Name : Advik Ashok Hegde **Div : D15C** **Roll No : 15**

Experiment No. - 4 :

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

Problem Statement: Perform the following Tests:Correlation Tests:
a) Pearson's Correlation Coefficient
b) Spearman's Rank Correlation
c) Kendall's Rank Correlation
d) Chi-Squared Test

Introduction Statistical hypothesis testing is a fundamental concept in data analysis and machine learning. It helps in determining relationships between variables and making data-driven decisions. In this experiment, we implement various statistical hypothesis tests using Python libraries such as SciPy and Scikit-learn.

For this experiment ,we are working with the same dataset that we obtained after cleaning in the last experiment named "cleaned_vehivles.csv".

Since all data cleaning and preprocessing operations are already performed ,we can directly start with performing the operations of Statistical Hypothesis Testing.

Pearson's Correlation Coefficient

- Measures the linear relationship between two continuous variables.
- Values range from -1 to 1, where 1 indicates a strong positive correlation, -1 indicates a strong negative correlation, and 0 indicates no correlation.

```python
from scipy.stats import pearsonr

# Calculate Pearson correlation
pearson_corr, pearson_p = pearsonr(df["engine_hp"], df["msrp"])

print(f"Pearson Correlation Coefficient: {pearson_corr}")
print(f"P-value: {pearson_p}")
```

```
Pearson Correlation Coefficient: 0.6587937229804306
P-value: 0.0
```

A value of **0.6588** suggests a **moderately strong positive linear relationship** between the two variables.This means that as one variable increases, the other tends to increase as well.

The **p-value of 0.0** (or a very small value close to zero) suggests that the correlation is **highly statistically significant**.This means there is strong evidence to reject the null hypothesis (which assumes no correlation between the variables).

Spearman's Rank Correlation

- A non-parametric test that assesses the monotonic relationship between two variables.
- Useful for measuring correlations in ordinal or non-normally distributed data.

```
from scipy.stats import spearmanr

# Calculate Spearman correlation
spearman_corr, spearman_p = spearmanr(df["popularity"], df["city_mpg"])

print(f"Spearman's Rank Correlation: {spearman_corr}")
print(f"P-value: {spearman_p}")
```

```
Spearman's Rank Correlation: 0.027328076748436195
P-value: 0.003833171587034418
```

A value of **0.0273** is **very close to 0**, indicating an **extremely weak positive association** between the variables.Since **p < 0.05**, the correlation is **statistically significant**.This means that, despite being weak, the relationship is unlikely to be due to random chance.

Kendall's Rank Correlation

- Another non-parametric test that measures the strength of association between two variables.
- More robust for small datasets compared to Spearman's correlation.

```
from scipy.stats import kendalltau

# Calculate Kendall correlation
kendall_corr, kendall_p = kendalltau(df["number_of_doors"], df["highway_mpg"])

print(f"Kendall's Rank Correlation: {kendall_corr}")
print(f"P-value: {kendall_p}")
```

```
Kendall's Rank Correlation: 0.1119635631132
P-value: 6.856199111675777e-47
```

A value of **0.1119** indicates a **very weak positive association** between the variables.The **p-value is extremely small** (almost 0), meaning the correlation is **highly statistically significant**.This suggests that the observed weak correlation is **unlikely to be due to random chance**.

Chi-Squared Test

- Used to test the independence between categorical variables.
- Helps in determining whether distributions of categorical variables differ from one another.

```python
from scipy.stats import chi2_contingency

# Create a contingency table
contingency_table = pd.crosstab(df["transmission_type"], df["driven_wheels"])

# Perform Chi-Squared test
chi2_stat, chi2_p, chi2_dof, chi2_expected = chi2_contingency(contingency_table)

print(f"Chi-Squared Statistic: {chi2_stat}")
print(f"P-value: {chi2_p}")
print(f"Degrees of Freedom: {chi2_dof}")
print(f"Expected Frequencies Table:\n {chi2_expected}")
```

```
Chi-Squared Statistic: 526.7198264496208
P-value: 4.5300427647599666e-105
Degrees of Freedom: 12
Expected Frequencies Table:
 [[1.14018581e+02 6.54569412e+01 2.14896373e+02 1.58628104e+02]
 [1.63460997e+03 9.38413436e+02 3.08082902e+03 2.27414758e+03]
 [1.40203681e+01 8.04895480e+00 2.64248705e+01 1.95058067e+01]
 [5.42876898e+02 3.11660264e+02 1.02318653e+03 7.55276309e+02]
 [2.47418260e+00 1.42040379e+00 4.66321244e+00 3.44220118e+00]]
```

Chi-Squared Statistic (526.72) : The Chi-Squared statistic measures the difference between observed and expected frequencies in a contingency table.A higher value indicates a greater deviation from expected frequencies, meaning the variables are likely dependent.

P-value (4.53×10^{-105}) : Since $p < 0.05$, we reject the null hypothesis, meaning there is a significant relationship between the categorical variables.

Degrees of Freedom (12) : More degrees of freedom generally indicate a more complex relationship being tested.

There is strong statistical evidence that the two categorical variables are not independent. The difference between observed and expected values is significant, meaning there is a meaningful association between them.

Conclusion Through these statistical tests, we evaluated relationships between variables and determined their significance. Pearson's test was used for linear relationships, while Spearman and Kendall's tests were used for rank-based correlations. The Chi-Squared test helped assess categorical variable dependencies. These analyses are essential in feature selection and model evaluation in machine learning.