The Chi-Square Distribution

Goodness-of-Fit Test, Test of Independence, Test for Homogeneity

Noara Razzak

May 9, 2020

The Chi-Square Distribution

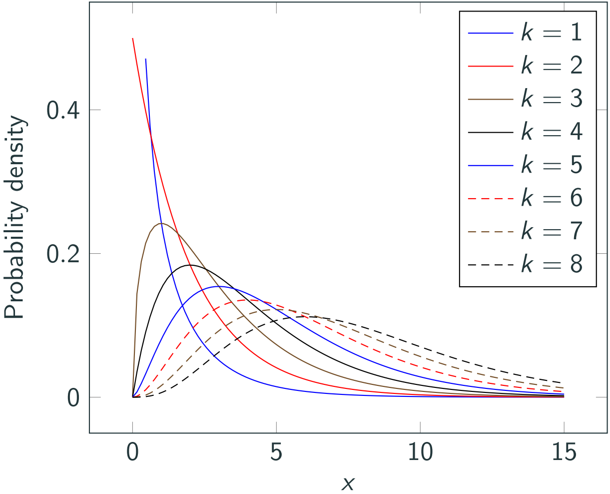
The Chi-Square Distribution (χ^2) is a continuous probability distribution that arises in statistical hypothesis testing, particularly in tests of goodness-of-fit, independence, and homogeneity.

- It is defined for positive real numbers and is skewed to the right.
- Its shape depends on the degrees of freedom (k). As k increases, the distribution becomes more symmetric.
- If $Z_1, Z_2, \ldots, Z_k$ are independent standard normal random variables, then:

$$Q = \sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$$

  where Q follows a chi-square distribution with k degrees of freedom.

Chi-Square Distribution



Chi-Square Goodness-of-Fit Test

The Goodness-of-Fit Test determines whether a sample matches a population with a specific distribution.

Hypotheses

- H_0: The sample follows the specified distribution.
- H_1: The sample does not follow the specified distribution.

Test Statistic

The test statistic is given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where:

- O_i = Observed frequency in category i
- E_i = Expected frequency in category i under H_0
- n = Number of categories

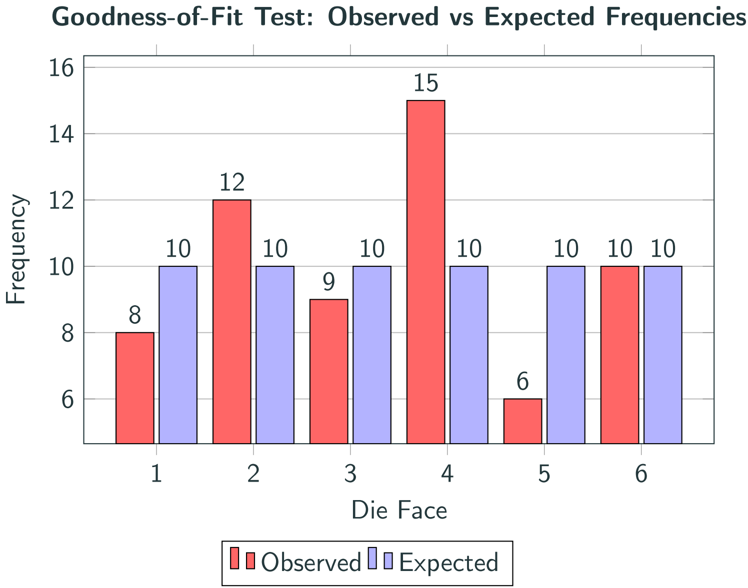
Decision Rule Reject H_0 if $\chi^2 > \chi^2_{\alpha,df}$, where $df = n - 1 -$ number of estimated parameters.

Example 1

Fairness of a Die: A casino wants to test if a six-sided die is fair (i.e., each face has an equal probability of landing face up). The die is rolled 60 times, and the observed frequencies are recorded:

Table 1: Observed vs. Expected Frequencies

Face	1	2	3	4	5	6
Observed (O_i)	8	12	9	15	6	10
Expected (E_i)	10	10	10	10	10	10


Goodness-of-Fit Test: Observed vs Expected Frequencies
16
14
12
10
8
6
Frequency
15
12
10
10
10
10
10
10
10
8
9
6
1
2
3
4
5
6
Die Face
Observed
Expected

1: State Hypotheses

- H_0: The die is fair (each face has probability $1/6$).
- H_1: The die is not fair (at least one face has a different probability).

2: Calculate the Test Statistic

The chi-square statistic is:

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(15-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(10-10)^2}{10}$$

$$\chi^2 = \frac{4}{10} + \frac{4}{10} + \frac{1}{10} + \frac{25}{10} + \frac{16}{10} + \frac{0}{10} = 0.4 + 0.4 + 0.1 + 2.5 + 1.6 + 0 = 5.0$$

3: Determine Degrees of Freedom

$$df = k - 1 = 6 - 1 = 5$$

where k is the number of categories (faces).

4: Compare to Critical Value At $\alpha = 0.05$ and $df = 5$, the critical value from the chi-square table is:

$$\chi^2_{0.05,5} = 11.07$$

Since our calculated $\chi^2 = 5.0 < 11.07$, we **fail to reject** H_0.

Conclusion There is **insufficient evidence** to conclude that the die is unfair at the 5% significance level.

Chi-Square Test of Independence

The Test of Independence assesses whether two categorical variables are independent.

Hypotheses

- H_0: The two variables are independent.
- H_1: The two variables are dependent.

Test Statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

- O_{ij} = Observed frequency in cell (i, j)
- $E_{ij} = \frac{(\text{Row } i \text{ Total}) \times (\text{Column } j \text{ Total})}{\text{Grand Total}}$
- r = Number of rows, c = Number of columns

Degrees of Freedom

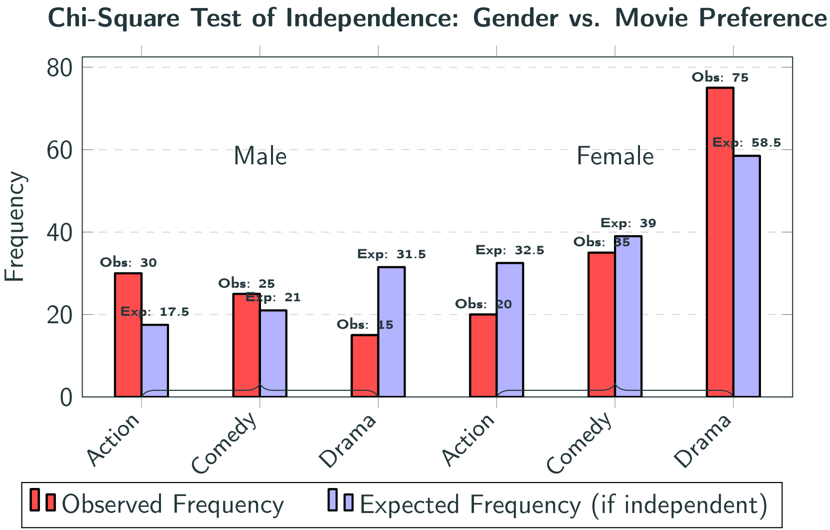
$$df = (r - 1)(c - 1)$$

Example 2

Gender vs. Movie Preference: A survey asks 200 people about their gender and preferred movie genre. Test whether gender and movie preference are independent ($\alpha = 0.05$).

Table 2: Observed Frequencies

	Action	Comedy	Drama	Total
Male	30	25	15	70
Female	20	35	75	130
Total	50	60	90	200


Chi-Square Test of Independence: Gender vs. Movie Preference
Frequency
80
60
40
20
0
Male
Female
Obs: 30
Exp: 17.5
Obs: 25
Exp: 21
Obs: 15
Exp: 31.5
Obs: 20
Exp: 32.5
Obs: 35
Exp: 39
Obs: 75
Exp: 58.5
Action
Comedy
Drama
Action
Comedy
Drama
Observed Frequency
Expected Frequency (if independent)

1: State Hypotheses

- H_0: Gender and movie preference are independent
- H_1: Gender and movie preference are dependent

2: Calculate Expected Frequencies

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

Table 3: Expected Frequencies

	Action	Comedy	Drama
Male	$\frac{70 \times 50}{200} = 17.5$	$\frac{70 \times 60}{200} = 21$	$\frac{70 \times 90}{200} = 31.5$
Female	$\frac{130 \times 50}{200} = 32.5$	$\frac{130 \times 60}{200} = 39$	$\frac{130 \times 90}{200} = 58.5$

3: Compute Chi-Square Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(30-17.5)^2}{17.5} + \frac{(25-21)^2}{21} + \cdots + \frac{(75-58.5)^2}{58.5} = 28.42$$

4: Determine Critical Value Degrees of freedom:
$df = (r-1)(c-1) = (2-1)(3-1) = 2$
Critical value: $\chi^2_{0.05,2} = 5.991$

5: Conclusion Since $28.42 > 5.991$, we reject H_0. There is significant evidence that gender and movie preference are associated.

Chi-Square Test for Homogeneity The test for Homogeneity checks whether different populations have the same distribution of a categorical variable.

Hypotheses

- H_0: The distributions are the same across populations.
- H_1: The distributions differ across populations.

Test Statistic

The test for Homogeneity is the same as the test of Independence:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

- O_{ij} = Observed frequency for population i and category j
- E_{ij} = Expected frequency under homogeneity

Degrees of Freedom

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

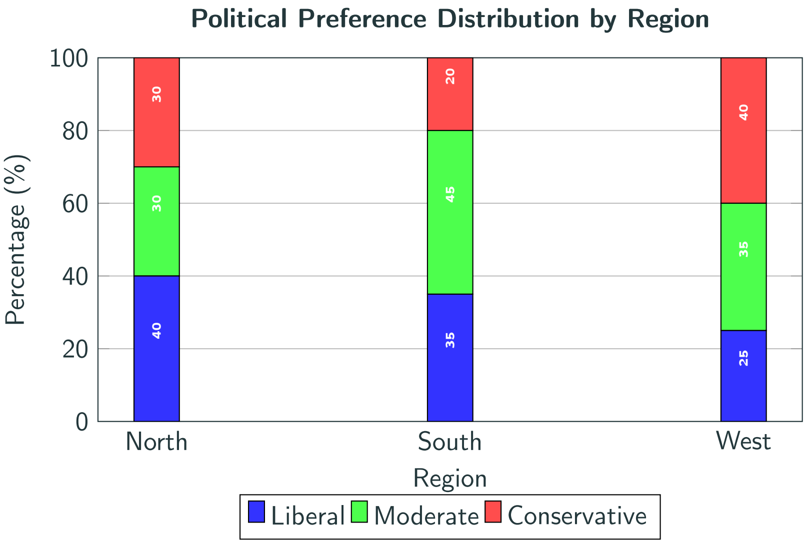
Note: The test of Independence and test for Homogeneity use the same formula but differ in their hypotheses and sampling design.

Example 3

Political Preference Across Regions: A researcher wants to test if the distribution of political preferences (Liberal, Moderate, Conservative) is the same across three regions (North, South, West). Data from 300 respondents:

Table 4: Observed Frequencies

	Liberal	Moderate	Conservative
North	40	30	30
South	35	45	20
West	25	35	40


Political Preference Distribution by Region
Percentage (%)
0
20
40
60
80
100
40
30
30
35
45
20
25
35
40
North
South
West
Region
Liberal
Moderate
Conservative

1: State Hypotheses

- H_0: The distribution of political views is *homogeneous* across regions
- H_1: At least one region has a different distribution

2: Calculate Expected Frequencies

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

Table 5: Expected Frequencies (if homogeneous)

	Liberal	Moderate	Conservative
North	$\frac{100\times100}{300} = 33.3$	$\frac{100\times110}{300} = 36.7$	$\frac{100\times90}{300} = 30$
South	$\frac{100\times100}{300} = 33.3$	$\frac{100\times110}{300} = 36.7$	$\frac{100\times90}{300} = 30$
West	$\frac{100\times100}{300} = 33.3$	$\frac{100\times110}{300} = 36.7$	$\frac{100\times90}{300} = 30$

3: Compute Test Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(40-33.3)^2}{33.3} + \cdots + \frac{(40-30)^2}{30} = 13.36$$

4: Determine Critical Value Degrees of freedom:
$df = (r-1)(c-1) = (3-1)(3-1) = 4$
Critical value ($\alpha = 0.05$): $\chi^2_{0.05,4} = 9.488$

5: Conclusion Since $13.36 > 9.488$, we reject H_0. The distribution of political views differs significantly across regions.

Table 6: Comparison of Goodness-of-Fit, Test of Independence, and Test for Homogeneity

Feature	Goodness-of-Fit Test	Test of Independence	Test for Homogeneity
Purpose	Checks if sample data fits a theoretical distribution	Determines if two categorical variables are independent	Checks if different populations have the same distribution for a categorical variable
Data Structure	One categorical variable with observed vs. expected frequencies	Two categorical variables in a contingency table	One categorical variable compared across multiple populations

Table 6: Comparison of Goodness-of-Fit, Test of Independence, and Test for Homogeneity

Feature	Goodness-of-Fit Test	Test of Independence	Test for Homogeneity
Null Hypothesis (H_0)	The observed distribution matches the expected distribution	The two variables are independent	The distributions are the same across different populations
Sample Requirements	One sample, one variable	One sample, two variables	Multiple samples, one variable

Table 6: Comparison of Goodness-of-Fit, Test of Independence, and Test for Homogeneity

Feature	Goodness-of-Fit Test	Test of Independence	Test for Homogeneity
Degrees of Freedom	$k-1$ (where $k =$ number of categories)	$(r-1)(c-1)$ (for an $r \times c$ table)	$(r-1)(c-1)$ (similar to independence)
Chi-Square Statistic	$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$	$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$	Same formula as Test of Independence

Next class we will cover F Distribution and One-Way ANOVA, Facts About the F Distribution, Test of Two Variances