Chi Square

Problem

You have the following contingency table showing the preferences of 100 people for two types of beverages (Tea and Coffee) based on their age group (Under 30 and 30 and Above):

	Tea	**Coffee**	**Total**
Under 30	20	30	50
30 and Above	10	40	50
Total	30	70	100

Write a Python program to determine if there is a significant association between age group and beverage preference using a chi-square test.

Solution

Here's the Python code to solve this problem:

```
import numpy as np
from scipy.stats import chi2_contingency

# Contingency table
data = np.array([[20, 30],
        [10, 40]])

# Perform chi-square test
chi2, p, dof, expected = chi2_contingency(data)

# Print the results
print(f"Chi-square Statistic: {chi2}")
print(f"P-value: {p}")
print(f"Degrees of Freedom: {dof}")
print("Expected Frequencies:")
print(expected)
```

```
# Conclusion
alpha = 0.05
if p < alpha:
   print("There is a significant association between age group and beverage preference.")
else:
   print("There is no significant association between age group and beverage preference.")
```

Explanation

1. **Import Libraries**: We import `numpy` for handling the data and `chi2_contingency` from `scipy.stats` for performing the chi-square test.
2. **Create Contingency Table**: The `data` variable contains the observed frequencies.
3. **Perform Chi-Square Test**: The `chi2_contingency` function returns the chi-square statistic, p-value, degrees of freedom, and expected frequencies.
4. **Print Results**: We print the chi-square statistic, p-value, degrees of freedom, and expected frequencies.
5. **Conclusion**: Based on the p-value, we determine if there is a significant association between age group and beverage preference.

Scenario

Imagine you want to know if there is an association between a person's favorite type of pet (Dog or Cat) and their living situation (Urban or Rural). You survey 200 people, and the results are tabulated as follows:

	Dog	Cat	Total
Urban	70	30	100
Rural	50	50	100
Total	120	80	200

Steps to Perform the Chi-Square Test

1. **State the Hypotheses:**
    - Null Hypothesis (H_0): There is no association between living situation and pet preference.
    - Alternative Hypothesis (H_1): There is an association between living situation and pet preference.

2. **Calculate the Expected Frequencies:**

   Use the formula:

$$E_{ij} = \frac{(Row\ Total_i \times Column\ Total_j)}{Grand\ Total}$$

- For Urban-Dog (E_{11}):

$$E_{11} = \frac{(100 \times 120)}{200} = 60$$

- For Urban-Cat (E_{12}):

$$E_{12} = \frac{(100 \times 80)}{200} = 40$$

- For Rural-Dog (E_{21}):

$$E_{21} = \frac{(100 \times 120)}{200} = 60$$

- For Rural-Cat (E_{22}):

$$E_{22} = \frac{(100 \times 80)}{200} = 40$$

The expected frequency table is:

	Dog (Expected)	Cat (Expected)	Total
Urban	60	40	100
Rural	60	40	100
Total	120	80	200

3. **Compute the Chi-Square Statistic:**
   Use the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

   Substituting the observed and expected values:

$$\chi^2 = \frac{(70-60)^2}{60} + \frac{(30-40)^2}{40} + \frac{(50-60)^2}{60} + \frac{(50-40)^2}{40}$$

$$\chi^2 = \frac{100}{60} + \frac{100}{40} + \frac{100}{60} + \frac{100}{40}$$

$$\chi^2 = \frac{5}{3} + \frac{5}{2} + \frac{5}{3} + \frac{5}{2}$$

$$\chi^2 \approx 1.67 + 2.5 + 1.67 + 2.5 = 8.34$$

4. **Determine the Degrees of Freedom:**

$$df = (r-1) \times (c-1) = (2-1) \times (2-1) = 1$$

5. **Compare with the Critical Value:**
   For $df = 1$ and a significance level $\alpha = 0.05$, the critical value from the chi-square distribution table is 3.841.

6. **Decision:**
   Since $\chi^2 = 8.34$ is greater than the critical value of 3.841, we reject the null hypothesis.

Conclusion

There is sufficient evidence to conclude that there is a significant association between living situation and pet preference.

Pearson Correlation Coefficient

Suppose we have two variables, X and Y, with the following data points:

- $X = [1, 2, 3, 4, 5]$
- $Y = [2, 4, 6, 8, 10]$

Step 1: Calculate the means of X and Y.

$$\text{Mean of } X = \bar{X} = \frac{1+2+3+4+5}{5} = 3$$

$$\text{Mean of } Y = \bar{Y} = \frac{2+4+6+8+10}{5} = 6$$

Step 2: Subtract the mean from each value to get the deviation scores.

For X:

$$X - \bar{X} = [1-3, 2-3, 3-3, 4-3, 5-3] = [-2, -1, 0, 1, 2]$$

For Y:

$$Y - \bar{Y} = [2-6, 4-6, 6-6, 8-6, 10-6] = [-4, -2, 0, 2, 4]$$

Step 3: Multiply the corresponding deviation scores.

$$(X - \bar{X})(Y - \bar{Y}) = [-2 \times -4, -1 \times -2, 0 \times 0, 1 \times 2, 2 \times 4] = [8, 2, 0, 2, 8]$$

Step 4: Sum these products.

$$\text{Sum} = 8 + 2 + 0 + 2 + 8 = 20$$

Step 5: Calculate the square of the deviation scores for both X and Y.

For X:

$$(X - \bar{X})^2 = [-2^2, -1^2, 0^2, 1^2, 2^2] = [4, 1, 0, 1, 4]$$

$$\text{Sum of squares for } X = 4 + 1 + 0 + 1 + 4 = 10$$

For Y:

$$(Y - \bar{Y})^2 = [-4^2, -2^2, 0^2, 2^2, 4^2] = [16, 4, 0, 4, 16]$$

$$\text{Sum of squares for } Y = 16 + 4 + 0 + 4 + 16 = 40$$

Step 6: Plug these values into the Pearson correlation coefficient formula:

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

$$r = \frac{20}{\sqrt{10 \times 40}} = \frac{20}{\sqrt{400}} = \frac{20}{20} = 1$$

Conclusion:

The Pearson correlation coefficient r is 1, which indicates a perfect positive linear relationship between the variables X and Y.