Name: Mohit Kerkar **Div: D15C** **Roll No: 23**

<u>AIDS Exp 04</u>

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

Theory:

Correlation Analysis of AQI Dataset

1. Introduction

Air Quality Index (AQI) is an important measure of air pollution levels, influenced by pollutants such as SO_2, NOx, RSPM, and CO_2. Understanding the correlation between these pollutants and AQI can help determine which factors significantly impact air quality.
This experiment aims to perform Pearson's, Spearman's, Kendall's correlation, and the Chi-Squared test to analyze the relationship between SO_2 levels and AQI using statistical methods.
The following image is the image of my first few instances of my AQI dataset

	A	B	C	D	E	F	G	H
1	Date	SO2 µg/m3	Nox µg/m3	RSPM µg/m3	SPM	CO2 µg/m3	AQI	Location
2	2009-01-01 0:00	15	53	179			153	MPCB-KR
3	2009-02-01 0:00	15	48	156			137	MPCB-KR
4	2009-03-01 0:00	13	51	164			143	MPCB-KR
5	2009-04-01 0:00	8	37	135			123	MPCB-KR
6	2009-07-01 0:00	13	36	140			127	MPCB-KR
7	2009-08-01 0:00	10	30	135			123	MPCB-KR
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

PNQ_AQI.csv

2. Theoretical Background

2.1 Pearson's Correlation Coefficient (r)

Pearson's correlation measures the linear relationship between two continuous variables. It ranges from -1 to 1:

Formula:

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

Where:

- X_i, Y_i are individual data points
- $\bar{X}, \bar{Y}$ are the means of X and Y
- $\sum$ represents summation

- r > 0 → Positive correlation
- r < 0 → Negative correlation
- r = 0 → No linear correlation

Significance:

- Determines the strength and direction of the relationship.
- Requires normally distributed data.

2.2 Spearman's Rank Correlation (ρ)

Spearman's correlation measures the monotonic relationship between two variables based on their ranked values.

- Works for both linear and nonlinear relationships.
- Less sensitive to outliers.

Significance:

- Useful when data is not normally distributed.
- Helps identify whether an increase in one variable generally corresponds to an increase in another.

Formula:

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

Where:

- d_i = difference between ranks of corresponding values
- n = number of data points

Interpretation:

- $\rho = 1$ → Perfect positive monotonic relationship
- $\rho = -1$ → Perfect negative monotonic relationship
- $\rho \approx 0$ → No monotonic relationship

2.3 Kendall's Rank Correlation (τ)

Kendall's Tau is similar to Spearman's correlation but focuses on the ordinal association between two variables. It compares the number of concordant and discordant pairs.

Significance:

- Measures how well the ranks of one variable correspond to the ranks of another.
- More robust for small datasets.

Formula:

$$\tau = \frac{C - D}{C + D}$$

Where:

- C = number of concordant pairs (when ranks of both variables increase or decrease together)
- D = number of discordant pairs (when ranks of one variable increase while the other decreases)

Interpretation:

- $\tau > 0$ → Positive association
- $\tau < 0$ → Negative association
- $\tau = 0$ → No association

2.4 Chi-Squared Test (χ^2)

The Chi-Squared test evaluates whether there is a statistically significant association between two categorical variables.
Significance:

- Helps determine whether AQI levels are dependent on SO_2 concentrations.
- Works for categorical (binned) data.

Formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- O_i = Observed frequency
- E_i = Expected frequency under independence assumption

Interpretation:

- If p-value $< \alpha$ (e.g., 0.05), reject the null hypothesis → Variables are dependent
- If p-value $> \alpha$, fail to reject the null hypothesis → No significant relationship

3. Experimental Methodology

Load and Preprocess the Data

```
import pandas as pd
import numpy as np
from scipy.stats import pearsonr, spearmanr, kendalltau, chi2_contingency

# Load dataset
df = pd.read_csv('/content/sample_data/PNQ_AQI.csv', encoding='utf-8')

# Convert relevant columns to numeric
df['SO2 µg/m3'] = pd.to_numeric(df['SO2 µg/m3'], errors='coerce')
df['AQI'] = pd.to_numeric(df['AQI'], errors='coerce')

# Drop NaN values
df = df.dropna()
```

```
                  Date  SO2 µg/m3  Nox µg/m3  RSPM µg/m3  SPM  CO2 µg/m3  \
0  2009-01-01 00:00:00       15.0       53.0       179.0  NaN        NaN
1  2009-02-01 00:00:00       15.0       48.0       156.0  NaN        NaN
2  2009-03-01 00:00:00       13.0       51.0       164.0  NaN        NaN
3  2009-04-01 00:00:00        8.0       37.0       135.0  NaN        NaN
4  2009-07-01 00:00:00       13.0       36.0       140.0  NaN        NaN

     AQI Location             AQI_category SO2_category
0  153.0  MPCB-KR                Unhealthy          Low
1  137.0  MPCB-KR  Unhealthy for Sensitive          Low
2  143.0  MPCB-KR  Unhealthy for Sensitive          Low
3  123.0  MPCB-KR  Unhealthy for Sensitive     Very Low
4  127.0  MPCB-KR  Unhealthy for Sensitive          Low
```

Pearson's Correlation

pearson_corr, pearson_p = pearsonr(df['SO2 µg/m3'], df['AQI'])
print(f"Pearson Correlation: {pearson_corr:.4f}, p-value: {pearson_p:.4f}")

```
Pearson Correlation: 0.1868, p-value: 0.0000
```

Interpretation:

- If $p < 0.05$, the correlation is statistically significant.
- The closer r is to 1 or -1, the stronger the relationship.

Spearman's Rank Correlation

spearman_corr, spearman_p = spearmanr(df['SO2 µg/m3'], df['AQI'])
print(f"Spearman Correlation: {spearman_corr:.4f}, p-value: {spearman_p:.4f}")

```
Spearman Correlation: 0.1979, p-value: 0.0000
```

Interpretation:

- A positive ρ indicates that as SO_2 increases, AQI tends to increase.
- Works well even if the relationship is nonlinear.

Kendall's Rank Correlation

kendall_corr, kendall_p = kendalltau(df['SO2 µg/m3'], df['AQI'])
print(f"Kendall Correlation: {kendall_corr:.4f}, p-value: {kendall_p:.4f}")

```
Kendall Correlation: 0.1337, p-value: 0.0000
```

Interpretation:
- Measures how well ranks match.
- More stable for small datasets.

Chi-Squared Test

Before applying Chi-Square, we categorize SO_2 and AQI into bins:

```
# Categorize SO2 and AQI into bins
df['SO2_category'] = pd.cut(df['SO2 µg/m3'], bins=3, labels=['Low', 'Medium', 'High'])
df['AQI_category'] = pd.cut(df['AQI'], bins=3, labels=['Good', 'Moderate', 'Unhealthy'])

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

# Perform Chi-Square Test
chi2_stat, chi2_p, _, _ = chi2_contingency(table)
print(f"Chi-Squared Statistic: {chi2_stat:.4f}, p-value: {chi2_p:.4f}")
```

```
Chi-Squared Statistic: 486.6191, p-value: 0.0000
```

Interpretation:
- If $p < 0.05$, AQI levels are significantly dependent on SO_2.

4. **Results & Discussion**

Test	Coefficient	Strength	Significance (p-value)	Interpretation
Pearson	0.1868	Weak	0.0000	Weak linear correlation
Spearman	0.1979	Weak	0.0000	Weak monotonic correlation
Kendall	0.1337	Very Weak	0.0000	Weak ordinal correlation
Chi-Square	486.6191	Significant	0.0000	SO_2 significantly impacts AQI

Key Findings:
- Pearson, Spearman, and Kendall correlations show a weak positive relationship between SO_2 and AQI.
- Chi-Square test confirms that AQI depends on SO_2 levels in a statistically significant way.
- SO_2 alone is not a strong predictor of AQI, so other pollutants (NOx, RSPM, etc.) likely play a major role.

Name: Mohit Kerkar **Div: D15C** **Roll No: 23**

5. Conclusion

This experiment involved manually calculating the correlation between SO_2 and AQI using different statistical tests. Through step-by-step computations, we found that while SO_2 has a weak correlation with AQI, the Chi-Square test suggested a significant relationship. However, since AQI is influenced by multiple pollutants, it became evident that SO_2 alone does not determine air quality.

By working through these calculations, we gained a deeper understanding of how different statistical methods reveal relationships between variables. Future manual analyses could focus on NOx, CO_2, and RSPM to further explore their individual effects on AQI and refine our understanding of air pollution dynamics.