Chi-Square Test of Difference of More Than Two Proportions

Dr. Kalyan N

October 14, 2024

Chi-Square Test for Difference of More Than Two Proportions

- The **Chi-Square Test of Difference of More than Two Proportions** is used to determine whether the proportions of multiple groups differ significantly.
- It compares the *observed frequencies* and *expected frequencies*.
- Hypotheses:
  - H_0: The proportions are equal.
  - H_A: At least one proportion is different.
- Test Statistic:

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

where O_i is the observed frequency and E_i is the expected frequency.

Problem 1: Coffee Preferences

A coffee shop surveyed 200 customers about their preferences for three types of coffee:

- Espresso
- Latte
- Cappuccino

The observed frequencies were:

- Espresso: 80
- Latte: 70
- Cappuccino: 50

Test whether the proportions of customers preferring these types of coffee are significantly different at the 5% significance level.

Solution: Coffee Preferences - Step 1

Step 1: Formulate Hypotheses

- H_0: The proportions of customers preferring Espresso, Latte, and Cappuccino are equal.
- H_A: At least one proportion is different.

Solution: Coffee Preferences - Step 2

Step 2: Calculate Expected Frequencies

The total number of customers is $N = 200$. Under H_0, the expected frequency for each coffee type is:

$$E = \frac{N}{3} = \frac{200}{3} \approx 66.67$$

So, the expected frequencies are:

- Espresso: $E = 66.67$
- Latte: $E = 66.67$
- Cappuccino: $E = 66.67$

Solution: Coffee Preferences - Step 3

Step 3: Compute the Chi-Square Test Statistic

The observed frequencies O_i for Espresso, Latte, and Cappuccino are 80, 70, and 50, respectively. Using the formula:

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

We calculate:

$$\chi^2 = \frac{(80 - 66.67)^2}{66.67} + \frac{(70 - 66.67)^2}{66.67} + \frac{(50 - 66.67)^2}{66.67}$$

$$\chi^2 = \frac{13.33^2}{66.67} + \frac{3.33^2}{66.67} + \frac{-16.67^2}{66.67}$$

$$\chi^2 = \frac{177.69}{66.67} + \frac{11.09}{66.67} + \frac{278.09}{66.67}$$

$$\chi^2 = 2.67 + 0.17 + 4.17 = 7.01$$

Solution: Coffee Preferences - Step 4

Step 4: Compare with Critical Value

- Degrees of freedom: $df = k - 1 = 3 - 1 = 2$
- From the Chi-Square table, the critical value at $\alpha = 0.05$ for $df = 2$ is 5.991.

Conclusion:

- Since $\chi^2 = 7.01 > 5.991$, we reject the null hypothesis.
- There is sufficient evidence to conclude that the proportions of customers preferring the three types of coffee are significantly different.

Problem 2: Voting Preferences

In a survey of 300 people, their preferences for three political parties were recorded:

- Party A: 100
- Party B: 120
- Party C: 80

Test if the proportions of voters for these three parties are significantly different at a 1% significance level.

Solution: Voting Preferences - Step 1

Step 1: Formulate Hypotheses

- H_0: The proportions of voters for Party A, Party B, and Party C are equal.
- H_A: At least one proportion is different.

Solution: Voting Preferences - Step 2

Step 2: Calculate Expected Frequencies

The total number of people is $N = 300$. Under H_0, the expected frequency for each party is:

$$E = \frac{N}{3} = \frac{300}{3} = 100$$

The expected frequencies are:

- Party A: $E = 100$
- Party B: $E = 100$
- Party C: $E = 100$

Solution: Voting Preferences - Step 3

Step 3: Compute the Chi-Square Test Statistic

The observed frequencies O_i for Party A, Party B, and Party C are 100, 120, and 80, respectively. Using the formula:

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

We calculate:

$$\chi^2 = \frac{(100 - 100)^2}{100} + \frac{(120 - 100)^2}{100} + \frac{(80 - 100)^2}{100}$$

$$\chi^2 = \frac{0^2}{100} + \frac{20^2}{100} + \frac{-20^2}{100}$$

$$\chi^2 = 0 + \frac{400}{100} + \frac{400}{100} = 8$$

Solution: Voting Preferences - Step 4

Step 4: Compare with Critical Value

- Degrees of freedom: $df = k - 1 = 3 - 1 = 2$
- From the Chi-Square table, the critical value at $\alpha = 0.01$ for $df = 2$ is 9.210.

Conclusion:

- Since $\chi^2 = 8 < 9.210$, we fail to reject the null hypothesis.
- There is insufficient evidence to conclude that the proportions of voters for the three parties are significantly different at the 1% significance level.

Chi-Square Test of Independence of Attributes

- The **Chi-Square Test of Independence** tests whether two categorical variables are independent.
- It compares the observed frequencies with the expected frequencies, assuming independence between the variables.
- Hypotheses:
  - H_0: The two variables are independent.
  - H_A: The two variables are dependent.
- Test Statistic:

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

where O_i is the observed frequency and E_i is the expected frequency.

Problem 1: Gender and Voting Preference

A survey was conducted to check if gender influences voting preferences. The data collected from 200 people is as follows:

Gender	**Party A**	**Party B**	**Total**
Male	40	60	100
Female	30	70	100
Total	70	130	200

Test whether gender and voting preference are independent at the 5% significance level.

Solution: Gender and Voting Preference - Step 1

Step 1: Formulate Hypotheses

- H_0: Gender and voting preference are independent.
- H_A: Gender and voting preference are dependent.

Solution: Gender and Voting Preference - Step 2

Step 2: Calculate Expected Frequencies

The expected frequency E for each cell is calculated as:

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

For example, the expected frequency for Males voting for Party A is:

$$E_{(Male,A)} = \frac{100 \times 70}{200} = 35$$

Similarly, calculate the expected frequencies for the other cells:

$$E_{(Male,B)} = \frac{100 \times 130}{200} = 65$$

$$E_{(Female,A)} = \frac{100 \times 70}{200} = 35$$

$$E_{(Female,B)} = \frac{100 \times 130}{200} = 65$$

Solution: Gender and Voting Preference - Step 3

Step 3: Compute the Chi-Square Test Statistic

Using the formula:

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

We calculate the contributions for each cell:

$$\chi^2 = \frac{(40-35)^2}{35} + \frac{(60-65)^2}{65} + \frac{(30-35)^2}{35} + \frac{(70-65)^2}{65}$$

$$\chi^2 = \frac{25}{35} + \frac{25}{65} + \frac{25}{35} + \frac{25}{65} = 0.71 + 0.38 + 0.71 + 0.38 = 2.18$$

Solution: Gender and Voting Preference - Step 4

Step 4: Compare with Critical Value

- Degrees of freedom: $df = (r-1)(c-1) = (2-1)(2-1) = 1$
- From the Chi-Square table, the critical value at $\alpha = 0.05$ for $df = 1$ is 3.841.

Conclusion:

- Since $\chi^2 = 2.18 < 3.841$, we fail to reject the null hypothesis.
- There is insufficient evidence to conclude that gender and voting preference are dependent.

Problem 2: Smoking and Lung Disease

A study was conducted to determine if smoking is associated with lung disease. The data from 150 people is shown below:

Smoking Status	**Lung Disease**	**No Lung Disease**	**Total**
Smoker	30	20	50
Non – Smoker	10	90	100
Total	40	110	150

Test whether smoking status and lung disease are independent at the 1% significance level.

Solution: Smoking and Lung Disease - Step 1

Step 1: Formulate Hypotheses

- H_0: Smoking status and lung disease are independent.
- H_A: Smoking status and lung disease are dependent.

Solution: Smoking and Lung Disease - Step 2

Step 2: Calculate Expected Frequencies

Using the same method as in Problem 1, calculate the expected frequencies:

$$E_{(Smoker,Disease)} = \frac{50 \times 40}{150} = 13.33$$

$$E_{(Smoker,NoDisease)} = \frac{50 \times 110}{150} = 36.67$$

$$E_{(Non-Smoker,Disease)} = \frac{100 \times 40}{150} = 26.67$$

$$E_{(Non-Smoker,NoDisease)} = \frac{100 \times 110}{150} = 73.33$$

Solution: Smoking and Lung Disease - Step 3

Step 3: Compute the Chi-Square Test Statistic

Now, calculate χ^2:

$$\chi^2 = \frac{(30 - 13.33)^2}{13.33} + \frac{(20 - 36.67)^2}{36.67} + \frac{(10 - 26.67)^2}{26.67} + \frac{(90 - 73.33)^2}{73.33}$$

$$\chi^2 = \frac{278.89}{13.33} + \frac{286.89}{36.67} + \frac{278.89}{26.67} + \frac{278.89}{73.33}$$

$$\chi^2 = 20.92 + 7.83 + 10.46 + 3.8 = 43.01$$

Solution: Smoking and Lung Disease - Step 4

Step 4: Compare with Critical Value

- Degrees of freedom: $df = (2-1)(2-1) = 1$
- From the Chi-Square table, the critical value at $\alpha = 0.01$ for $df = 1$ is 6.635.

Conclusion:

- Since $\chi^2 = 43.01 > 6.635$, we reject the null hypothesis.
- There is sufficient evidence to conclude that smoking status and lung disease are dependent.

Chi-Square Test of Goodness of Fit

- The **Chi-Square Goodness of Fit** test determines whether an observed frequency distribution fits an expected distribution.
- It is used to compare observed frequencies with expected frequencies based on a theoretical model.
- Hypotheses:
  - H_0: The observed data fits the expected distribution.
  - H_A: The observed data does not fit the expected distribution.
- Test Statistic:

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

where O_i is the observed frequency and E_i is the expected frequency.

Problem 1: Dice Fairness Test

A six-sided die was rolled 60 times with the following results:

Outcome	Observed Frequency
1	8
2	10
3	12
4	9
5	11
6	10

Test if the die is fair at the 5% significance level.

Solution: Dice Fairness Test - Step 1

Step 1: Formulate Hypotheses

- H_0: The die is fair, i.e., all outcomes are equally likely.
- H_A: The die is not fair.

Solution: Dice Fairness Test - Step 2

Step 2: Calculate Expected Frequencies

If the die is fair, each outcome has an equal probability of $\frac{1}{6}$. Since the die was rolled 60 times, the expected frequency for each outcome is:

$$E = \frac{60}{6} = 10$$

Thus, the expected frequency for each outcome is 10.

Solution: Dice Fairness Test - Step 3

Step 3: Compute the Chi-Square Test Statistic

Using the formula:

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

We calculate:

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(10-10}{10}$$

$$\chi^2 = \frac{4}{10} + 0 + \frac{4}{10} + \frac{1}{10} + \frac{1}{10} + 0 = 1$$

Solution: Dice Fairness Test - Step 4

Step 4: Compare with Critical Value

- Degrees of freedom: $df = k - 1 = 6 - 1 = 5$
- From the Chi-Square table, the critical value at $\alpha = 0.05$ for $df = 5$ is 11.07.

Conclusion:

- Since $\chi^2 = 1 < 11.07$, we fail to reject the null hypothesis.
- There is insufficient evidence to conclude that the die is unfair.

Problem 2: Color Distribution in MM's

A candy manufacturer claims that the color distribution of MM's is:

- Red: 30%
- Green: 20%
- Blue: 20%
- Yellow: 15%
- Brown: 15%

A sample of 100 MM's was taken, and the observed frequencies were:

Color	**Observed Frequency**
Red	28
Green	22
Blue	18
Yellow	14
Brown	18

Test whether the observed distribution fits the claimed distribution at the 5% significance level.

Solution: MM's Color Distribution - Step 1

Step 1: Formulate Hypotheses

- H_0: The observed color distribution fits the expected distribution.
- H_A: The observed color distribution does not fit the expected distribution.

Solution: MM's Color Distribution - Step 2

Step 2: Calculate Expected Frequencies

The expected frequencies based on the claimed percentages are:

$$E_{Red} = 0.30 \times 100 = 30, \quad E_{Green} = 0.20 \times 100 = 20$$

$$E_{Blue} = 0.20 \times 100 = 20, \quad E_{Yellow} = 0.15 \times 100 = 15$$

$$E_{Brown} = 0.15 \times 100 = 15$$

Solution: MM's Color Distribution - Step 3

Step 3: Compute the Chi-Square Test Statistic

Now, calculate χ^2:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(28-30)^2}{30} + \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(14-15)^2}{15} + \frac{(}{}$$

$$\chi^2 = \frac{4}{30} + \frac{4}{20} + \frac{4}{20} + \frac{1}{15} + \frac{9}{15} = 0.13 + 0.2 + 0.2 + 0.067 + 0.6 = 1.197$$

Solution: MM's Color Distribution - Step 4

Step 4: Compare with Critical Value

- Degrees of freedom: $df = k - 1 = 5 - 1 = 4$
- From the Chi-Square table, the critical value at $\alpha = 0.05$ for $df = 4$ is 9.488.

Conclusion:

- Since $\chi^2 = 1.197 < 9.488$, we fail to reject the null hypothesis.
- There is insufficient evidence to conclude that the color distribution is different from the claimed distribution.

Characteristics of Chi-Square Test

(i) This test (as a non-parametric test) is based on frequencies and not on parameters like mean and standard deviation.

(ii) The test is used for testing the hypothesis and is not useful for estimation.

(iii) This test possesses the additive property as has already been explained.

(iv) This test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.

(v) This test is an important non-parametric test as no rigid assumptions are necessary regarding the type of population, no need for parameter values, and relatively less mathematical details are involved.

Proper Application of Chi-Square Test

- The chi-square test is no doubt a most frequently used test, but its correct application is equally an uphill task.
- It should be borne in mind that the test is to be applied only when the individual observations of the sample are independent, which means that the occurrence of one individual observation (event) has no effect upon the occurrence of any other observation (event) in the sample under consideration.

Common Mistakes in Chi-Square Test Application

- Small theoretical frequencies, if these occur in certain groups, should be dealt with under special care.
- Other possible reasons concerning the improper application or misuse of this test can be:
  - (i) Neglect of frequencies of non-occurrence.
  - (ii) Failure to equalize the sum of observed and the sum of the expected frequencies.
  - (iii) Wrong determination of the degrees of freedom.
  - (iv) Wrong computations, and the like.

Exercise Problems

The table given below shows the data obtained during outbreak of smallpox:

	Attacked	*Not attacked*	*Total*
Vaccinated	31	469	500
Not vaccinated	185	1315	1500
Total	216	1784	2000

Test the effectiveness of vaccination in preventing the attack from smallpox. Test your result with the help of χ^2 at 5 per cent level of significance.

Exercise Problems

Two research workers classified some people in income groups on the basis of sampling studies. Their results are as follows:

Investigators	*Income groups*			*Total*
	Poor	*Middle*	*Rich*	
A	160	30	10	200
B	140	120	40	300
Total	300	150	50	500

Exercise Problems

Eight coins were tossed 256 times and the following results were obtained:

Numbers of heads	0	1	2	3	4	5	6	7	8
Frequency	2	6	30	52	67	56	32	10	1

Are the coins biased? Use χ^2 test.

Exercise Problems

The following values of χ^2 from different investigations carried to examine the effectiveness of a recently invented medicine for checking malaria are obtained:

Investigation	χ^2	*d.f.*
1	2.5	1
2	3.2	1
3	4.1	1
4	3.7	1
5	4.5	1

What conclusion would you draw about the effectiveness of the new medicine on the basis of the five investigations taken together?