Course: Mathematical statistics

Week 10: chi-square test(test of goodness of fit, contingency table)

Lecturer: Nagulama Moses

Kumi University

May 11, 2025

Outline

Intended learning outcomes

- Apply the Chi-square goodness-of-fit test to assess how well observed categorical data match an expected distribution.
- Compute the Chi-square statistic from observed and expected frequencies.
- Justify conclusions about independence, homogeneity based on calculated statistics and significance levels.

Test involving contingency table

1. Test of independence
2. Test of Homogeneity

- Test of independence deals with determining whether two variables are independent or related to each other when a single sample is selected

- Test of homogeneity of proportions is used to determine whether the proportions for a variable are equal when several samples are selected from different population

When to use the Test

- You have two categorical variables (e.g., gender and preference).
- Your data is presented in a contingency table (cross-tabulation).
- The data is count data (frequencies, not percentages or means).
- You want to test for independence or association between the variables.

Test of independence

- consider n elements from a sample of n from a criteria e.g. salary of degree holder and salary of diploma holder
- assume that 1st criteria of classification has r levels
- assume that 2nd criteria of classification has c levels
- O_{ij} be the observed frequency for level i of the 1st criteria and level j for 2nd criteria

Table: Contingency Table for Chi-Square Test of Independence

Criteria 1	**criteria 2**			**Row Total**
	1	**2**	**c**	
1	O_{11}	O_{12}	O_{1c}	
2	O_{21}	O_{22}	O_{2c}	
r	O_{r1}	O_{r2}	O_{rc}	
Column Total				

$r \times c$ contingency tables where r and c level of 1st and 2nd criteria classification respectively

for large n the test statistics

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

approximately chi square distribution with$(r-1)(c-1)$ degrees of freedom.

- to test the null hypothesis
- H_0 : two variables are independent
- H_1 :two variables are not independent
- α
- critical region $\chi^2 \geq \chi^2_{\alpha,(r-1)(c-1)}$

- column totals are referred to as marginal frequencies
- p_{ij} to be the probability that a randomly selected element falls in the ij^{th} cell.
- Then $P_{ij} = U_i V_j$ because of the assumption of independent

$$U_i = \frac{1}{n} \sum_{i=1}^{r} O_{ij}$$

$$V_j = \frac{1}{n} \sum_{j=1}^{c} O_{ij}$$

- expected frequency of each cell

$$e_{ij} = \frac{rowsum \times columnsum}{grandtotal}$$

Example

The table below shows the educational qualification of employees and their job performance rating for employees in a certain factory

educational qualification		level I	level II	level III	
	BSC	18(20.4)	35(35.36)	15(12.24)	68
	MSE	12(9.6)	17(16.68)	3(5.76)	32
		30	52	18	100

Determine at 0.05 level of significance whether the job performance rating is independent of educational level.

solution

variable of interest is the job performance level among degree classification

H_0 : job performance level is independent of the type of degree

H_1 : job performance level is not independent of the type of degree

$\alpha = 0.05$

critical region

$$\chi^2 > \chi^2_{0.05,(2-1)(3-1)} = \chi^2_{0.05,2} = 5.99$$

expected frequency

$$= \frac{rowsum \times columnsum}{grandtotal}$$

$$= \frac{68 \times 30}{100} = 20.4$$

$$= \frac{68 \times 52}{100} = 25.8$$

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(18-20.4)^2}{20.4} + \frac{(35-35.36)^2}{35.36} +$$
$$\frac{(15-12.24)^2}{12.24} + \frac{(12-9.6)^2}{9.6} + \frac{(17-16.68)^2}{16.68} +$$
$$\frac{(3-5.76)^2}{5.76}$$
$$= 2.84$$

decision: we fail to reject the null hypothesis
conclusion: job performance rating in independent of the level of degree

Example

Use the data shown in the following table to test at 0.01 level of significance whether a persons interest statistics is independent of his or her ability in mathematics

Ability in statistics	**Ability in mathematics**			**Row Total**
	low	**Average**	**High**	
low	63(45)	42(50)	15(25)	120
Average	58(56.25)	61(62.5)	31(31.25)	150
High	14(33.75)	47(37.5)	29(18.75)	90
Column Total	135	150	75	360

variable of interest is the ability in mathematics
H_0 : ability in mathematics is independent
H_1 : ability in mathematics is not independent
$\alpha = 0.01$
critical region $\chi^2 > \chi^2_{0.01(2)(2)}$

$$\chi^2 > \chi^2_{0.01,4} = 13.28$$

$$\chi^2 = \frac{(63-45)^2}{45} + \frac{(42-50)^2}{50} + ... + \frac{(29-18.75)^2}{18.75} = 32.14$$

we reject the null hypothesis. ability in mathematics is independent.

test of homogeneity of proportions

samples are taken from different population and we determine whether proportion of elements having a common attribute characteristics is uniform or the same for each population

$$H_0 : p_1 = p_2 = ... = p_n$$

$$H_1 : p_1 \neq p_2 \neq ... = p_n$$

the test is carriedout in the same way for indepedence with $(r-1)(c-1)$ degrees of freedom

Example

the political affiliation of 500 politicians and their opinion about the new law to be introduced in uganda parliament are summarised in the table below

new law	political affiliation			Row Total row totals
	NRM	FDC	NUP	
FOR	82(85.6)	70(64.2)	62(64.2)	214
AGAINST	93(88.8)	62(66.6)	67(66.2)	222
UNDECIDED	25(25.6)	18(19.2)	21(19.2)	64
Column Total	200	150	150	500

Test at 0.01 level of significance that the opinion concerning the new law is the same within each political group

solution

interest in For each opinion in the proportion of politicians

H_0 : for each opinion the proportion is the same $p_1 = p_2 = p_3$

H_1 : for each opinion $p_1 \neq p_2 \neq p_3$

$\alpha = 0.01$

critical region$\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$

$$\chi^2_{0.01,(3-1)(3-1)} = \chi^2_{0.01,4} = 13.28$$

$$\chi^2 = \frac{(85-85.6)^2}{85.6} + \frac{(70-64.2)^2}{64.2} + \frac{(62-64.2)^2}{64.2} +$$
$$\frac{(93-88.8)^2}{88.8} + \frac{(62-66.6)^2}{66.6} + \frac{(67-66.2)^2}{66.2} +$$
$$\frac{(25-25.6)^2}{25.6} + \frac{(18-19.2)^2}{19.2} + \frac{(21-19.2)^2}{19.2}$$
$$= 1.56$$

since $1.56 < 13.28$
we fail to reject reject H_0
the proportion of politicians is the same across the political parties

Example

a researcher selected 150 students from 3 schools in which he is carrying out his study he then asked the student a question. will you like or hate maths. the results are shown in the table below.

feeling	**School**			**Row Total** row totals
	I	**II**	**III**	
like	18(21.3)	26(23.9)	20(18.8)	64
hates	32(28.7)	30(32.1)	24(25.2)	86
Column Total	50	56	44	150

At 0.05 level of significance test the claim that the proportion of students who like or hate mathematics are the same in all the 3 schools

interest if for proportion of students who like or hate mathematics
H_0 :proportion of students who like or hate mathematics is $p_1 = p_2 = p_3$
H_1 : proportion of students who like or hate mathematics is $p_1 \neq p_2 \neq p_3$
$\alpha = 0.05$
critical region $\chi^2 > \chi^2_{0.05,2} = 5.991$

$$\chi^2 = \frac{(18-21.3)^2}{21.3} + \frac{(26-23.9)^2}{23.9} + \frac{(20-18.8)^2}{18.8} + \frac{(32-28.7)^2}{28.7} + \frac{(30-32.1)^2}{32.1} + \frac{(24-25.2)^2}{25.2} = 1.35$$

we fail to reject the null hypothesis.

A researcher wants to know whether preferences for fruit types differ across three different regions. They survey 150 people, 50 from each region, asking their favorite fruit among Apple, Banana, and Orange. The results are summarized below:

Region	**Apple**	**Banana**	**Orange**	**Total**
Region A	20	15	15	50
Region B	10	20	20	50
Region C	15	25	10	50
Total	45	60	45	150

Question: At the 5% significance level, test whether the distribution of fruit preferences is the same across the three regions.

- H_0: The distribution of fruit preferences is the same across all three regions.
- H_1: At least one region has a different distribution of preferences.
- All data are frequencies (counts)
- Groups are independent samples
- Expected frequency in each cell should be ≥ 5

Use the formula:

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

Example: Expected frequency for Region A, Apple:

$$E_{A,\text{Apple}} = \frac{50 \times 45}{150} = 15$$

Expected frequency table:

Region	**Apple**	**Banana**	**Orange**	**Total**
Region A	15	20	15	50
Region B	15	20	15	50
Region C	15	20	15	50
Total	45	60	45	150

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

Region A:

$$\frac{(20-15)^2}{15} = \frac{25}{15} = 1.667, \quad \frac{(15-20)^2}{20} = \frac{25}{20} = 1.25, \quad \frac{(15-15)^2}{15} = 0$$

Region B:

$$\frac{(10-15)^2}{15} = 1.667, \quad \frac{(20-20)^2}{20} = 0, \quad \frac{(20-15)^2}{15} = 1.667$$

Region C:

$$\frac{(15-15)^2}{15} = 0, \quad \frac{(25-20)^2}{20} = 1.25, \quad \frac{(10-15)^2}{15} = 1.667$$

Total Chi-Square:

$$\chi^2 = 1.667 + 1.25 + 0 + 1.667 + 0 + 1.667 + 0 + 1.25 + 1.667 = 9.168$$

$$df = (r-1)(c-1) = (3-1)(3-1) = 2 \times 2 = 4$$

From the Chi-Square distribution table at $\alpha = 0.05$, with 4 degrees of freedom:

$$\chi^2_{0.05,4} = 9.488$$

Since $\chi^2_{calculated} = 9.168 < 9.488$,
we **fail to reject** H_0.
At the 5% significance level, there is **no sufficient evidence** to suggest that the fruit preferences differ significantly across the three regions.
The distributions appear to be the same.

References

- Hogg,R;Mckean,J;Craig,A(2012).Introduction to mathematical statistics, 7th edition, pearson Prentice Hall, 2012.

- Hastings K.J,(1997) Probability and statistics, Addison Wesley reading,massachusetts.

Thank You!

Next Lecture: correlation coefficient