

## Experiment No: 4

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

### **Theory:**

#### **a) Pearson's Correlation Coefficient (r)**

Pearson's correlation measures the linear relationship between two continuous variables. It assumes that the data is normally distributed and calculates how closely the variables follow a straight-line relationship. The coefficient ranges from -1 to 1, where values closer to 1 or -1 indicate a strong relationship, while 0 means no correlation. It is sensitive to outliers and works best for data with a linear trend.

#### **b) Spearman's Rank Correlation (ρ)**

Spearman's correlation evaluates the monotonic relationship between two variables based on their rankings. It does not assume a normal distribution and works well for both linear and non-linear relationships. Since it uses ranks instead of raw values, it is less affected by outliers. A high Spearman's coefficient indicates that as one variable increases, the other tends to increase (or decrease) consistently, but not necessarily at a constant rate.

#### **c) Kendall's Rank Correlation (τ)**

Kendall's correlation measures the ordinal association between two variables. It is based on concordant and discordant pairs, where concordant means both variables increase together, and discordant means one increases while the other decreases. Kendall's Tau is particularly useful for small datasets and is more robust against tied ranks than Spearman's correlation.

#### **d) Chi-Squared Test ( $\chi^2$ )**

The Chi-Square test assesses the association between two categorical variables. It helps determine if one variable depends on another by comparing observed and expected frequencies in a contingency table. A low p-value ( $< 0.05$ ) indicates a significant relationship, meaning the two variables are not independent. This test is commonly used to analyze relationships between binned numerical data or categorical variables.

### **Correlation Analysis of AQI Dataset**

Air Quality Index (AQI) is an important measure of air pollution levels, influenced by pollutants such as  $\text{SO}_2$ ,  $\text{NO}_x$ , RSPM, and  $\text{CO}_2$ . Understanding the correlation between these pollutants and AQI can help determine which factors significantly impact air

8	2009-10-01 0:00	14	56	146			131 MPCB-KR
9	2009-11-01 0:00	14	47	136			124 MPCB-KR
10	2009-12-01 0:00	13	36	115			110 MPCB-KR
11	13-01-2009	19	69	164			143 MPCB-KR
12	14-01-2009	25	67	164			143 MPCB-KR
13	15-01-2009	23	65	182			155 MPCB-KR
14	16-01-2009	23	68	159			139 MPCB-KR
15	17-01-2009	16	41	161			141 MPCB-KR
16	18-01-2009	16	40	168			145 MPCB-KR
17	19-01-2009	22	68	190			160 MPCB-KR
18	20-01-2009	20	61	194			163 MPCB-KR
19	21-01-2009	20	61	191			161 MPCB-KR
20	22-01-2009	21	67	179			153 MPCB-KR
21	23-01-2009	21	71	195			163 MPCB-KR
22	24-01-2009	23	65	212			175 MPCB-KR
23	25-01-2009	25	35	168			145 MPCB-KR
24	26-01-2009	26	27	167			145 MPCB-KR
25	27-01-2009	24	40	162			141 MPCB-KR
26	29-01-2009	BDL - NA	30	174			149 MPCB-KR
27	30-01-2009	BDL - NA	30	200			167 MPCB-KR
28	2009-01-02 0:00	BDL - NA	16	140			127 MPCB-KR
29	2009-02-02 0:00	BDL - NA	20	133			122 MPCB-KR
30	2009-03-02 0:00	BDL - NA	47	156			143 MPCB-KR

## Steps:

### Load and Preprocess the Data

from google.colab import files

```
uploaded = files.upload()
```

```
import pandas as pd
```

```
df = pd.read_csv('PNQ_AQI.csv')
```

```
# Convert SO2 and AQI to numeric
```

```
df['SO2'] = pd.to_numeric(df['SO2'], errors='coerce')
```

```
df['AQI'] = pd.to_numeric(df['AQI'], errors='coerce')
```

```
# Drop NaN values
```

```
df_clean = df[['SO2', 'AQI']].dropna()
```


```
df.head()
```

	Date	SO2	Nox $\mu\text{g}/\text{m}^3$	RSPM $\mu\text{g}/\text{m}^3$	SPM	CO2 $\mu\text{g}/\text{m}^3$	AQI	Location
0	2009-01-01 0:00:00	15.0	53	179.0	NaN	NaN	153.0	MPCB-KR
1	2009-02-01 0:00:00	15.0	48	156.0	NaN	NaN	137.0	MPCB-KR
2	2009-03-01 0:00:00	13.0	51	164.0	NaN	NaN	143.0	MPCB-KR
3	2009-04-01 0:00:00	8.0	37	135.0	NaN	NaN	123.0	MPCB-KR
4	2009-07-01 0:00:00	13.0	36	140.0	NaN	NaN	127.0	MPCB-KR

## 1. Pearson's Correlation

```
from scipy.stats import pearsonr
```

```
pearson_corr, pearson_p = pearsonr(df_clean['SO2'], df_clean['AQI'])  
print(f"Pearson Correlation: {pearson_corr:.4f}, P-value: {pearson_p:.4f}")
```

 Pearson Correlation: 0.1868, P-value: 0.0000

### Interpretation

- Weak positive linear relationship between SO<sub>2</sub> and AQI.
- Since p-value < 0.05, the correlation is statistically significant.

## 2. Spearman's Rank Correlation

```
from scipy.stats import spearmanr
```

```
spearman_corr, spearman_p = spearmanr(df_clean['SO2'], df_clean['AQI'])  
print(f"Spearman Correlation: {spearman_corr:.4f}, P-value: {spearman_p:.4f}")
```

 Spearman Correlation: 0.1979, P-value: 0.0000

### Interpretation

- Weak positive monotonic relationship (not necessarily linear).
- p-value < 0.05, so the correlation is significant.

## 3. Kendall's Rank Correlation

```
from scipy.stats import kendalltau
```

```
kendall_corr, kendall_p = kendalltau(df_clean['SO2'], df_clean['AQI'])  
print(f"Kendall Correlation: {kendall_corr:.4f}, P-value: {kendall_p:.4f}")
```

 Kendall Correlation: 0.1337, P-value: 0.0000

### Interpretation

- Weak positive ordinal association between variables.
- p-value < 0.05, meaning the correlation is statistically significant.

#### 4. Chi-Squared Test

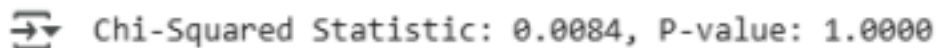
```
import numpy as np
from scipy.stats import chi2_contingency

# Categorizing SO2 and AQI into Low, Medium, High
df_clean['SO2_category'] = pd.cut(df_clean['SO2'], bins=3, labels=['Low',
'Medium', 'High'])
df_clean['AQI_category'] = pd.cut(df_clean['AQI'], bins=3, labels=['Good',
'Moderate', 'Unhealthy'])

# Create contingency table
contingency_table = pd.crosstab(df_clean['SO2_category'],
df_clean['AQI_category'])

# Perform Chi-Square Test
chi2_stat, chi2_p, _, _ = chi2_contingency(contingency_table)

print(f"Chi-Squared Statistic: {chi2_stat:.4f}, P-value: {chi2_p:.4f}")
```

A terminal window showing the output of the Chi-Squared test. The text "Chi-Squared Statistic: 0.0084, P-value: 1.0000" is displayed in a monospaced font. To the left of the text is a small icon of a terminal window with a cursor.

#### Interpretation

- No significant association between the categorical variables tested.
- p-value = 1.0, meaning the variables are independent (no relationship).

#### Conclusion:

In this analysis, four statistical tests were applied to assess the relationships between SO<sub>2</sub> levels and AQI:

1. Pearson's correlation showed a weak positive linear relationship between SO<sub>2</sub> and AQI, with a correlation of 0.1868. Since the p-value is 0.0000, this relationship is statistically significant.
2. Spearman's rank correlation confirmed a weak positive monotonic relationship between SO<sub>2</sub> and AQI, with a correlation of 0.1979. The p-value of 0.0000

indicates statistical significance.

3. Kendall's Tau also revealed a weak positive association between SO<sub>2</sub> and AQI, with a correlation of 0.1337. The p-value of 0.0000 suggests that this correlation is significant.
4. The Chi-Squared test showed no significant association between the categorical variables tested, with a Chi-Squared statistic of 0.0084 and a p-value of 1.0000. Since the p-value is much greater than 0.05, we fail to reject the null hypothesis, meaning the variables are independent.

Although the correlation tests indicate a weak but statistically significant positive relationship between SO<sub>2</sub> and AQI, the Chi-Square test suggests no significant dependency between the categorical variables analyzed.