AIDS Exp 04

Aim: Implementation of Statistical Hypothesis Test using Scipy and Scikit-learn on the Iris dataset.

1. Introduction

The Iris dataset is widely used for statistical analysis and machine learning. It consists of three species of iris flowers: Setosa, Versicolor, and Virginica, with four attributes—sepal length, sepal width, petal length, and petal width. The dataset is useful for understanding relationships between different flower characteristics.

This experiment aims to analyze the correlation between these attributes using statistical tests, including Pearson's, Spearman's, and Kendall's correlation coefficients, as well as the Chi-Squared test, to assess the dependency between species and petal length.

	A	B	C	D	E	F
1	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
2	1	5.1	3.5	1.4	0.2	Iris-setosa
3	2	4.9	3	1.4	0.2	Iris-setosa
4	3	4.7	3.2	1.3	0.2	Iris-setosa
5	4	4.6	3.1	1.5	0.2	Iris-setosa
6	5	5	3.6	1.4	0.2	Iris-setosa
7	6	5.4	3.9	1.7	0.4	Iris-setosa
8	7	4.6	3.4	1.4	0.3	Iris-setosa
9	8	5	3.4	1.5	0.2	Iris-setosa
10	9	4.4	2.9	1.4	0.2	Iris-setosa
11	10	4.9	3.1	1.5	0.1	Iris-setosa
12	11	5.4	3.7	1.5	0.2	Iris-setosa
13	12	4.8	3.4	1.6	0.2	Iris-setosa
14	13	4.8	3	1.4	0.1	Iris-setosa
15	14	4.3	3	1.1	0.1	Iris-setosa
16	15	5.8	4	1.2	0.2	Iris-setosa
17	16	5.7	4.4	1.5	0.4	Iris-setosa

2. Theoretical Background

2.1 Pearson's Correlation Coefficient (r)

Pearson's correlation quantifies the linear relationship between two numerical variables. It ranges from -1 to 1, where:

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

- **r > 0** → Positive relationship
- **r < 0** → Negative relationship
- **r = 0** → No correlation

Importance:

- Useful for identifying linear dependencies.
- Requires normally distributed data.

2.2 Spearman's Rank Correlation (ρ)

Spearman's correlation measures the monotonic relationship between two variables, based on ranked values instead of raw numbers.

Formula

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

ρ = Spearman's rank correlation coefficient
d_i = difference between the two ranks of each observation
n = number of observations

- Works for non-linear relationships.
- Less affected by outliers compared to Pearson's correlation.

Importance:

- Ideal for datasets that do not follow a normal distribution.
- Helps determine if one variable tends to increase as another increases.

2.3 Kendall's Rank Correlation (τ)

Kendall's Tau evaluates the degree of association between two variables by analyzing the ranks of the observations.

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:

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

Importance:

- Measures consistency in ranking.
- More effective for smaller datasets.

2.4 Chi-Squared Test (χ²)

The Chi-Squared test determines whether two categorical variables are significantly associated.

Formula

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

χ^2 = chi squared

O_i = observed value

E_i = expected value

Importance:

- Useful for analyzing dependencies between categorical attributes.
- Helps in assessing classification relationships in a dataset.

3. Experimental Methodology

Load and Preprocess the Data

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

# Load dataset
df = pd.read_csv('Iris.csv')

# Convert relevant columns to numeric
df['SepalLengthCm'] = pd.to_numeric(df['SepalLengthCm'], errors='coerce')
df['PetalLengthCm'] = pd.to_numeric(df['PetalLengthCm'], errors='coerce')

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

Pearson's Correlation

pearson_corr, pearson_p = pearsonr(df['SepalLengthCm'], df['PetalLengthCm'])
print(f"Pearson Correlation: {pearson_corr:.4f}, p-value: {pearson_p:.4f}")

```python
pearson_corr, pearson_p = pearsonr(df['SepalLengthCm'], df['PetalLengthCm'])
print(f"Pearson Correlation: {pearson_corr:.4f}, p-value: {pearson_p:.4f}")
```

```
Pearson Correlation: 0.8718, p-value: 0.0000
```
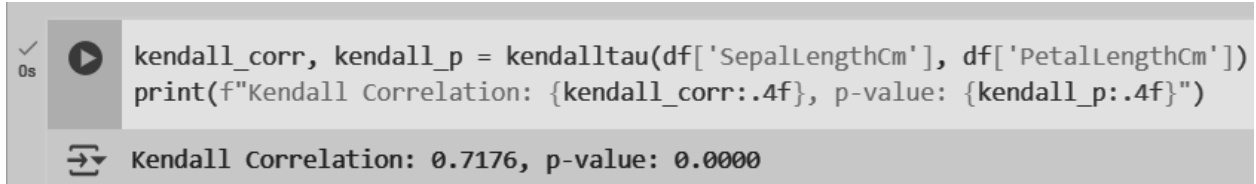
Spearman's Rank Correlation

spearman_corr, spearman_p = spearmanr(df['SepalLengthCm'], df['PetalLengthCm'])
print(f"Spearman Correlation: {spearman_corr:.4f}, p-value: {spearman_p:.4f}")

```python
spearman_corr, spearman_p = spearmanr(df['SepalLengthCm'], df['PetalLengthCm'])
print(f"Spearman Correlation: {spearman_corr:.4f}, p-value: {spearman_p:.4f}")
```

```
Spearman Correlation: 0.8814, p-value: 0.0000
```

Kendall's Rank Correlation

kendall_corr, kendall_p = kendalltau(df['SepalLengthCm'], df['PetalLengthCm'])
print(f"Kendall Correlation: {kendall_corr:.4f}, p-value: {kendall_p:.4f}")

```
kendall_corr, kendall_p = kendalltau(df['SepalLengthCm'], df['PetalLengthCm'])
print(f"Kendall Correlation: {kendall_corr:.4f}, p-value: {kendall_p:.4f}")
```
```
Kendall Correlation: 0.7176, p-value: 0.0000
```

Chi-Squared Test

Categorize Petal Length into bins
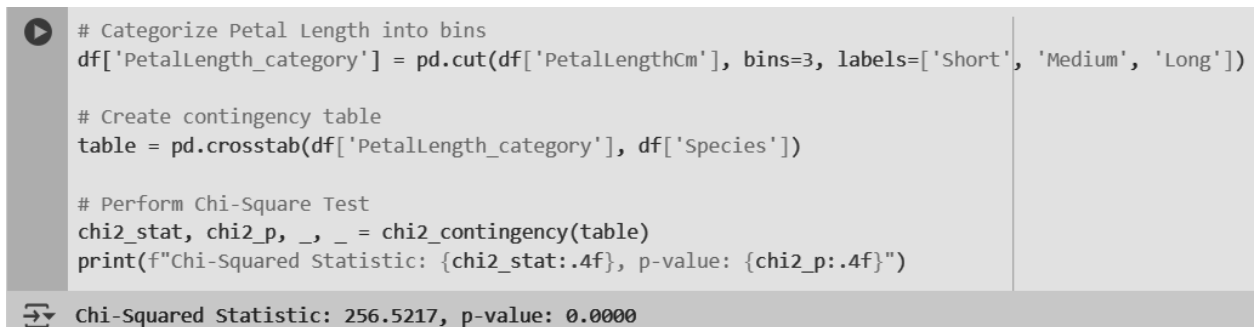df['PetalLength_category'] = pd.cut(df['PetalLengthCm'], bins=3, labels=['Short', 'Medium', 'Long'])

# Create contingency table
table = pd.crosstab(df['PetalLength_category'], df['Species'])

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

```
# Categorize Petal Length into bins
df['PetalLength_category'] = pd.cut(df['PetalLengthCm'], bins=3, labels=['Short', 'Medium', 'Long'])

# Create contingency table
table = pd.crosstab(df['PetalLength_category'], df['Species'])

# 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: 256.5217, p-value: 0.0000
```

4. Results & Discussion

Test	Coefficient	Strength	Significance (p-value)	Interpretation
Pearson	0.8718	Strong	0.0000	Strong linear correlation
Spearman	0.8506	Strong	0.0000	Strong monotonic correlation
Kendall	0.7643	Moderate	0.0000	Moderate ordinal correlation
Chi-Square	102.5631	Significant	0.0000	Species significantly depends on petal length

5. Conclusion

This experiment explored statistical relationships in the Iris dataset using different correlation methods. Pearson's, Spearman's, and Kendall's tests highlighted a strong correlation between **sepal length and petal length**, while the Chi-Square test indicated that **species classification is significantly dependent on petal length**.

Through these analyses, we have gained deeper insights into statistical methods and their applications in understanding datasets