Experiment 4

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

Theory:

Correlation measures the statistical relationship between two variables. It indicates how strongly and in what direction one variable changes concerning another. The result is a correlation coefficient that ranges from -1 to +1.

- Strength: How closely the data points fit a trend (linear or monotonic).
- Direction: Whether an increase in one variable results in an increase or decrease in the other.

1. Positive Correlation (r>0)

- When one variable increases, the other also increases.
- Example:
    - The relationship between study time and exam scores.

2. Negative Correlation (r<0)

- When one variable increases, the other decreases.
- Example:
    - The relationship between exercise and body fat percentage.

3. No Correlation (r=0)

- No discernible relationship between the two variables.
- Example:
    - The relationship between shoe size and intelligence.

```python
import pandas as pd
import scipy.stats as stats
from scipy.stats import chi2_contingency
```

Types of correlation tests:

1) Pearson's Correlation Coefficient

Pearson's correlation coefficient (rrr) is a statistical measure that quantifies the linear relationship between two continuous numerical variables. It ranges from −1 to +1, where +1 indicates a perfect positive correlation, −1 represents a perfect negative correlation, and 0 suggests no correlation. Pearson's correlation assumes that both variables are normally distributed and that the relationship between them is linear. It is widely used in fields such as finance, biology, and social sciences to analyze relationships like income vs. spending habits or height vs. weight.

If the absolute value of rrr is close to 1, it indicates a strong relationship, whereas values near 0 suggest a weak or no relationship. However, Pearson's correlation is sensitive to outliers, which can distort the results significantly.

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

```python
import math

col1="Vehicle Year"
col2="Vehicle Weight"

X = df_cleaned[col1].values
Y = df_cleaned[col2].values
mean_X = sum(X) / len(X)
mean_Y = sum(Y) / len(Y)
X_diff = [xi - mean_X for xi in X]
Y_diff = [yi - mean_Y for yi in Y]
numerator = sum(xi * yi for xi, yi in zip(X_diff, Y_diff))
denominator_X = sum(xi ** 2 for xi in X_diff)
denominator_Y = sum(yi ** 2 for yi in Y_diff)
correlation = numerator / math.sqrt(denominator_X * denominator_Y)
print(f"Pearson Correlation between {col1} and {col2}: {correlation:.3f}")
```

```
Pearson Correlation between Vehicle Year and Vehicle Weight: 0.314
```

```python
# Selecting numerical columns for correlation tests
numerical_cols = ["Vehicle Year", "Vehicle Weight", "Vehicle Declared Gross Weight", "Vehicle Recorded GVWR"]
df_cleaned=df[numerical_cols].copy()
df_cleaned.replace(0, pd.NA, inplace=True)  # Convert 0s to NaN if needed
df_cleaned.dropna(inplace=True)  # Drop rows with new NaNs

print(df_cleaned.head())
# Pearson Correlation
pearson_corr = df_cleaned.corr(method="pearson")
print("Pearson Correlation Matrix:\n", pearson_corr)
```

2. Spearman's Rank Correlation

Spearman's Rank Correlation (ρ) is a non-parametric test that assesses the strength and direction of a monotonic relationship between two variables. Unlike Pearson's correlation, Spearman's method does not assume a linear relationship; instead, it is based on ranked values, making it suitable for ordinal data. The coefficient ranges from
−1 to +1, where +1 indicates a perfect increasing rank agreement, −1 signifies a perfect decreasing rank agreement, and 0 means no correlation.

Spearman's correlation is ideal when dealing with non-normally distributed data or when the relationship is non-linear but still follows a consistent increasing or decreasing pattern. Common applications include ranking students based on performance or analyzing customer satisfaction scores vs. product ratings.

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

```python
import numpy as np

df_manual = df[['Vehicle Year', 'Vehicle Weight']].dropna()

df_manual['Year Rank'] = df_manual['Vehicle Year'].rank(method='average')
df_manual['Weight Rank'] = df_manual['Vehicle Weight'].rank(method='average')

df_manual['d'] = df_manual['Year Rank'] - df_manual['Weight Rank']
df_manual['d^2'] = df_manual['d'] ** 2

n = len(df_manual)
spearman_manual = 1 - (6 * df_manual['d^2'].sum()) / (n * (n**2 - 1))

print(f"Manual Spearman's Correlation: {spearman_manual:.4f}")
```

Manual Spearman's Correlation: 0.3906

```python
[23] #INBUILD
df_spearman = df[['Vehicle Year', 'Vehicle Weight']].dropna()

spearman_corr, p_value = stats.spearmanr(df_spearman['Vehicle Year'], df_spearman['Vehicle Weight'])

print(f"Spearman's Correlation: {spearman_corr:.4f}, p-value: {p_value:.4f}")
```

Spearman's Correlation: 0.0286, p-value: 0.0000

3. Kendall's Rank Correlation

Kendall's Tau (τ) is another non-parametric correlation test that measures the association between two ordinal variables. It is particularly useful for small datasets and is based on the number of concordant and discordant pairs in ranked data. A concordant pair means that the ranks of both variables move in the same direction, while a discordant pair moves in opposite directions. The correlation coefficient ranges from −1 (strong negative correlation) to +1 (strong positive correlation), with 0 indicating no association. Kendall's Tau is more robust than Spearman's correlation for small sample sizes and is commonly used in social sciences, psychology, and business research where ranking-based comparisons are necessary, such as employee performance evaluation vs. years of experience.

$$\tau = \frac{(C - D)}{\frac{1}{2}n(n - 1)}$$

```
[24] kendall_corr, p_value = kendalltau(df_cleaned["Vehicle Year"], df_cleaned["Vehicle Weight"])
     print(f"Kendall's Correlation: {kendall_corr:.4f}, p-value: {p_value:.4f}")

⤓  Kendall's Correlation: 0.2562, p-value: 0.0000
```

4. Chi-Squared Test for Categorical Data

The Chi-Squared (χ2) test is a statistical method used to determine the independence between two categorical variables. Unlike the other correlation tests, which are used for numerical or ordinal data, the Chi-Squared test assesses whether the distribution of one categorical variable is related to another. It compares the observed frequencies of data with the expected frequencies to see if any significant relationship exists. A high χ2value suggests a strong association, while a low value indicates independence. This test is widely applied in market research, healthcare, and sociology to analyze relationships such as gender vs. product preference, education level vs. voting behavior, or smoking habits vs. disease occurrence. The test assumes that each category has a sufficiently large sample size (typically, an expected frequency of at least 5 per category) to ensure reliability.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

```python
# Use correct column names from your dataset
category1 = df["Vehicle Make"]    # Manufacturer
category2 = df["Vehicle Model"]   # Model type

# Create a contingency table
contingency_table = pd.crosstab(category1, category2)

# Perform Chi-Square test
chi2_stat, p_value, dof, expected = stats.chi2_contingency(contingency_table)

# Print results
print(f"Chi-Square Statistic: {chi2_stat:.4f}")
print(f"Degrees of Freedom: {dof}")
print(f"P-value: {p_value:.4f}")
print("Expected Frequencies:\n", expected)
```

```
Chi-Square Statistic: 3468518.5933
Degrees of Freedom: 37812
P-value: 0.0000
Expected Frequencies:
 [[7.59142928e-05 2.56590310e-02 7.59142928e-05 ... 7.59142928e-05
  7.59142928e-05 7.59142928e-05]
 [1.93581447e-03 6.54305289e-01 1.93581447e-03 ... 1.93581447e-03
  1.93581447e-03 1.93581447e-03]
 [1.79157731e-02 6.05553131e+00 1.79157731e-02 ... 1.79157731e-02
  1.79157731e-02 1.79157731e-02]
 ...
 [7.40164354e-04 2.50175552e-01 7.40164354e-04 ... 7.40164354e-04
  7.40164354e-04 7.40164354e-04]
 [1.89785732e-05 6.41475774e-03 1.89785732e-05 ... 1.89785732e-05
  1.89785732e-05 1.89785732e-05]
 [1.89785732e-05 6.41475774e-03 1.89785732e-05 ... 1.89785732e-05
  1.89785732e-05 1.89785732e-05]]
```

Conclusion:

In this experiment, we analyzed relationships between variables using statistical hypothesis tests. Pearson's correlation measured linear relationships, while Spearman's and Kendall's rank correlations identified monotonic associations, making them suitable for non-linear data. The Chi-Squared test determined the independence of categorical variables. Our results highlighted the importance of choosing the right correlation method based on data type and distribution. Understanding these statistical tests is essential for accurate data analysis in various fields, including machine learning, business intelligence, and scientific research.