Experiment 4

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

Theory:

1. Pearson's Correlation Coefficient

- Measures linear correlation between two continuous variables.
- Values range from -1 to 1:
  - +1: Perfect positive correlation
  - -1: Perfect negative correlation
  - 0: No correlation

Formula:

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

2. Spearman's Rank Correlation

- Measures monotonic relationships (increasing or decreasing trends).
- Useful for non-linear relationships.
- Based on ranked data rather than raw values.
- No assumption about normality.

Formula:

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

where d_i is the difference between ranks of X and Y.

3. Kendall's Rank Correlation

- Measures the ordinal association between two variables.
- Useful for small datasets or ordinal data.
- Measures the consistency of the rank ordering.

Formula:

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

where C is the number of concordant pairs, and D is the number of discordant pairs.

4. Chi-Squared Test (χ^2 Test)

- Measures association between two categorical variables.
- Compares observed and expected frequencies in a contingency table.
- Null hypothesis states that there is no association between the variables.
- If the p-value is small (<0.05), we reject the null hypothesis (there is a significant association).

Formula:

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

where O is the observed frequency, and E is the expected frequency.

Steps:

1. **Pearson's Correlation Coefficient**

```
[23] from scipy.stats import pearsonr

     x = df["Item_MRP"].values
     y = df["Item_Outlet_Sales"].values
     n = len(x)

     mean_x = np.mean(x)
     mean_y = np.mean(y)

     numerator = sum((x - mean_x) * (y - mean_y))
     denominator = np.sqrt(sum((x - mean_x) ** 2) * sum((y - mean_y) ** 2))

     pearson_corr = numerator / denominator

     # Calculate p-value using scipy
     pearson_corr_scipy, p_value = pearsonr(x, y)

     print(f"Pearson Correlation (Manual): {pearson_corr:.4f}")
     print(f"Pearson Correlation (Scipy): {pearson_corr_scipy:.4f}")
     print(f"P-value: {p_value}")
```

```
Pearson Correlation (Manual): 0.5676
Pearson Correlation (Scipy): 0.5676
P-value: 0.0
```

- The Pearson correlation coefficient between Item_MRP (Maximum Retail Price) and Item_Outlet_Sales is 0.5676.
- This indicates a moderate positive linear relationship between MRP and sales.
- As the MRP increases, sales tend to increase as well, but the relationship is not perfectly strong.
- A higher MRP is somewhat associated with higher sales, but other factors might also be influencing sales.
- The p-value is 0.0, meaning the correlation is highly significant and unlikely due to random chance.

2. **Spearman's Rank Correlation**

```
[25] from scipy.stats import spearmanr

x = df["Item_MRP"].values
y = df["Item_Outlet_Sales"].values
n = len(x)

x_ranks = np.argsort(np.argsort(x))
y_ranks = np.argsort(np.argsort(y))

d_squared_sum = sum((x_ranks - y_ranks) ** 2)
spearman_corr_manual = 1 - (6 * d_squared_sum) / (n * (n**2 - 1))

#Compute Spearman's Correlation & p-value using SciPy
spearman_corr_scipy, p_value = spearmanr(x, y)

print(f"Spearman Correlation (Manual): {spearman_corr_manual:.4f}")
print(f"Spearman Correlation (SciPy): {spearman_corr_scipy:.4f}")
print(f"P-value: {p_value}")
```

```
Spearman Correlation (Manual): 0.5630
Spearman Correlation (SciPy): 0.5630
P-value: 0.0
```

- The Spearman correlation (0.5630) is very close to the Pearson correlation (0.5676).
- This suggests that the relationship is nearly linear, meaning that higher MRP values are generally associated with higher sales, and the ranking order remains consistent.
- If the Spearman correlation had been significantly different from Pearson's, it would indicate a non-linear relationship.
- The p-value is 0.0, indicating a statistically significant relationship, meaning the correlation is unlikely due to chance.

3. Kendall's Rank Correlation

```
from scipy.stats import kendalltau

# Define ordinal mapping for Outlet_Size
size_mapping = {"Small": 1, "Medium": 2, "High": 3}
df["Outlet_Size_Ordinal"] = df["Outlet_Size"].map(size_mapping)

# Extract necessary columns
x = df["Outlet_Size_Ordinal"].values  # Now numerical
y = df["Item_Outlet_Sales_Capped"].values
n = len(x)

C = 0  # Concordant pairs
D = 0  # Discordant pairs

# Manual Kendall's Tau calculation
for i in range(n):
    for j in range(i + 1, n):
        if (x[i] - x[j]) * (y[i] - y[j]) > 0:
            C += 1
        elif (x[i] - x[j]) * (y[i] - y[j]) < 0:
            D += 1

kendall_tau_manual = (C - D) / (0.5 * n * (n - 1))

# Compute Kendall's Tau using SciPy
kendall_tau_scipy, p_value = kendalltau(x, y)

print(f"Kendall's Rank Correlation (Manual): {kendall_tau_manual:.4f}")
print(f"Kendall's Rank Correlation (SciPy): {kendall_tau_scipy:.4f}")
print(f"P-value: {p_value:.4f}")
```

```
Kendall's Rank Correlation (Manual): 0.1226
Kendall's Rank Correlation (SciPy): 0.1633
P-value: 9.675939169456887e-82
```

- Kendall's Tau (0.1633) indicates a weak positive ordinal association between Outlet_Size and Item_Outlet_Sales_Capped.
- Compared to Pearson and Spearman correlations, Kendall's Tau is lower, suggesting a less consistent ranking relationship.
- Despite the weak correlation, the result is statistically significant, meaning the relationship is unlikely due to random chance.
- While larger outlet sizes may have slightly higher sales, the effect is weak, implying other factors like location, type, or promotions are more influential.

4. Chi-Square Test:

```python
import numpy as np
import pandas as pd
from scipy.stats import chi2, chi2_contingency

# Create contingency table
contingency_table = pd.crosstab(df["Outlet_Size"], df["Sales_Bin"])
observed = contingency_table.values

row_totals = observed.sum(axis=1).reshape(-1, 1)
col_totals = observed.sum(axis=0)
grand_total = observed.sum()
expected = (row_totals @ col_totals.reshape(1, -1)) / grand_total

chi_squared_manual = np.sum((observed - expected) ** 2 / expected)

# Degrees of Freedom
df_degrees = (observed.shape[0] - 1) * (observed.shape[1] - 1)

# Compute p-value manually
p_value_manual = 1 - chi2.cdf(chi_squared_manual, df_degrees)

# Compute using SciPy
chi2_scipy, p_value_scipy, df_scipy, expected_scipy = chi2_contingency(contingency_table)

# Print results
print(f"Chi-Square Statistic (Manual): {chi_squared_manual:.4f}")
print(f"Chi-Square Statistic (SciPy): {chi2_scipy:.4f}")
print(f"P-value (Manual): {p_value_manual:.4f}")
print(f"P-value (SciPy): {p_value_scipy:.4f}")
print(f"Degrees of Freedom: {df_degrees}")
print("\nExpected Frequencies:")
print(expected)
```

```python
# Hypothesis Decision
alpha = 0.05  # Significance level
if p_value_scipy < alpha:
    print("\nConclusion: Null hypothesis (H₀) is REJECTED.")
    print("There is a significant relationship between Outlet Size and Sales.")
else:
    print("\nConclusion: Null hypothesis (H₀) is NOT rejected.")
    print("There is no significant relationship between Outlet Size and Sales.")
```

```
Chi-Square Statistic (Manual): 356.9009
Chi-Square Statistic (SciPy): 356.9009
P-value (Manual): 0.0000
P-value (SciPy): 0.0000
Degrees of Freedom: 4

Expected Frequencies:
[[ 311.10407134  310.33861316  310.5573155 ]
 [ 932.31080605  930.01689546  930.67229849]
 [1601.58512261 1597.64449138 1598.77038601]]

Conclusion: Null hypothesis (H₀) is REJECTED.
There is a significant relationship between Outlet Size and Sales.
```

- The p-value (0.0000) is less than 0.05, meaning we reject the null hypothesis and conclude that Outlet Size and Sales are related.
- The chi-square statistic (356.9009) indicates a strong deviation from expected values, suggesting a considerable association between Outlet Size and Sales.
- Given the degrees of freedom, the result is statistically valid for the given contingency table structure.
- While the test confirms a relationship, it does not indicate how strong or in which direction the relationship is.

Conclusion:

The experiment analyzed relationships between different variables using statistical hypothesis tests. Pearson's Correlation (0.5676) and Spearman's Rank Correlation (0.5630) showed a moderate positive relationship between Item MRP and Sales, indicating that higher MRP tends to be associated with higher sales. Kendall's Rank Correlation (0.1633) suggested a weak but statistically significant positive association between Outlet Size and Sales, implying that while larger outlets may have slightly higher sales, other factors likely play a more significant role. The Chi-Square Test (p-value = 0.0000) confirmed a strong dependency between Outlet Size and Sales, but it does not indicate the strength or direction of the relationship. Overall, while MRP strongly influences sales, Outlet Size also has an impact, though it may not be the primary determining factor.