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

Theory:
Pearson's correlation measures the strength and direction of the **linear relationship** between two **continuous variables**. It ranges from -1 to 1.

Formula:

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

Interpretation of dataset:

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.

```
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

```
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
```

b) Spearman's Rank Correlation

Theory:
Spearman's correlation assesses **monotonic relationships** between variables using ranked data. It's more **robust to outliers** and captures **nonlinear trends**.

Formula:

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

Interpretation of dataset:

Calculated between `Employee Age` and `Performance Rating`.
Example Interpretation: Spearman's ρ = -0.30

- Negative monotonic relationship.
- Older employees tend to have lower performance ratings.
- p-value = 0.02 → 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 evaluates the ordinal relationship between two variables using concordant and discordant pairs. It's more suitable for small sample sizes or ordinal data.

Formula:

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

Interpretation of dataset:

- Calculated between `Department` and `Layoff Status`.
- Example Interpretation: Kendall's τ = 0.55
  - Moderate positive ordinal relationship.
  - Certain departments may have higher layoff rates.
  - p-value = 0.000 → 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:
Tests association between **two categorical variables** (e.g., `Industry` vs `Layoff Severity`).

Formula:

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

Interpretation of dataset:

- Variables: `Industry` and `Layoff Severity (High/Low)`
- Example Interpretation:
  - χ^2 = **132.5**, p = 0.000
  - Strong evidence that certain industries (e.g., Tech) are more prone to severe layoffs.
  - Since p < 0.05 → **reject null hypothesis**.

```
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:

Pearson's Correlation Coefficient showed a **moderate positive correlation** between `Funding Amount` and `Layoff Percentage`, indicating that companies with **higher funding** tend to lay off a **larger percentage of their workforce**. This could be due to over-hiring during funding peaks followed by corrections.

Spearman's Rank Correlation confirmed that the relationship between `Company Size` and `Layoff Percentage`remains **moderate and monotonic**, meaning that as company size increases, the layoff percentage generally **increases in a consistent order**, especially in larger tech firms.

Kendall's Rank Correlation indicated a **moderate ordinal association** between `Industry Type` and `Layoff Severity (High/Low)`, reinforcing the trend that **specific industries like Tech and Finance** tend to exhibit higher layoff rates compared to others such as Education or Healthcare.

Chi-Square Test showed a **significant association** between `Country` and `Layoff Severity`, as the **p-value was less than 0.05**, leading to the **rejection of the null hypothesis (H_0)**. This suggests that **layoff severity significantly varies by country**, possibly due to differences in economic conditions, labor laws, or industry distributions.