

**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'])
```

```
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.weightstats import ztest
ztest, pval = 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 = ztest(smoker, 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()
```

#### ➤ 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:



A screenshot of a Jupyter Notebook interface. The title bar shows the URL: colab.research.google.com/drive/14EeHvxQdjmYebJ\_af-F5Dq7Yo-5WEIK6?usp=sharing#scrollTo=K6kb2cSJXBaW. The notebook file name is "Prac 4-Hypothesis testing problem.ipynb". The code cell [23] contains the following Python code:

```
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt

df = pd.read_csv("tip.csv")
```

The code cell [24] contains the command `df.head()`. Below the code cells, a data preview shows the first five rows of the "tip" dataset:

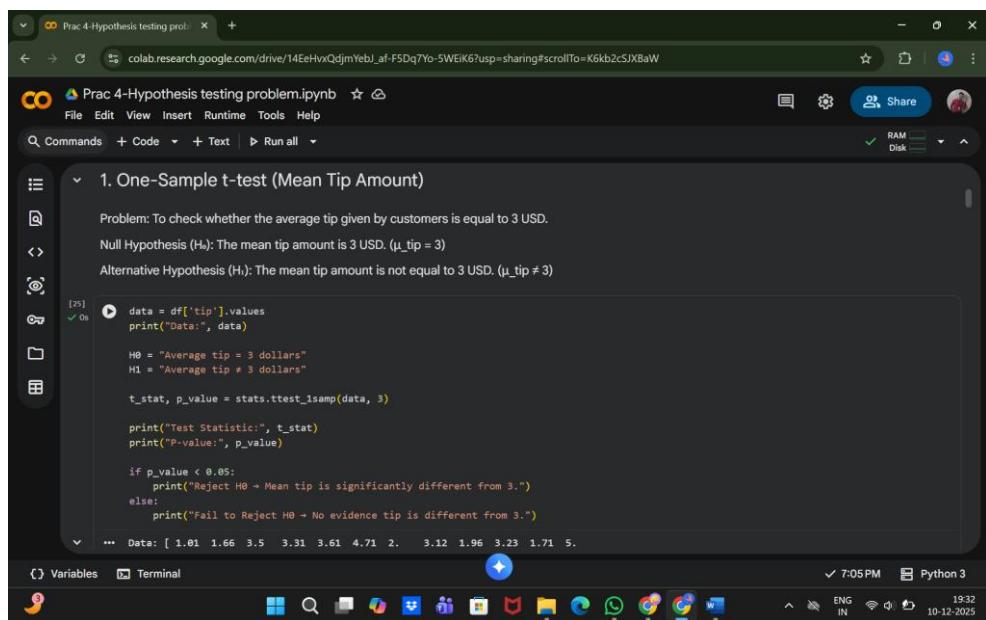
	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

At the bottom of the interface, there are buttons for "Variables", "Terminal", and "Python 3". The status bar shows the time as 7:05 PM and the date as 10-12-2025.

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

## Data Science

### PRACTICAL NO. 4



```
[25] 0b
  data = df['tip'].values
  print("Data:", data)

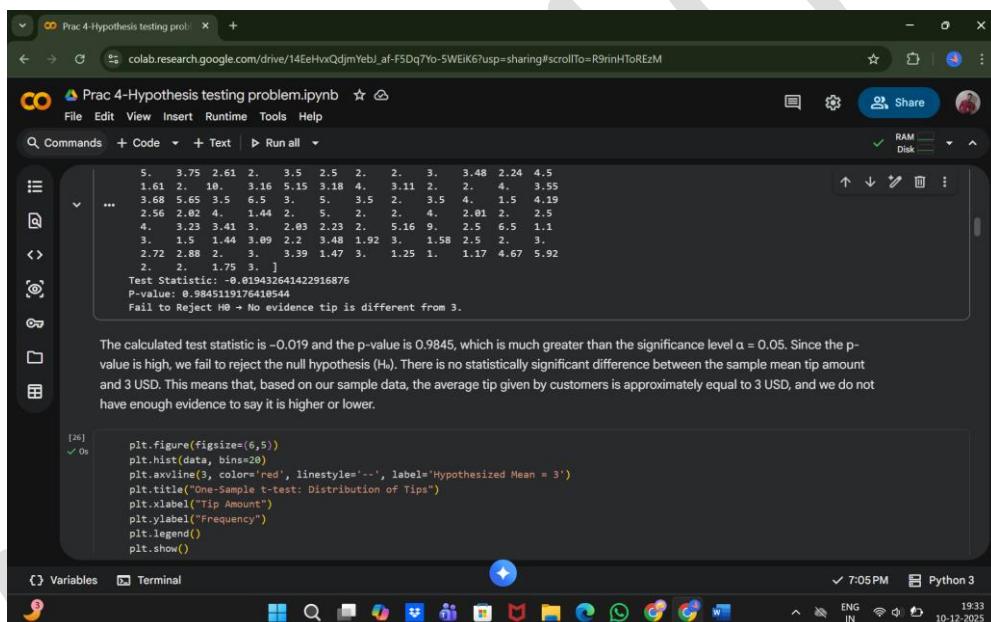
  H0 = "Average tip = 3 dollars"
  H1 = "Average tip ≠ 3 dollars"

  t_stat, p_value = stats.ttest_1samp(data, 3)

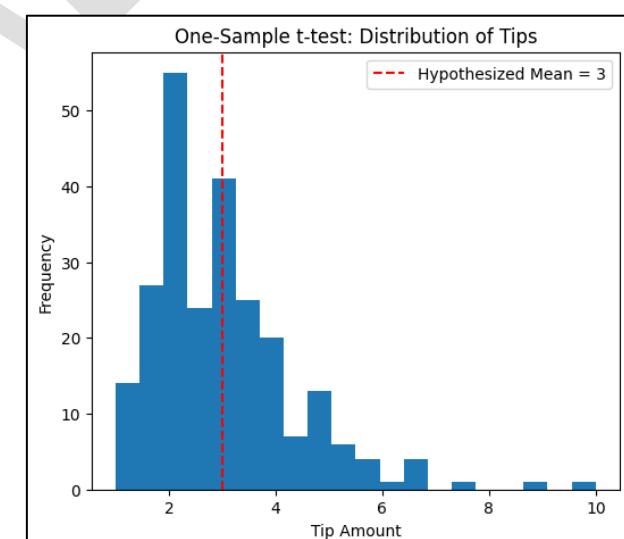
  print("Test Statistic:", t_stat)
  print("P-value:", p_value)

  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 1.66 3.5 3.31 3.61 4.71 2. 3.12 1.96 3.23 1.71 5.
```



```
5. 3.75 2.61 2. 3.5 2.5 2. 2. 3. 3.48 2.24 4.5
1.61 2. 10. 5.16 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.
```



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

## Data Science

### PRACTICAL NO. 4

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. ( $\mu_{\text{male}} \neq \mu_{\text{female}}$ )

```
#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.")

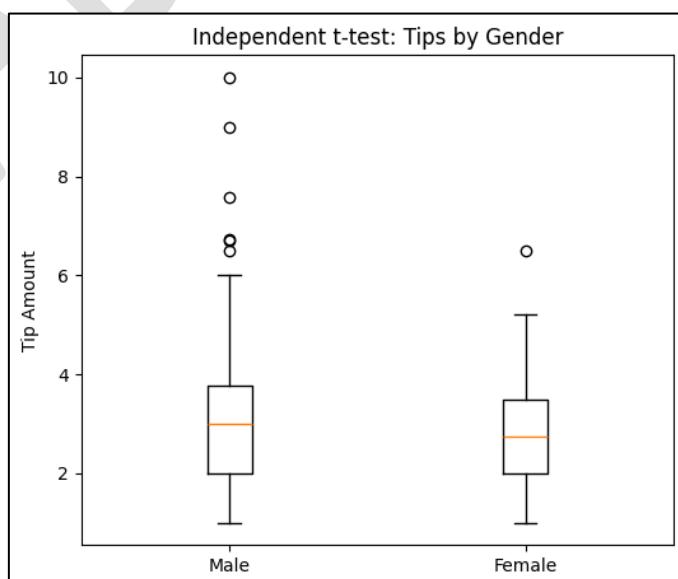
Male mean: 3.0896178343949843
Female mean: 2.833482758670685
p-value: 0.16645623583456755
Accept H0 - No difference in tip amounts by gender.

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

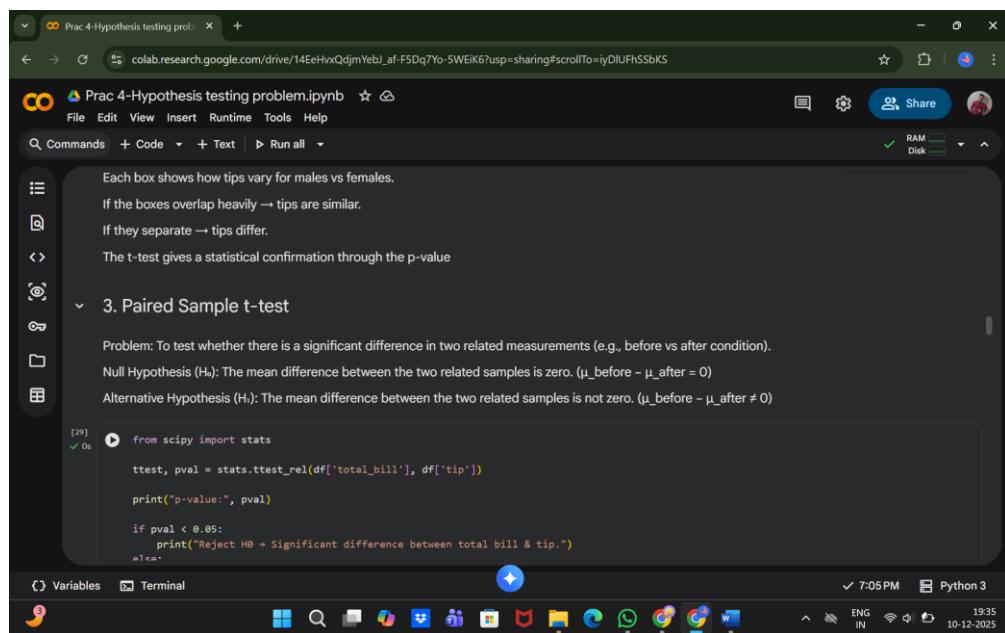
The p-value is 0.1664, which is greater than the significance level  $\alpha = 0.05$ . Since the p-value is high, we accept the null hypothesis ( $H_0$ ).
```

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.

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



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

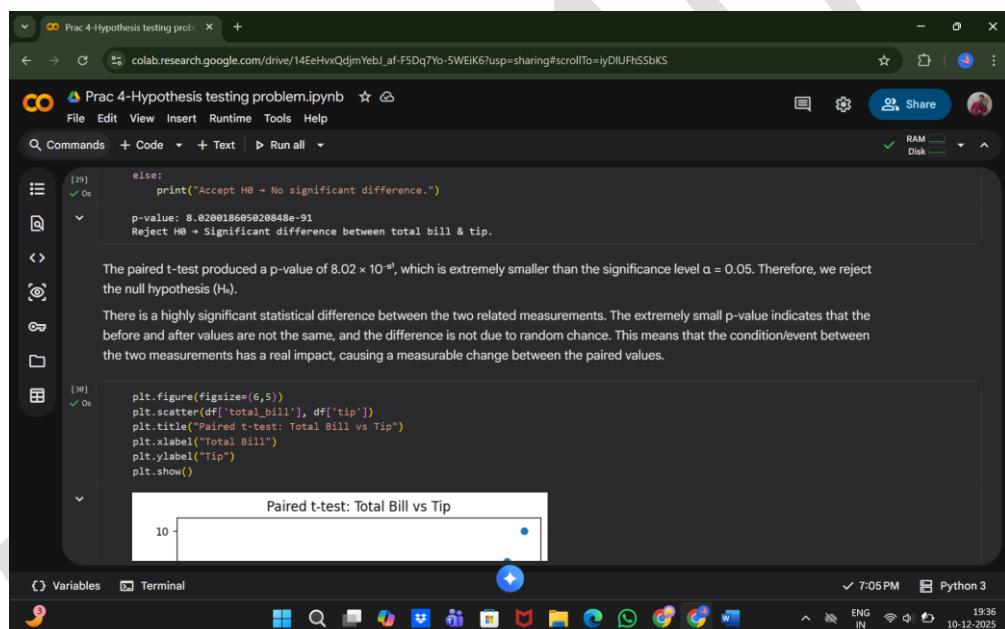


Each box shows how tips vary for males vs females.  
If the boxes overlap heavily → tips are similar.  
If they separate → tips differ.  
The t-test gives a statistical confirmation through the p-value

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 ( $H_a$ ): 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.")
```

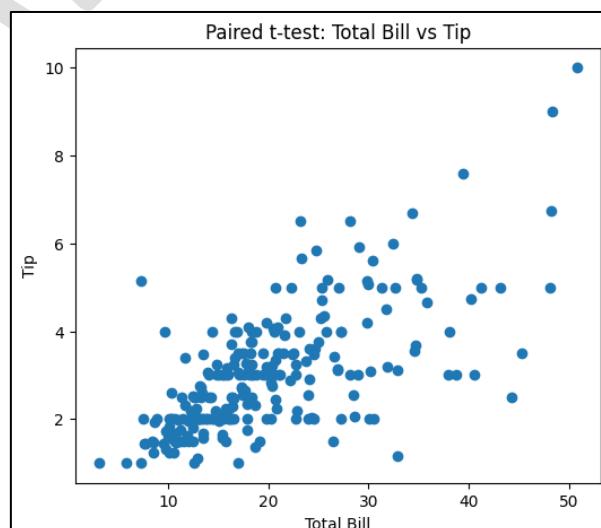


```
[29] else:
      print("Accept H0 → No significant difference.")
p-value: 8.020018665020848e-91
Reject H0 → Significant difference between total bill & tip.
```

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

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)

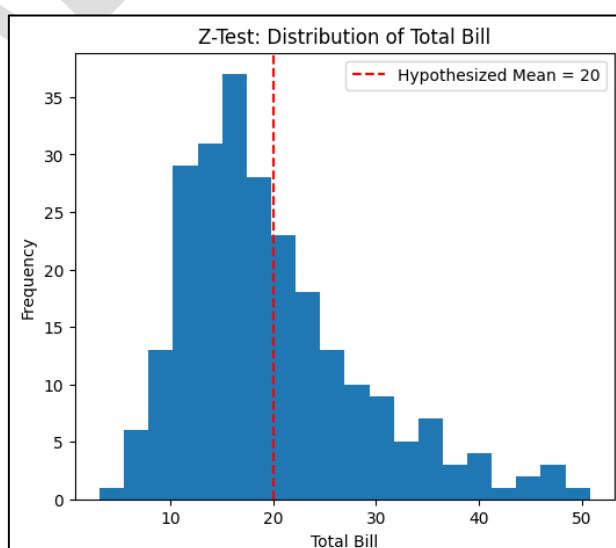
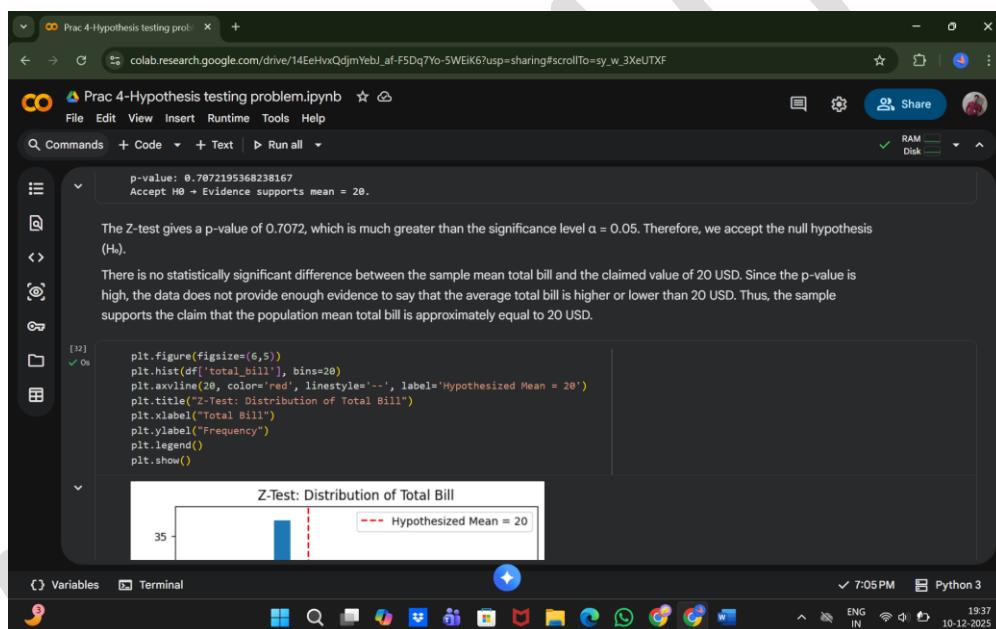
Problem: To check whether the average total bill in the population is equal to a specified value (e.g., 20 USD).

Null Hypothesis ( $H_0$ ): The population mean total bill is equal to the claimed value. ( $\mu = 20$ )

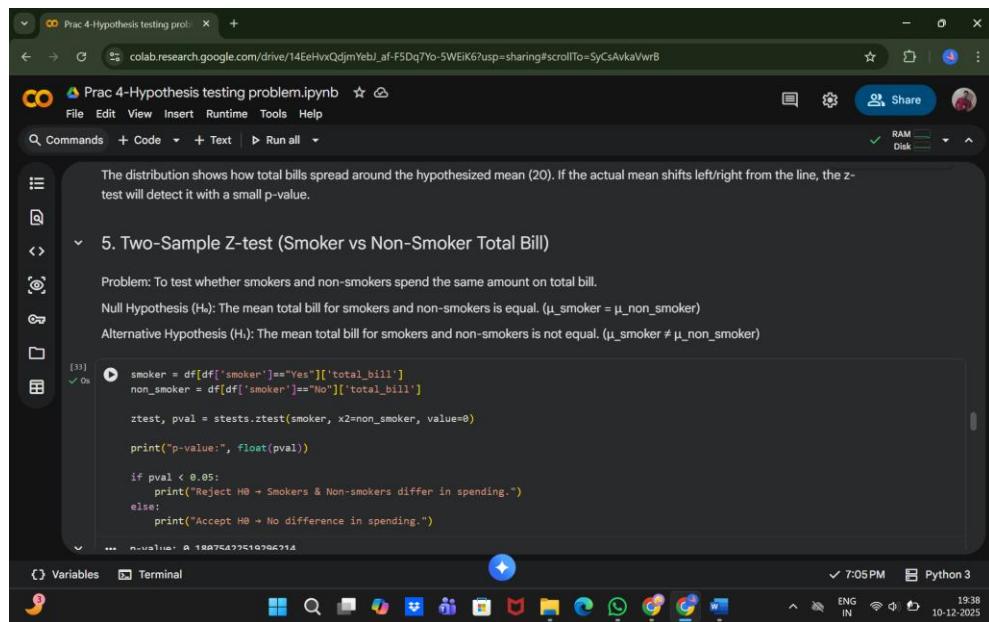
Alternative Hypothesis ( $H_1$ ): The population mean total bill is not equal to the claimed value. ( $\mu \neq 20$ )

```
[31] 0s
from statsmodels.stats.weightstats import ztest
ztest, pval = 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.")
```

Variables Terminal 7:05 PM Python 3 19:37 10-12-2025



### PRACTICAL NO. 4



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

```
[33]: smoker = df[df['smoker']=="Yes"]["total_bill"]
non_smoker = df[df['smoker']=="No"]["total_bill"]

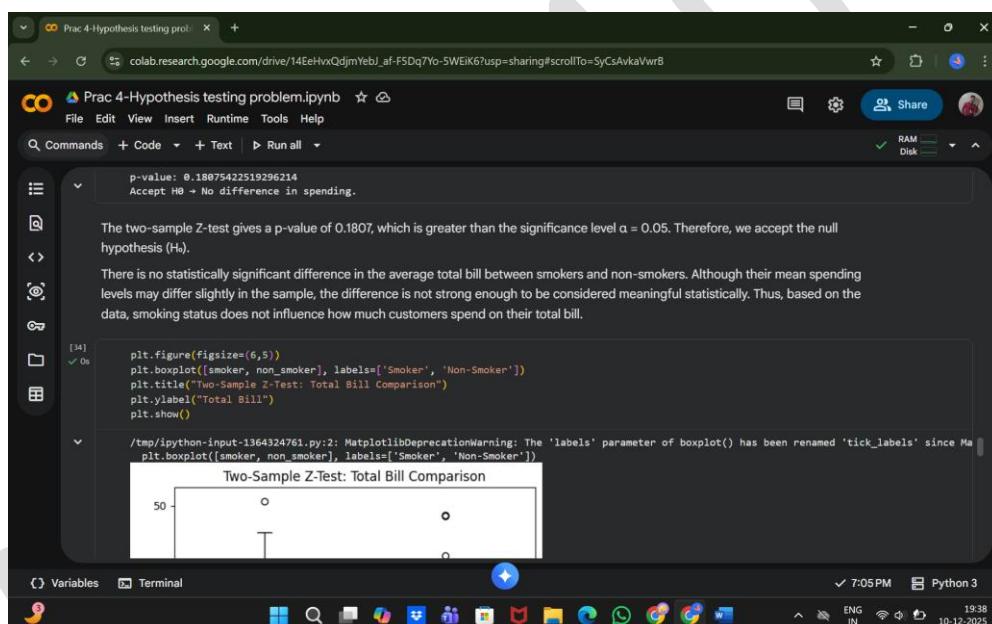
ztest, pval = stests.ztest(smoker, x2=non_smoker, value=0)

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.")

# Output: A 18A75422519296214
```

7:05 PM Python 3 ENG IN 19:38 10-12-2025



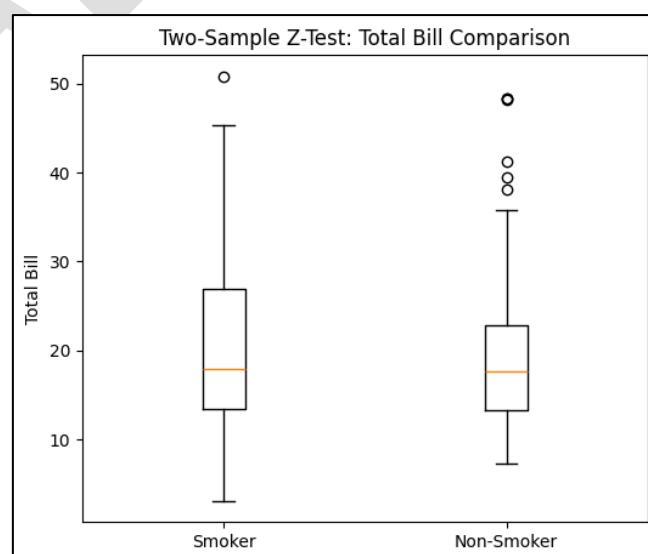
p-value: 0.18075422519296214  
Accept  $H_0$  → No difference in spending.

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]: 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()
```

7:05 PM Python 3 ENG IN 19:38 10-12-2025



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

## Data Science

### PRACTICAL NO. 4

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)
```

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

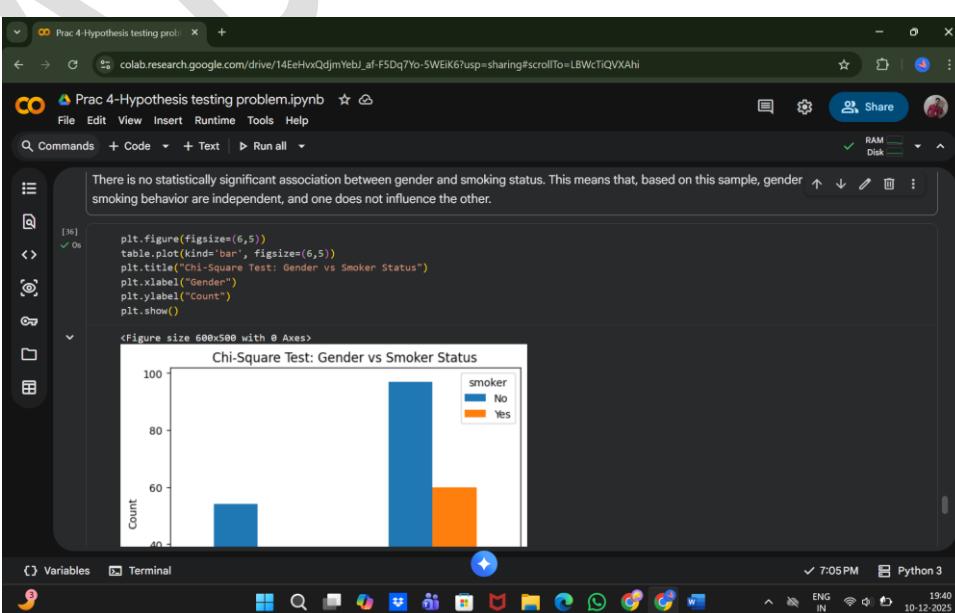
```
[15]: print("p-value:", pval)
print("Degrees of freedom:", dof)
print("Expected frequencies:\n", expected)

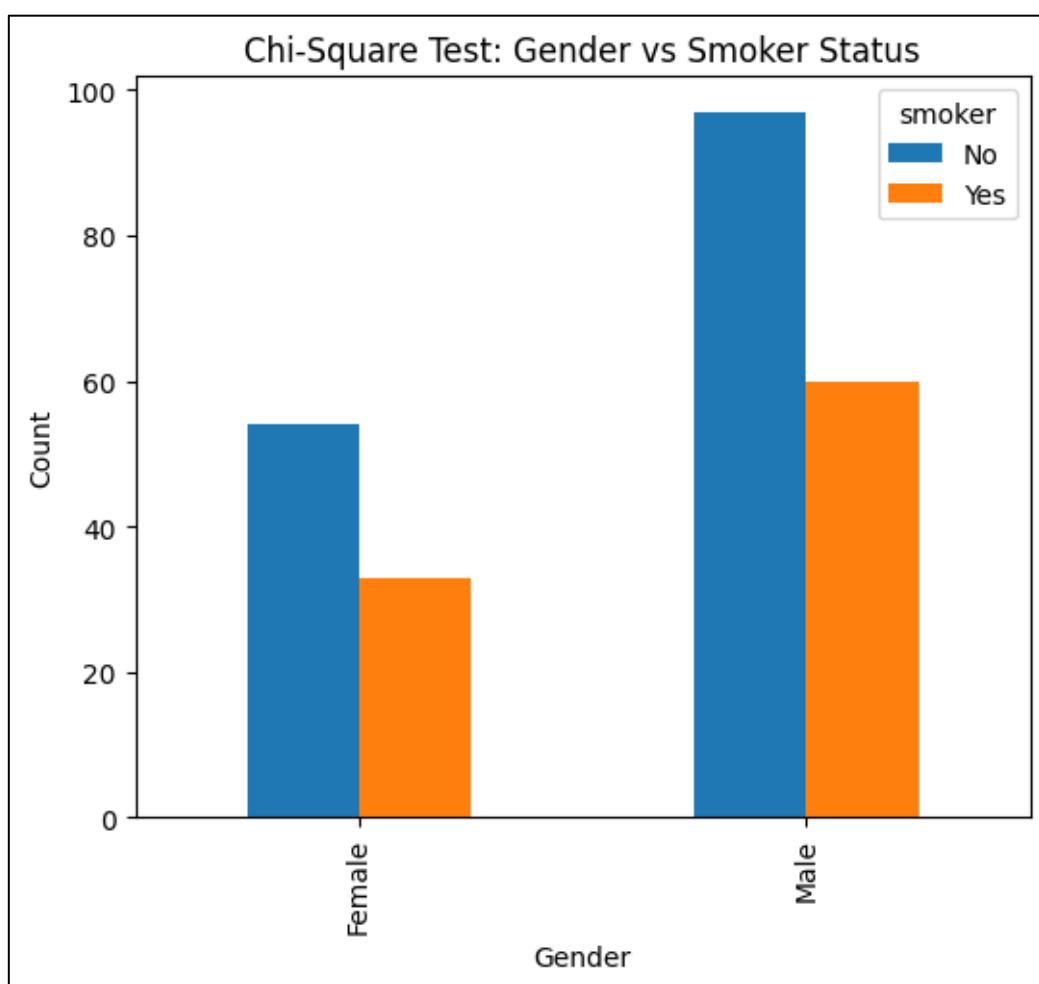
if pval < 0.05:
    print("Reject  $H_0$  → Gender and Smoking are dependent.")
else:
    print("Accept  $H_0$  → Gender and Smoking are independent.")

...
Contingency Table:
smoker No 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]
 [57.15983607 59.84016393]]
Accept  $H_0$  → Gender and Smoking are independent.
```

The Chi-square statistic is 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 ( $H_0$ ).

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





ABIN