Name: Shivam.R.Prajapati **Div: D15C** **Roll No:41**

Experiment No: 4

AIM: Implementation of **Statistical Hypothesis Test** using Scipy and Sci-kit learn.

THEORY:

1]Statistical Hypothesis Test:

A statistical hypothesis test is a **method to check if a claim about data is true using probability**. It involves two hypotheses: the **Null hypothesis (H_0)** (no effect) and the **alternative hypothesis (H_1)** (some effect). Data is collected and analyzed using statistical tests, and a **p-value determines the result**—if $p < 0.05$, H_0 is rejected, meaning there is strong evidence for H_1; otherwise, H_0 is not rejected. It helps in making decisions based on sample data.

2] Correlation Tests:

Correlation tests measure the relationship between two variables

1. Pearson's Correlation Coefficient (r) :

Pearson's correlation measures the **linear relationship** between two continuous variables. It checks if an increase in one variable leads to an increase or decrease in another.

- **r = +1** → Strong positive correlation
- **r = 0** → No correlation
- **r = -1** → Strong negative correlation

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

2. Spearman's Rank Correlation (ρ):

Spearman's correlation is used when data is not normally distributed or when variables are ordinal (ranked). It checks for a **monotonic relationship** (whether one variable increases, the other also increases but not necessarily at a constant rate).

- **ρ = +1** → Perfect positive rank correlation
- **ρ = 0** → No correlation
- **ρ = -1** → Perfect negative rank correlation

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where,

- di = Difference between the ranks of corresponding values
- n = Number of observations

3. Kendall's Rank Correlation (τ):

Kendall's correlation is used for **small datasets** and measures the **strength of association between two variables .**

- **τ = +1** → Perfect agreement
- **τ = 0** → No association
- **τ = -1** → Perfect disagreement

$$\tau = \frac{(C-D)}{\frac{1}{2}n(n-1)}$$

Where,

- C = Number of **concordant pairs** [A pair of observations (Xi,Yi) and (Xj,Yj)is concordant if both values increase or both decrease together.]
- D = Number of **discordant pairs** [A pair is discordant if one variable increases while the other decreases.]
- n = Number of observations

4. Chi-Squared Test for Independence:

The Chi-Square test is used to **check if there is a significant relationship between two categorical variables**.

- If **p-value < 0.05**, the variables are **dependent** (related).

- If **p-value ≥ 0.05**, the variables are **independent** (no relationship).

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where,

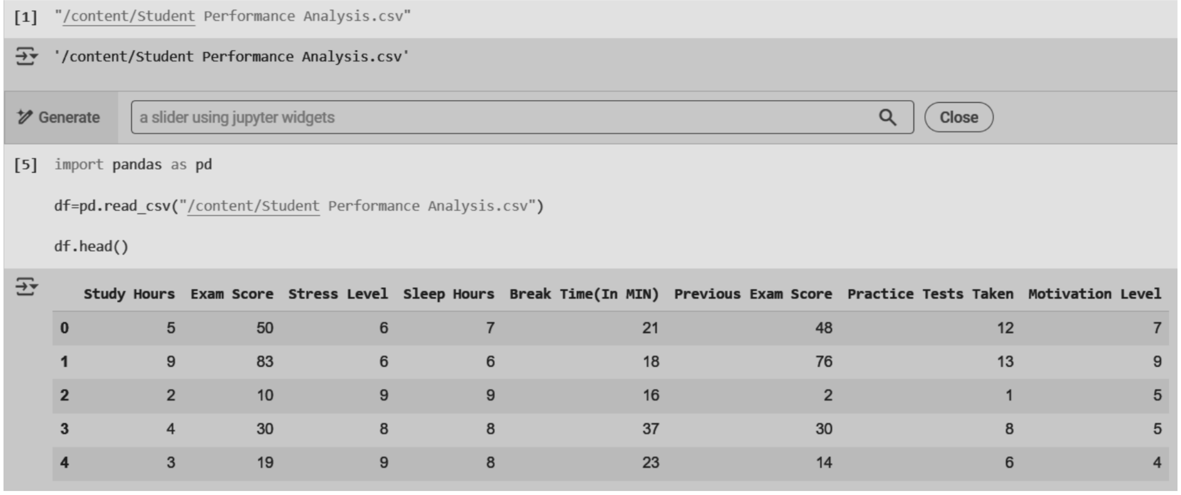
- O = Observed frequency
- E= Expected frequency

DATASET:

The **Student Performance Analysis** dataset contains **50 records with 8 attributes related to academic performance**. It includes Study Hours, Exam Score, Stress Level, Sleep Hours, Break Time (in minutes), Previous Exam Score, Practice Tests Taken, and Motivation Level. The dataset helps analyze how factors like study habits, stress, sleep, and motivation impact student exam performance.

STEPS:

Step 1: Load the Dataset into Google Colab



```
[1] "/content/Student Performance Analysis.csv"
```

```
'/content/Student Performance Analysis.csv'
```

```
[5] import pandas as pd

    df=pd.read_csv("/content/Student Performance Analysis.csv")

    df.head()
```

	Study Hours	Exam Score	Stress Level	Sleep Hours	Break Time(In MIN)	Previous Exam Score	Practice Tests Taken	Motivation Level
0	5	50	6	7	21	48	12	7
1	9	83	6	6	18	76	13	9
2	2	10	9	9	16	2	1	5
3	4	30	8	8	37	30	8	5
4	3	19	9	8	23	14	6	4

The dataset "**Student Performance Analysi**s" is loaded into a Google Colab Notebook using Pandas. It contains **8 columns**: Study Hours, Exam Score, Stress Level, Sleep Hours, Break Time (in Minutes), Previous Exam Score, Practice Tests Taken, and Motivation Level. The **data represents student study habits, stress, and motivation, helping analyze their impact on exam performance**.

Step 2: Performing **Pearson Correlation Test**

Code:

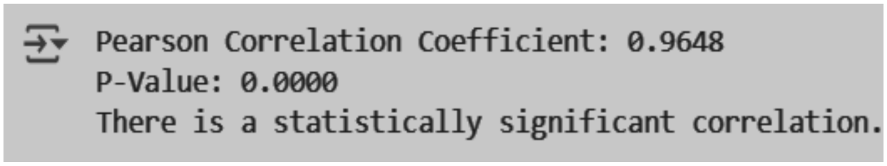
```
from scipy.stats import pearsonr

corr, p_value = pearsonr(df["Study Hours"], df["Exam Score"])

print(f"Pearson Correlation Coefficient: {corr:.4f}")
print(f"P-Value: {p_value:.4f}")

if p_value < 0.05:
   print("There is a statistically significant correlation.")
else:
   print("There is no significant correlation.")
```

Output:



```
Pearson Correlation Coefficient: 0.9648
P-Value: 0.0000
There is a statistically significant correlation.
```

The **output shows a strong positive correlation (0.9648) between Study Hours and Exam Score**, meaning that as study hours increase, exam scores also increase. The **p-value (0.0000) is less than 0.05**, indicating the correlation is statistically significant and unlikely due to chance

Step 3: Performing **Spearman's Rank Correlation Test**

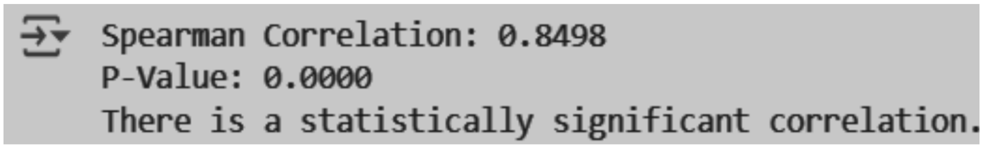
Code:

```
from scipy.stats import spearmanr
corr, p_value = spearmanr(df["Motivation Level"], df["Exam Score"])

print(f"Spearman Correlation: {corr:.4f}")
print(f"P-Value: {p_value:.4f}")

if p_value < 0.05:
    print("There is a statistically significant correlation.")
else:
    print("There is no significant correlation.")
```

Output:



```
Spearman Correlation: 0.8498
P-Value: 0.0000
There is a statistically significant correlation.
```

The **output shows a strong positive Spearman correlation (0.8498) between Motivation Level and Exam Score**, indicating that as motivation increases, exam scores tend to increase. **The p-value (0.0000) is less than 0.05**, confirming the correlation is statistically significant and not due to random chance.

Step 4: Performing **Kendall Correlation Test**

Code:

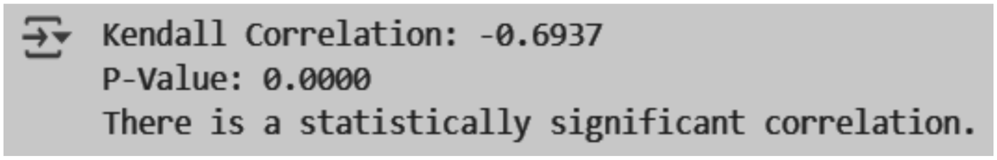
```
from scipy.stats import kendalltau

corr, p_value = kendalltau(df["Exam Score"], df["Stress Level"])

print(f"Kendall Correlation: {corr:.4f}")
print(f"P-Value: {p_value:.4f}")

if p_value < 0.05:
    print("There is a statistically significant correlation.")
else:
    print("There is no significant correlation.")
```

Output:



```
Kendall Correlation: -0.6937
P-Value: 0.0000
There is a statistically significant correlation.
```

The **output shows a strong negative Kendall correlation (-0.6937) between Exam Score and Stress Level**, meaning that as stress increases, exam scores tend to decrease. The **p-value (0.0000) is less than 0.05**, indicating that this negative correlation is statistically significant and not due to random chance

Step 5: Chi-Square Test

Code:

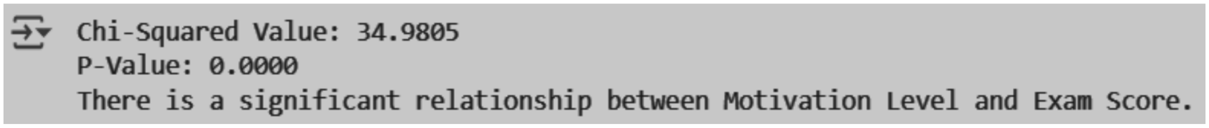
```
from scipy.stats import chi2_contingency
df["Motivation Category"] = pd.qcut(df["Motivation Level"], q=3, labels=["Low",
"Medium", "High"])
df["Score Category"] = pd.qcut(df["Exam Score"], q=3, labels=["Low",
"Medium", "High"])
contingency_table = pd.crosstab(df["Motivation Category"], df["Score
Category"])

chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print(f"Chi-Squared Value: {chi2:.4f}")
print(f"P-Value: {p_value:.4f}")

if p_value < 0.05:
    print("There is a significant relationship between Motivation Level and Exam
Score.")
else:
    print("There is no significant relationship between Motivation Level and Exam
Score.")
```

Output:



```
Chi-Squared Value: 34.9805
P-Value: 0.0000
There is a significant relationship between Motivation Level and Exam Score.
```

The output shows a Chi-Squared value of 34.9805 with a **p-value of 0.0000**, indicating a statistically significant relationship between Motivation Level and

Exam Score. This means that motivation level and exam performance are not independent and are likely related.

CONCLUSION:

In this experiment we have implemented statistical hypothesis testing using Scipy and Scikit-Learn to analyze relationships between different variables. The Pearson's coefficient correlation measured the strength of the linear relationships,while the Spearman's rank correlation and Kendall Rank Correlation assessed monotonic relationships making them useful for nonlinear data.Additionally the Chi-Square Test was used to determine if two categorical Variables are independent or not.These tests provide deeper insight into data association, helping hypothesis validation and informed decision making.