Chi-Square Test for Independence Documentation

# Overview

The Chi-Square Test for Independence is a statistical test used to determine whether there is a significant association between two categorical variables. It is based on the comparison of observed frequencies in the data with the frequencies that would be expected if the variables were independent.

# Mathematical Logic

The Chi-Square statistic (χ²) is calculated using the following formula:  
χ² = Σ((O\_i - E\_i)² / E\_i)  
  
Where:  
- O\_i = Observed frequency for the i-th category  
- E\_i = Expected frequency for the i-th category  
  
The expected frequency (E\_i) for each category is calculated under the assumption of independence between the variables, using:  
E\_i = (Row Total \* Column Total) / Grand Total

# Steps to Perform Chi-Square Test

**1. Create a Contingency Table:** Summarize the frequencies of the two categorical variables in a matrix format.  
**2. Calculate Expected Frequencies:** Compute the expected frequencies for each cell of the table assuming the variables are independent.  
**3. Compute the Chi-Square Statistic:** Use the formula to calculate the χ² value.  
**4. Determine the Degrees of Freedom (df):** Calculated as (number of rows - 1) \* (number of columns - 1).  
**5. Find the P-Value:** Compare the χ² value with the Chi-Square distribution table to find the p-value.  
**6. Interpret the Result:** If the p-value is less than the significance level (typically 0.05), reject the null hypothesis of independence.

# Example Calculation

Consider a dataset with two variables: Gender (Male, Female) and Preference (Sports, Reading). Suppose we have the following observed frequencies in a contingency table:  
  
| | Sports | Reading | Row Total |  
|-------------|--------|---------|-----------|  
| Male | 30 | 20 | 50 |  
| Female | 20 | 30 | 50 |  
| Column Total| 50 | 50 | 100 |  
  
1. Expected Frequencies:  
E(Male, Sports) = (50 \* 50) / 100 = 25  
E(Male, Reading) = (50 \* 50) / 100 = 25  
E(Female, Sports) = (50 \* 50) / 100 = 25  
E(Female, Reading) = (50 \* 50) / 100 = 25  
  
2. Chi-Square Statistic:  
χ² = Σ((O\_i - E\_i)² / E\_i) = ((30 - 25)² / 25) + ((20 - 25)² / 25) + ((20 - 25)² / 25) + ((30 - 25)² / 25)  
χ² = (5² / 25) + ((-5)² / 25) + ((-5)² / 25) + (5² / 25) = 4  
  
3. Degrees of Freedom:  
df = (2-1)(2-1) = 1  
  
4. P-Value: Using the Chi-Square distribution table or a calculator, find the p-value corresponding to χ² = 4 and df = 1.

# Uses of Chi-Square Test

The Chi-Square Test for Independence is widely used in various fields, including:  
1. Biology: To determine if there is an association between different genetic traits.  
2. Social Sciences: To analyze survey data and examine relationships between demographic variables.  
3. Market Research: To evaluate consumer preferences and behaviors based on categorical variables like gender, age group, etc.  
4. Medical Research: To investigate the relationship between risk factors and health outcomes.

# Interpretation

P-Value < 0.05: Reject the null hypothesis; there is a significant association between the variables.  
P-Value ≥ 0.05: Fail to reject the null hypothesis; no significant association exists between the variables.  
  
This test allows researchers and analysts to make informed decisions based on the relationships between categorical variables in their data.

# Coding Implementation

## 1. Generate Dataset

First, we generate a dataset with 200 samples, each having two categorical variables: Gender and Preference. The dataset is saved to a CSV file for further analysis.

```python  
import pandas as pd  
import numpy as np  
  
# Seed for reproducibility  
np.random.seed(42)  
  
# Generate sample data  
n\_samples = 200  
genders = np.random.choice(['Male', 'Female'], size=n\_samples)  
preferences = np.random.choice(['Sports', 'Reading'], size=n\_samples)  
  
# Create a DataFrame  
data = {  
 'Gender': genders,  
 'Preference': preferences  
}  
df = pd.DataFrame(data)  
  
# Save DataFrame to CSV in the current environment  
csv\_file\_path = 'large\_sample\_data.csv'  
df.to\_csv(csv\_file\_path, index=False)  
print(f"CSV file created at: {csv\_file\_path}")  
```

## 2. Data Visualization

Next, we visualize the distribution of the Gender and Preference variables, and create a heatmap of the contingency table.

```python  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
# Read the CSV file  
df = pd.read\_csv('large\_sample\_data.csv')  
  
# Plot the distribution of Gender  
plt.figure(figsize=(10, 4))  
  
plt.subplot(1, 2, 1)  
sns.countplot(data=df, x='Gender')  
plt.title('Distribution of Gender')  
  
# Plot the distribution of Preference  
plt.subplot(1, 2, 2)  
sns.countplot(data=df, x='Preference')  
plt.title('Distribution of Preference')  
  
plt.tight\_layout()  
plt.show()  
  
# Create a contingency table  
contingency\_table = pd.crosstab(df['Gender'], df['Preference'])  
  
# Plot the heatmap of the contingency table  
plt.figure(figsize=(8, 6))  
sns.heatmap(contingency\_table, annot=True, fmt='d', cmap='YlGnBu')  
plt.title('Contingency Table Heatmap')  
plt.show()  
```

## 3. Chi-Square Test for Independence

Finally, we perform the Chi-Square Test for Independence to check if there is a significant association between Gender and Preference.

```python  
from scipy.stats import chi2\_contingency  
  
# Perform the Chi-Square Test for Independence  
chi2, p, dof, expected = chi2\_contingency(contingency\_table)  
  
print(f"Chi-Square Statistic: {chi2}")  
print(f"P-Value: {p}")  
print(f"Degrees of Freedom: {dof}")  
print("Expected Frequencies:")  
print(expected)  
  
# Interpret the p-value  
alpha = 0.05  
if p < alpha:  
 print("There is a significant association between Gender and Preference.")  
else:  
 print("There is no significant association between Gender and Preference.")  
```