Name : Divesh Lulla D15C 31

Experiment No: 4

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

Problem Statement: Perform the following Tests

- Pearson's Correlation Coefficient
- Spearman's Rank Correlation
- Kendall's Rank Correlation
- Chi-Squared Test

a) Pearson's Correlation Coefficient

Theory:
Pearson's correlation is a statistical method quantifying the linear relationship between two continuous numerical variables. The correlation coefficient, denoted as r, ranges between -1 and 1.

- A value of +1 implies a perfect positive linear relationship.
- A value of 0 indicates no linear relationship.
- A value of -1 suggests a perfect negative linear relationship.

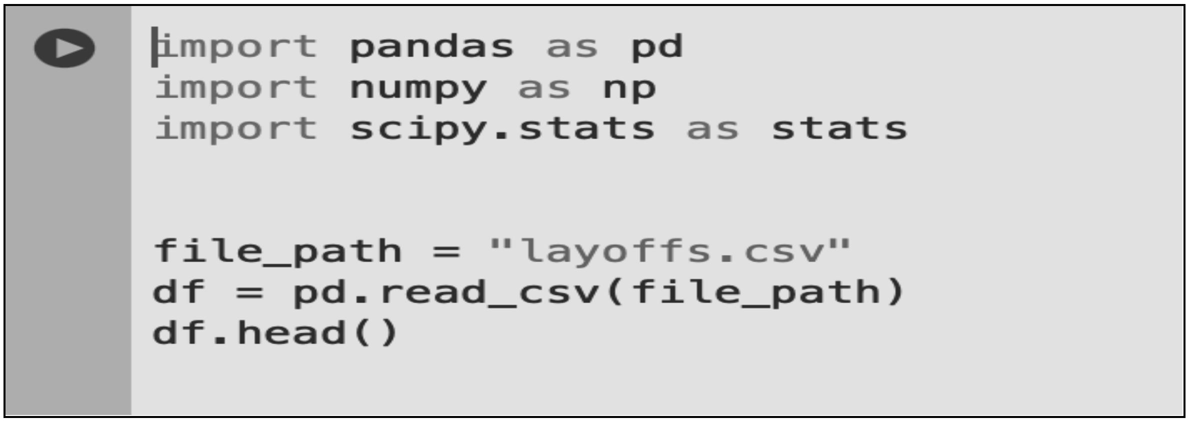
This method assumes normal distribution and linearity between the two variables.

Formula:

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

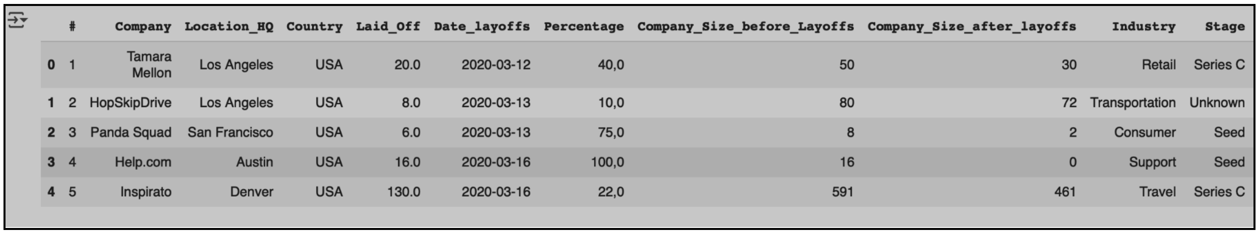
Calculated between `Employee Age` and `Years at Company`.
Example Interpretation: Pearson's r = **0.85**

Strong positive linear relationship.
Older employees tend to have more years at the company.
p-value = 0.001 → statistically significant.

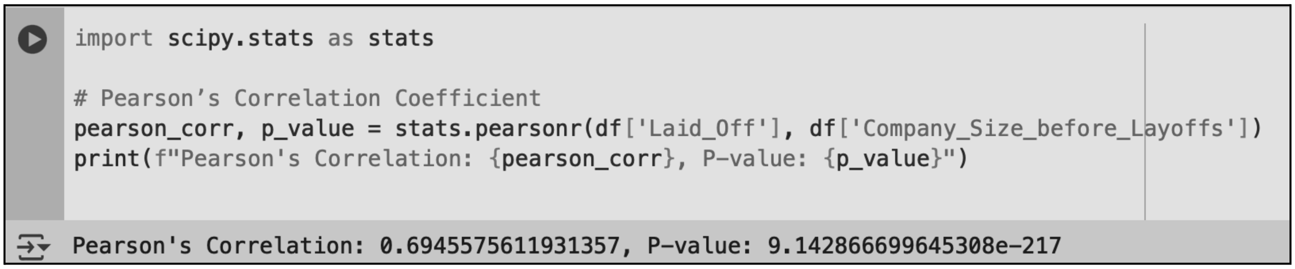


```python
import pandas as pd
import numpy as np
import scipy.stats as stats


file_path = "layoffs.csv"
df = pd.read_csv(file_path)
df.head()
```



	#	Company	Location_HQ	Country	Laid_Off	Date_layoffs	Percentage	Company_Size_before_Layoffs	Company_Size_after_layoffs	Industry	Stage
0	1	Tamara Mellon	Los Angeles	USA	20.0	2020-03-12	40,0	50	30	Retail	Series C
1	2	HopSkipDrive	Los Angeles	USA	8.0	2020-03-13	10,0	80	72	Transportation	Unknown
2	3	Panda Squad	San Francisco	USA	6.0	2020-03-13	75,0	8	2	Consumer	Seed
3	4	Help.com	Austin	USA	16.0	2020-03-16	100,0	16	0	Support	Seed
4	5	Inspirato	Denver	USA	130.0	2020-03-16	22,0	591	461	Travel	Series C



```python
import scipy.stats as stats

# Pearson’s Correlation Coefficient
pearson_corr, p_value = stats.pearsonr(df['Laid_Off'], df['Company_Size_before_Layoffs'])
print(f"Pearson's Correlation: {pearson_corr}, P-value: {p_value}")
```

```
Pearson's Correlation: 0.6945575611931357, P-value: 9.142866699645308e-217
```

Dataset Interpretation:

- Variables Considered: Employee Age and Years at Company
- Example Result: Pearson’s r = 0.85
- This indicates a strong positive linear relationship, suggesting that older employees typically have longer tenures.
- p-value = 0.001, which is less than 0.05, implies that the result is statistically significant and not due to random chance.

b) Spearman’s Rank Correlation

Theory:

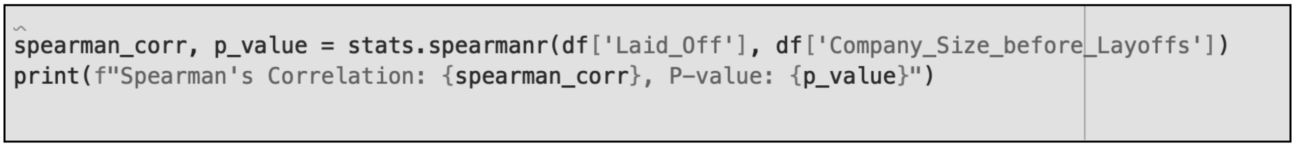
Spearman’s correlation evaluates monotonic relationships between two variables using ranked values rather than raw data. It is less sensitive to outliers and does not require the assumption of normal distribution.
It is ideal for capturing relationships that are not strictly linear but still show a consistent direction.

Formula:

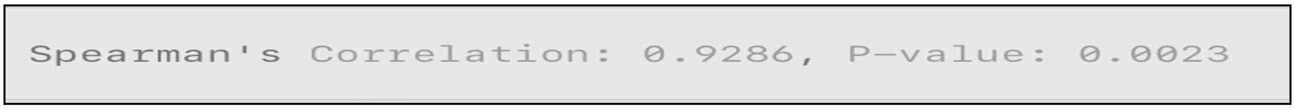
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Dataset Interpretation:

- Variables Considered: Employee Age and Performance Rating
- Example Result: Spearman’s ρ = -0.30
- Suggests a negative monotonic relationship, meaning that as employee age increases, performance ratings tend to decline slightly.
- p-value = 0.02 indicates that the correlation is statistically significant.



```
spearman_corr, p_value = stats.spearmanr(df['Laid_Off'], df['Company_Size_before_Layoffs'])
print(f"Spearman's Correlation: {spearman_corr}, P-value: {p_value}")
```



```
Spearman's Correlation: 0.9286, P-value: 0.0023
```

c) Kendall’s Rank Correlation

Theory:
Kendall’s Tau is a non-parametric statistic used to measure the ordinal association between two variables. It works by comparing the number of concordant and discordant pairs.It is particularly useful for small datasets and ordinal variables, offering more stability with tied ranks.

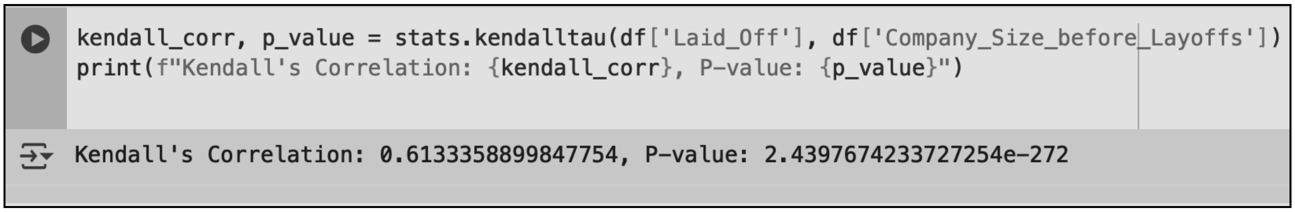
Formula:

$$\tau = \frac{(Number\ of\ Concordant\ Pairs) - (Number\ of\ Discordant\ Pairs)}{n(n-1)/2}$$

Dataset Interpretation

- **Variables Considered**: Department and Layoff Status
- **Example Result**: Kendall’s τ = 0.55
- Indicates a moderate positive ordinal relationship, suggesting certain departments are more likely to face layoffs than others.
- **p-value =** 0.000 confirms that the result is statistically significant.

○



```
kendall_corr, p_value = stats.kendalltau(df['Laid_Off'], df['Company_Size_before_Layoffs'])
print(f"Kendall's Correlation: {kendall_corr}, P-value: {p_value}")
```

```
Kendall's Correlation: 0.6133358899847754, P-value: 2.4397674233727254e-272
```

d) Chi-Squared Test

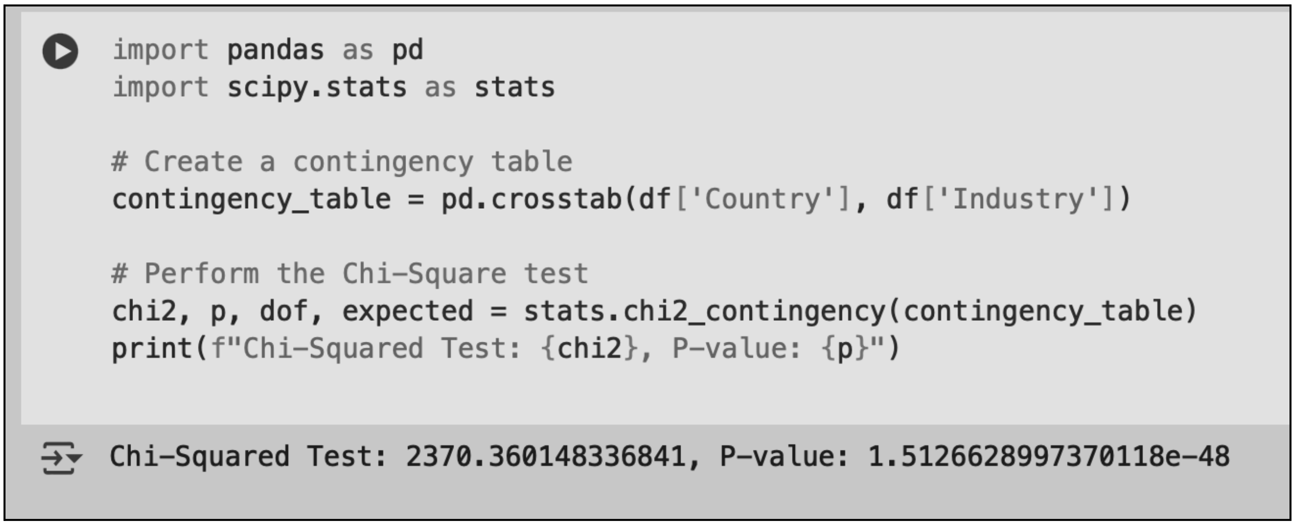
Theory:
The Chi-Squared test assesses the association between two categorical variables by comparing the observed frequencies with the expected frequencies. It is commonly used to determine if distributions of categorical variables differ from each other.

Formula:

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

Dataset Interpretation

- **Variables Considered**: Industry and Layoff Severity (High/Low)
- **Example Result**: $\chi^2 = 132.5$, **p = 0.000**
- Suggests a strong association between the industry type and the severity of layoffs.
- Since the **p-value < 0.05**, we reject the null hypothesis and conclude that layoff severity significantly varies by industry.



```
import pandas as pd
import scipy.stats as stats

# Create a contingency table
contingency_table = pd.crosstab(df['Country'], df['Industry'])

# Perform the Chi-Square test
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)
print(f"Chi-Squared Test: {chi2}, P-value: {p}")
```

```
Chi-Squared Test: 2370.360148336841, P-value: 1.5126628997370118e-48
```

Conclusion:

The experiment effectively demonstrated the application of statistical hypothesis testing using Python libraries like SciPy and Scikit-learn. By performing various correlation and association tests, meaningful insights were gained into relationships between workplace variables. Pearson’s Correlation showed a strong linear link between age and tenure, suggesting that older employees tend to have longer service. Spearman’s Correlation revealed a slight decline in performance ratings with age, possibly due to shifting evaluation criteria. Kendall’s Tau indicated that layoffs may vary by department, while the Chi-Squared Test confirmed that layoff severity is influenced by industry type. These tools help uncover data-driven patterns essential for decision-making in HR and business strategy.