6,5))
table.plot(kind='bar', figsize=(6,5))
plt.title("Chi-Square Test: Gender vs Smoker Status")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.show()
```

Output:



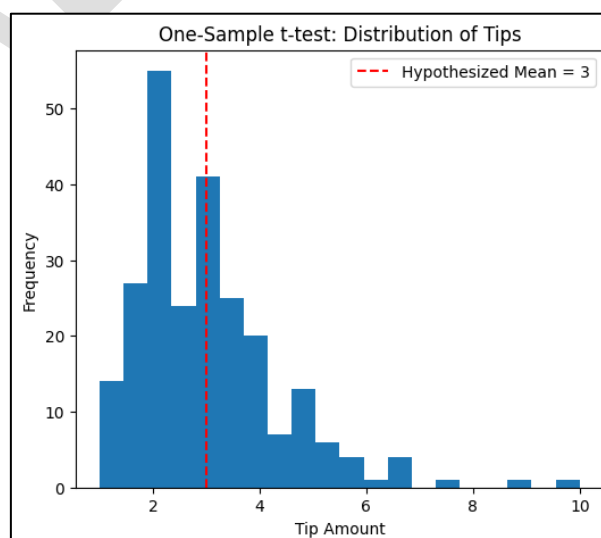
# Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce

## Data Science

### PRACTICAL NO. 4

```
Prac 4-Hypothesis testing problem.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
RAM Disk
1. One-Sample t-test (Mean Tip Amount)
Problem: To check whether the average tip given by customers is equal to 3 USD.
Null Hypothesis (H0): The mean tip amount is 3 USD. ( $\mu_{tip} = 3$ )
Alternative Hypothesis (H1): The mean tip amount is not equal to 3 USD. ( $\mu_{tip} \neq 3$ )
In [2]: data = df['tip'].values
print("Data:", data)
H0 = "Average tip = 3 dollars"
H1 = "Average tip != 3 dollars"
t_stat, p_value = stats.ttest_1samp(data, 3)
print("Test Statistic:", t_stat)
print("P-value:", p_value)
if p_value < 0.05:
    print("Reject H0 -> Mean tip is significantly different from 3.")
else:
    print("Fail to Reject H0 -> No evidence tip is different from 3.")
... Data: [ 1.01 1.65 3.5 3.31 3.61 4.71 2. 3.12 1.96 3.23 1.71 5.]
Variables Terminal 7:05 PM Python 3 19:32 10-12-2025
```

```
Prac 4-Hypothesis testing problem.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
RAM Disk
...
5. 3.75 2.61 2. 3.5 2.5 2. 2. 3. 3.48 2.24 4.5
1.61 2. 10. 3.16 5.15 3.18 4. 3.11 2. 2. 4. 3.55
3.68 5.65 3.5 6.5 3. 5. 3.5 2. 3.5 4. 1.5 4.19
2.56 2.02 4. 1.44 2. 5. 2. 2. 4. 2.01 2. 2.5
4. 3.23 3.41 3. 2.03 2.23 2. 5.16 9. 2.5 6.5 1.1
3. 1.5 1.44 3.09 2.2 3.48 1.92 3. 1.58 2.5 2. 3.
2.72 2.88 2. 3. 3.39 1.47 3. 1.25 1. 1.17 4.67 5.92
2. 2. 1.75 3. ]
Test Statistic: -0.019432641422916876
P-value: 0.9845119176410544
Fail to Reject H0 -> No evidence tip is different from 3.
The calculated test statistic is -0.019 and the p-value is 0.9845, which is much greater than the significance level  $\alpha = 0.05$ . Since the p-value is high, we fail to reject the null hypothesis (H0). There is no statistically significant difference between the sample mean tip amount and 3 USD. This means that, based on our sample data, the average tip given by customers is approximately equal to 3 USD, and we do not have enough evidence to say it is higher or lower.
In [26]: plt.figure(figsize=(6,5))
plt.hist(data, bins=20)
plt.axvline(3, color='red', linestyle='--', label='Hypothesized Mean = 3')
plt.title("One-Sample t-test: Distribution of Tips")
plt.xlabel("Tip Amount")
plt.ylabel("Frequency")
plt.legend()
plt.show()
Variables Terminal 7:05 PM Python 3 19:33 10-12-2025
```



## Data Science PRACTICAL NO. 4

Prac 4-Hypothesis testing problem.ipynb

The histogram shows how tips are distributed. The red dashed line marks the expected mean (3). If the bars pile far away from the line, it suggests the actual average differs from 3 which the t-test confirms using probability (p-value).

### 2. Independent Two-Sample t-test (Male vs Female Tip Amount)

Problem: To test whether male and female customers give the same average tip.

Null Hypothesis ( $H_0$ ): The mean tip from male customers is equal to the mean tip from female customers. ( $\mu_{\text{male}} = \mu_{\text{female}}$ )

Alternative Hypothesis ( $H_1$ ): The mean tip from male customers is different from the mean tip from female customers. ( $\mu_{\text{male}} \neq \mu_{\text{female}}$ )

```
[27] ✓ Os
from scipy.stats import ttest_ind

male = df[df['sex']=='Male']['tip']
female = df[df['sex']=='Female']['tip']

print("Male mean:", np.mean(male))
print("Female mean:", np.mean(female))

ttest, pval = ttest_ind(male, female)

print("p-value:", pval)

if pval < 0.05:
```

Variables Terminal 7:05 PM Python 3 1934 10-12-2025

Prac 4-Hypothesis testing problem.ipynb

```
[27] ✓ Os
print("Reject H0 - Males and Females give different tips.")
else:
    print("Accept H0 - No difference in tip amounts by gender.")
```

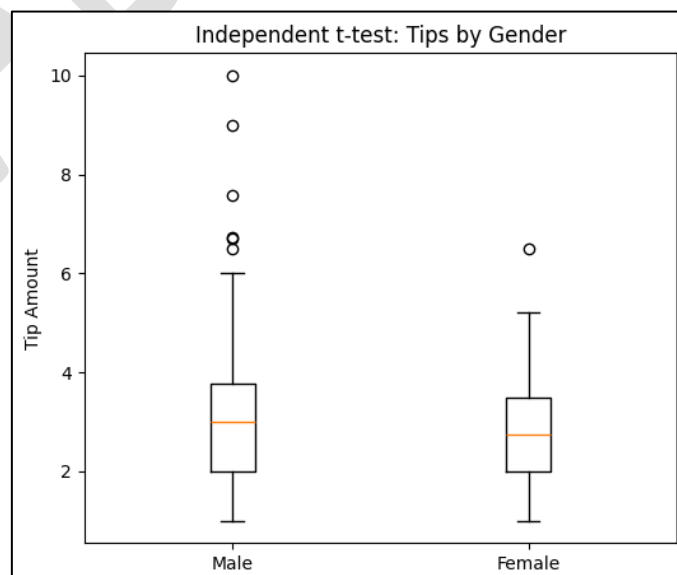
Male mean: 3.0896178343949843  
Female mean: 2.8334482758620685  
p-value: 0.16645623503456755  
Accept H0 - No difference in tip amounts by gender.

The mean tip amounts are:  
Male customers: 3.0896  
Female customers: 2.8334

The p-value is 0.1664, which is greater than the significance level  $\alpha = 0.05$ . Since the p-value is high, we accept the null hypothesis ( $H_0$ ). There is no statistically significant difference in tip amounts between male and female customers. Although males appear to tip slightly more on average, the difference is not strong enough to be considered meaningful based on statistical evidence. Therefore, gender does not influence tipping behavior in this dataset.

```
[28] ✓ Os
plt.figure(figsize=(6,5))
plt.boxplot([male, female], labels=['Male', 'Female'])
plt.title("Independent t-test: Tips by Gender")
plt.ylabel("Tip Amount")
plt.show()
```

Variables Terminal 7:05 PM Python 3 1934 10-12-2025



## Data Science PRACTICAL NO. 4

Each box shows how tips vary for males vs females.  
If the boxes overlap heavily → tips are similar.  
If they separate → tips differ.  
The t-test gives a statistical confirmation through the p-value

### 3. Paired Sample t-test

Problem: To test whether there is a significant difference in two related measurements (e.g., before vs after condition).  
Null Hypothesis ( $H_0$ ): The mean difference between the two related samples is zero. ( $\mu_{\text{before}} - \mu_{\text{after}} = 0$ )  
Alternative Hypothesis ( $H_1$ ): The mean difference between the two related samples is not zero. ( $\mu_{\text{before}} - \mu_{\text{after}} \neq 0$ )

```
[29] from scipy import stats

ttest, pval = stats.ttest_rel(df['total_bill'], df['tip'])

print("p-value:", pval)

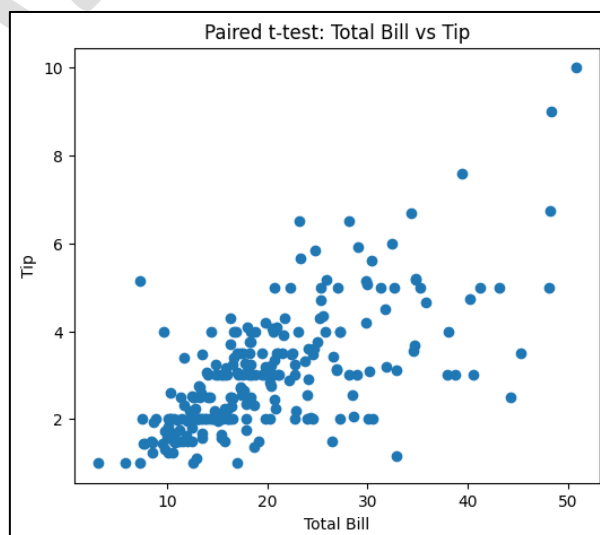
if pval < 0.05:
    print("Reject H0 → Significant difference between total bill & tip.")
else:
```

```
[29] else:
    print("Accept H0 → No significant difference.")

p-value: 8.020818605920848e-91
Reject H0 → Significant difference between total bill & tip.
```

The paired t-test produced a p-value of  $8.02 \times 10^{-91}$ , which is extremely smaller than the significance level  $\alpha = 0.05$ . Therefore, we reject the null hypothesis ( $H_0$ ).  
There is a highly significant statistical difference between the two related measurements. The extremely small p-value indicates that the before and after values are not the same, and the difference is not due to random chance. This means that the condition/event between the two measurements has a real impact, causing a measurable change between the paired values.

```
[30] plt.figure(figsize=(6,5))
plt.scatter(df['total_bill'], df['tip'])
plt.title("Paired t-test: Total Bill vs Tip")
plt.xlabel("Total Bill")
plt.ylabel("Tip")
plt.show()
```



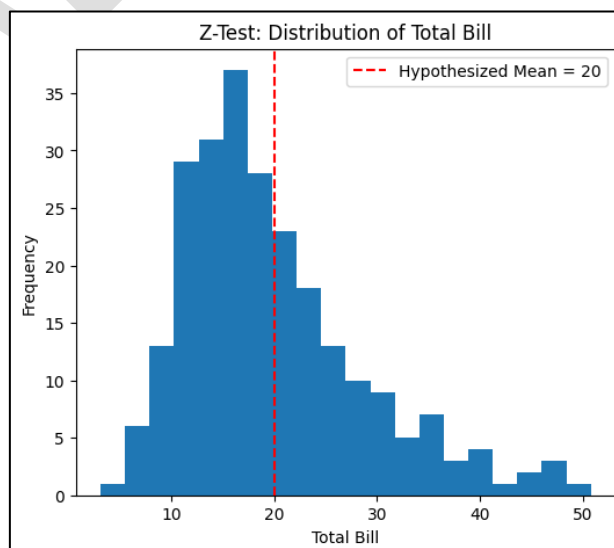
# Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce

## Data Science

### PRACTICAL NO. 4

```
Prac 4-Hypothesis testing problem.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
The scatterplot shows the relationship between bill and tip.
More spread → greater differences, closer to a straight line → similar values.
The paired t-test mathematically checks if the difference is significant.
4. One-Sample Z-test (Using Total Bill)
Problem: To check whether the average total bill in the population is equal to a specified value (e.g., 20 USD).
Null Hypothesis ( $H_0$ ): The population mean total bill is equal to the claimed value. ( $\mu = 20$ )
Alternative Hypothesis ( $H_1$ ): The population mean total bill is not equal to the claimed value. ( $\mu \neq 20$ )
[31] ✓ 0s
from statsmodels.stats import weightstats as stests
ztest, pval = stests.ztest(df['total_bill'], value=20)
print("p-value:", float(pval))
if pval < 0.05:
    print("Reject  $H_0$  → Mean total bill  $\neq 20$ .")
else:
    print("Accept  $H_0$  → Evidence supports mean = 20.")
Variables Terminal 7:05 PM Python 3 19:37 10-12-2025
```

```
Prac 4-Hypothesis testing problem.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
p-value: 0.7072195368238167
Accept  $H_0$  → Evidence supports mean = 20.
The Z-test gives a p-value of 0.7072, which is much greater than the significance level  $\alpha = 0.05$ . Therefore, we accept the null hypothesis ( $H_0$ ).
There is no statistically significant difference between the sample mean total bill and the claimed value of 20 USD. Since the p-value is high, the data does not provide enough evidence to say that the average total bill is higher or lower than 20 USD. Thus, the sample supports the claim that the population mean total bill is approximately equal to 20 USD.
[32] ✓ 0s
plt.figure(figsize=(6,5))
plt.hist(df['total_bill'], bins=20)
plt.axvline(20, color='red', linestyle='--', label='Hypothesized Mean = 20')
plt.title("Z-Test: Distribution of Total Bill")
plt.xlabel("Total Bill")
plt.ylabel("Frequency")
plt.legend()
plt.show()
Variables Terminal 7:05 PM Python 3 19:37 10-12-2025
```



## Data Science PRACTICAL NO. 4

Prac 4-Hypothesis testing problem.ipynb

The distribution shows how total bills spread around the hypothesized mean (20). If the actual mean shifts left/right from the line, the z-test will detect it with a small p-value.

### 5. Two-Sample Z-test (Smoker vs Non-Smoker Total Bill)

Problem: To test whether smokers and non-smokers spend the same amount on total bill.

Null Hypothesis ( $H_0$ ): The mean total bill for smokers and non-smokers is equal. ( $\mu_{\text{smoker}} = \mu_{\text{non\_smoker}}$ )

Alternative Hypothesis ( $H_1$ ): The mean total bill for smokers and non-smokers is not equal. ( $\mu_{\text{smoker}} \neq \mu_{\text{non\_smoker}}$ )

```
[33] ✓ On
smoker = df[df['smoker']=='Yes']['total_bill']
non_smoker = df[df['smoker']=='No']['total_bill']

ztest, pval = stats.ztest(smoker, x2=non_smoker, value=0)

print("p-value:", float(pval))

if pval < 0.05:
    print("Reject H0 -> Smokers & Non-smokers differ in spending.")
else:
    print("Accept H0 -> No difference in spending.")
```

Variables Terminal

7:05 PM Python 3

Prac 4-Hypothesis testing problem.ipynb

p-value: 0.18075422519296214  
Accept H0 -> No difference in spending.

The two-sample Z-test gives a p-value of 0.1807, which is greater than the significance level  $\alpha = 0.05$ . Therefore, we accept the null hypothesis ( $H_0$ ).

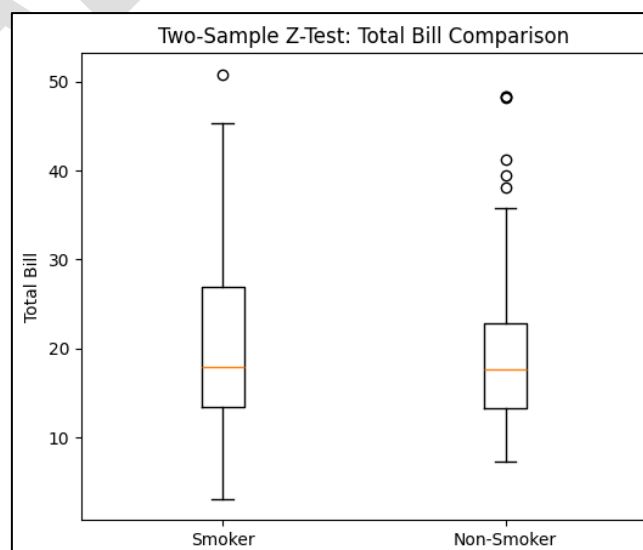
There is no statistically significant difference in the average total bill between smokers and non-smokers. Although their mean spending levels may differ slightly in the sample, the difference is not strong enough to be considered meaningful statistically. Thus, based on the data, smoking status does not influence how much customers spend on their total bill.

```
[34] ✓ On
plt.figure(figsize=(6,5))
plt.boxplot([smoker, non_smoker], labels=['Smoker', 'Non-Smoker'])
plt.title("Two-Sample Z-Test: Total Bill Comparison")
plt.ylabel("Total Bill")
plt.show()
```

/tmp/ipython-input-1364324761.py:2: MatplotlibDeprecationWarning: The 'labels' parameter of boxplot() has been renamed 'tick\_labels' since Ma

Variables Terminal

7:05 PM Python 3





# Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce

## Data Science

### PRACTICAL NO. 4

```
Prac 4-Hypothesis testing problem.ipynb
```

The boxplot compares how much smokers vs non-smokers spend.  
If boxes overlap → similar.  
If they separate → different.  
The z-test confirms statistically.

#### 6. Chi-Square Test (Gender vs Smoker Status)

Problem: To check whether gender and smoking status are related.  
Null Hypothesis ( $H_0$ ): Gender and smoking status are independent. (No relationship)  
Alternative Hypothesis ( $H_1$ ): Gender and smoking status are dependent. (There is a relationship)

```
[15]: from scipy.stats import chi2_contingency

table = pd.crosstab(df['sex'], df['smoker'])
print("Contingency Table:\n", table)

chi2, pval, dof, expected = chi2_contingency(table)

print("Chi-square:", chi2)
```

```
[15]: print("p-value:", pval)
print("Degrees of freedom:", dof)
print("Expected frequencies:\n", expected)

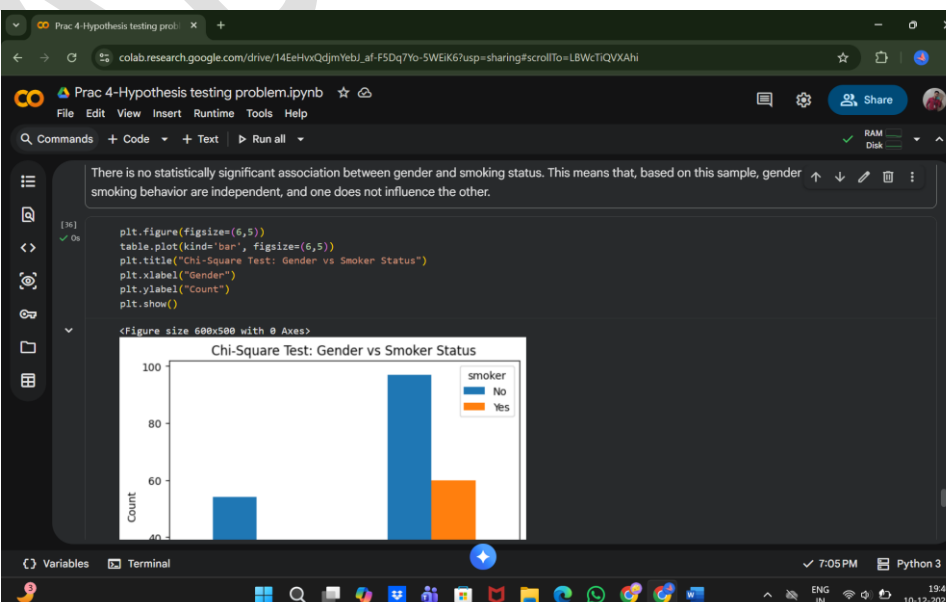
if pval < 0.05:
    print("Reject H0 → Gender and Smoking are dependent.")
else:
    print("Accept H0 → Gender and Smoking are independent.")
```

Contingency Table:

	smoker No	smoker Yes
sex		
Female	54	33
Male	97	60

Chi-square: 0.0  
p-value: 1.0  
Degrees of freedom: 1  
Expected frequencies:  
[[53.84016393 33.15983607]  
[97.15983607 59.84016393]]  
Accept H0 → Gender and Smoking are independent.

The Chi-square statistic = 0.0, p-value = 1.0, and degrees of freedom = 1.  
Since the p-value is much greater than 0.05, we accept the null hypothesis ( $H_0$ ).



Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce  
Data Science  
PRACTICAL NO. 4

