# Chi Square

- The **Chi-Square test** is a statistical test used to determine whether there is a significant association between two **categorical variables**.
- It compares observed frequencies with expected frequencies to check if any differences are due to chance.

## Types of Chi-Square Tests

### 1. Chi-Square Goodness-of-Fit Test

- Checks if a sample follows a specific distribution.
- Tests if observed frequencies match **expected frequencies for a single categorical variable**.
- Example: Does the distribution of colors in a bag of candies match the expected distribution?
- Example: Checking if a die is fair.

**Formula**

The Chi-Square statistic is calculated as:

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

Where:

- $O_i$: Observed frequency for category $i$.
- $E_i$: Expected frequency for category $i$.

**Degrees of Freedom**

$$df = k - 1$$

Where:

- $k$: Number of categories.

## 2. Chi-Square Test for Independence

- Checks if two categorical variables are related.
- To test if there's a relationship between two categorical variables
- Example: Examining if gender and voting preference are related.

## Hypotheses

- **Null Hypothesis ($H_0$):** The variables are independent (no association).
- **Alternative Hypothesis ($H_1$):** The variables are dependent (there is an association).

## Formula:

**Formula**

The Chi-Square statistic is calculated as:

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

Where:

- $O_{ij}$: Observed frequency in cell $(i, j)$.
- $E_{ij}$: Expected frequency in cell $(i, j)$, calculated as:

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

**Degrees of Freedom**

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

Where:

- $r$: Number of rows.
- $c$: Number of columns.

💡 If the **Chi-Square statistic is large**, it suggests that observed values significantly deviate from expected values, implying a relationship between variables.

## Degrees of Freedom (DOF):

$$DOF = (\text{Number of Rows} - 1) \times (\text{Number of Columns} - 1)$$

- It claims about population proportions.
- **Non-parametric test**: A **non-parametric test** is a type of statistical test that does not assume the data follows a specific distribution, like the normal distribution
  - When you are given any proportion, you use this.
- Performed on **categorical variables (nominal or ordinal) data.**
  - **Categorical variables** are types of data that represent categories or groups.
    - no meaningful order or numeric value
  - **Nominal Data** (Names or Labels): C**ategories with no order or ranking**.
    - Ex. color, gender, animal type
  - **Ordinal Data** (Ordered Categories): meaningful order or ranking
    - Ex. **Education Level**: High School, Bachelor's, Master's, PhD
    - **Rating Scale**: Poor, Fair, Good, Excellent

The exact difference between the categories is not clearly defined.

Q. *In 2000 Indian census, the ages of the individuals in a small town were found to be:*

*<18 → 20%*

*18-35 → 30%*

*>35 → 50%*

*In 2010, n=500 individuals were sampled:*

*<18 → 121*

*18-35 → 288*

*>35 → 91*

*Using alpha is equal to 0.05, would you conclude the population distribution of ages has changed in the last 10 years?*

**Potential Year-2000 expected**

| <18 | 18-35 | >35 |
|---|---|---|
| 20% | 30% | 50% |

**n=500, observed**

| <18 | 18-35 | >35 |
|---|---|---|
| 121 | 288 | 91 |

Expected 2000 *census* data with n=500

| <18 | 18-35 | >35 |
|---|---|---|
| 500*0.2= **100** | 500*0.3= **150** | 500*0.5=**250** |

By only seeing this data, we can tell there is a difference. But we have to take the **95% CI** in account.

| | | | |
|---|---|---|---|
| 121 | 288 | 91 | **Observed** |
| 100 | 150 | 250 | **Expected** |

$H_0$= Data meets the distribution of 2000 census.

$H_1$= Data does not meet the distribution of 2000 census.

α= 0.05

Degree of freedom= n-1 = 3-3=2

n is number of categories (<18,18-35,>35)

Check in chi square table

| df | $\chi^2_{.995}$ | $\chi^2_{.990}$ | $\chi^2_{.975}$ | $\chi^2_{.950}$ | $\chi^2_{.900}$ | $\chi^2_{.100}$ | $\chi^2_{.050}$ | $\chi^2_{.025}$ | $\chi^2_{.010}$ | $\chi^2_{.005}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.000 | 0.000 | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.833 | 15.086 | 16.750 |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |

χ2 = 5.991 (Chi Square)

If Chi square is more than 5.991, we reject the null hypothesis.

**Calculate Test Statistics:**

$$\chi^2 = \sum \frac{(f_O - f_E)^2}{f_E}$$

$f_O$ = observed frequencies
$f_E$ = expected frequencies

**(121-100)^2/100 + (288-150)^2/150+(91-250)^2/250**

= 232.494

232.494>5.99

Therefore, we will reject the H0.

Hence, *population distribution of ages has changed in the last 10 years?*

## Python Code: Test of Independence

***Q. A company surveys 200 employees to determine if job satisfaction is related to the department. The observed data is:***

| Department | Satisfied | Not Satisfied | Total |
|---|---|---|---|
| IT | 40 | 30 | 70 |
| HR | 25 | 15 | 40 |
| Sales | 50 | 40 | 90 |
| Total | 115 | 85 | 200 |

We want to test:

- **Null Hypothesis ($H_0$):** Job satisfaction and department are independent.
- **Alternative Hypothesis ($H_1$):** Job satisfaction depends on the department.

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

# Observed frequency table
observed = np.array([[40, 30],   # IT
                     [25, 15],   # HR
                     [50, 40]])  # Sales

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

# Print results
print(f"Chi-Square Statistic: {chi2_stat:.4f}")  # Chi-square value
print(f"P-Value: {p_value:.4f}")  # Probability of getting this result if H0 is true
print(f"Degrees of Freedom: {dof}")  # DOF calculation
print("Expected Frequencies:\n", expected)  # Table of expected frequencies
```

**Output:**

*Chi-Square Statistic: 0.5521*
*P-Value: 0.7588*
*Degrees of Freedom: 2*
*Expected Frequencies:*
*[[40.25 29.75]*
*[23.   17.  ]*
*[51.75 38.25]]*

**If p<0.05, job satisfaction depends on the department.**

**If p>0.05, job satisfaction and department are independent.**

## Python:  Goodness-of-Fit Test

```
from scipy.stats import chisquare

# Observed frequencies
observed = [30, 20, 25, 35]

# Expected frequencies (hypothesized distribution)
expected = [25, 25, 25, 25]

# Perform Chi-Square test
chi2_stat, p_value = chisquare(observed, f_exp=expected)

print(f"Chi-Square Statistic: {chi2_stat:.4f}")
print(f"P-value: {p_value:.4f}")

# Interpret the result
alpha = 0.05
if p_value < alpha:
    print("Reject H₀: The observed frequencies do not match the expected freq
uencies.")
else:
```

```
    print("Fail to reject H₀: The observed frequencies match the expected frequencies.")
```

**Output**:

```
Chi-Square Statistic: 6.0000
P-value: 0.1116
Fail to reject H₀: The observed frequencies match the expected frequencies.
```

**Another example:**

A **restaurant owner** claims that customers order different dishes in the following proportions:

- **Pizza:** 40%
- **Burger:** 35%
- **Pasta:** 25%

We surveyed **200** customers and recorded their actual orders:

- **Pizza:** 85
- **Burger:** 70
- **Pasta:** 45

We test if the observed data follows the expected proportions.

→

## Step 1: Define Observed and Expected Counts

- **Observed counts:** Actual number of customer orders.
- **Expected counts:** Compute using total sample size and claimed proportions.

For **Pizza**:

$$E = 200 \times 0.4 = 80$$

Similarly, calculate for **Burger** and **Pasta**.

### Step 2: Compute Chi-Square Statistic

Formula:

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

Where:

- $O$ = Observed count
- $E$ = Expected count

---

### Step 3: Compute P-Value and Compare with Alpha

- If **p-value < 0.05**, reject $H_0$ (data does not follow expected proportions).
- If **p-value > 0.05**, fail to reject $H_0$ (data matches expected proportions).

## Python Code:

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

# Observed frequencies (actual customer orders)
observed = np.array([85, 70, 45])

# Expected frequencies based on claimed proportions
expected_proportions = np.array([0.4, 0.35, 0.25])  # Given proportions
sample_size = np.sum(observed)  # Total customers surveyed
```

```
expected = expected_proportions * sample_size  # Compute expected counts

# Perform Chi-Square Goodness-of-Fit Test
chi2_stat, p_value = stats.chisquare(f_obs=observed, f_exp=expected)

# Print results
print(f"Chi-Square Statistic: {chi2_stat:.4f}")
print(f"P-Value: {p_value:.4f}")

# Interpretation
alpha = 0.05  # Significance level
if p_value < alpha:
    print("Reject the null hypothesis: The data does not follow the expected distribution.")
else:
    print("Fail to reject the null hypothesis: The data follows the expected distribution.")
```

**Output:**

*Chi-Square Statistic: 0.8125*
*P-Value: 0.6661*
*Fail to reject the null hypothesis: The data follows the expected distribution.*