Course: Mathematical statistics

Week 9: chi-square test(test of goodness of fit, contingency table)

Lecturer: Nagulama Moses

Kumi University

May 12, 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, or goodness-of-fit based on calculated statistics and significance levels.

Introduction

A random variable x is called a chi square distribution if the probability density function (PDF) of the chi-square distribution with k degrees of freedom is given by:

$$f(x;k) = \frac{1}{2^{k/2}\,\Gamma(k/2)}\, x^{(k/2-1)} e^{-x/2}, \quad x > 0$$

Where:

- x is the chi-square test statistic (with $x > 0$),
- k is the degrees of freedom,
- $\Gamma(\cdot)$ is the gamma function, where $\Gamma(n) = (n-1)!$ for a positive integer n.

Theorem

if a random sample of size n is obtained from a normal distributed population with mean μ and variance σ^2 then

$$W = \frac{(n-1)\chi^2}{\sigma^2}$$

is a chi square distribution with $(n-1)$ degrees of freedom.

$$\chi^2 = chi - square$$

characteristics of chi-square distribution

- It is not symmetric
- The shape of the distribution depends on the degree of freedom
- As the number of degrees of freedom increases, the chi square distribution becomes more symmetric
- The values of chi square are non negative

Goodness of fit Test

- In the chi square goodness of fit test the underlined distribution of the population is not known but wish to test using sample data that a

  particular distribution will be satisfied as a population model or put

  different that a given frequency distribution fits a specified pattern

Test procedure

- Make use of a random sample of size n from a population whose distribution is not known.
- The n observations are arranged in a frequency table having k^+ intervals
- Let θ_i be the observed frequency class for i^{th} interval.

- From the hypothesised probability distribution, the

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

with degree of freedom $k - p - 1$ where k is the number of frequency class interval, p is the number of parameters to be estimated in the hypothesised distribution

- We reject the null hypothesis if $\chi^2 > \chi^2_{\alpha,k-p-1}$.

Assumption of a chi square goodness of fit test

- The data are obtained from a random sample
- The expected frequency for each category must be 5 and above i.e $e_i \geq 5$
- The chi-square test does not use raw data or measurements its only observed frequencies are used.

Example

A die is tossed 120 times and a number of times a given face appears was recorded and the results are shown below

Face: 1 , 2, 3, 4, 5, 6

Frequency: 17,18,24,26,21,14

can we conclude at 0.05 level of significance that the die is fair.

solution

1. interest is on the fairness of the die
2. $H_0 : P(X = x) = \frac{1}{6}, x = 1, 2, ..., 6$
3. $H_1 : P(X = x) \neq \frac{1}{6}$ for atleast one value of x
4. $\alpha = 0.05$
5. test statistic $\chi^2 = \sum_{i=1}^{6} \frac{(\theta_i - e_i)^2}{e_i}$

6. critical region $\chi^2 > \chi^2_{0.05,6-1} = \chi^2_{0.05,5} = 11.07$

face	θ_i	$e_i = np_i$
1	17	20
2	18	20
3	24	20
4	26	20
5	21	20
6	14	20

$$\chi^2 = \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(26-20)^2}{20} + \frac{(21-20)^2}{20} + \frac{(14-20)^2}{20}$$

$$= \frac{1}{20}(9 + 4 + 16 + 36 + 1 + 30)$$

$$= \frac{102}{20} = 5.1$$

since $5.1 < 11.07$

we fail to reject the null hypothesis. The die is balanced

Note

- The distribution of a test statistic is approximately a chi-square to ensure that approximation is adequate we need to have $e_i \geq 5$ for all$i$
- If the expected frequency in any category is less than 5, that category should be combined with one or more of the neighbouring category
- If the null hypothesis depends on unknown parameter estimated from the data. a one degree is deducted for each parameter obtained from the data.
- Thus the degree of freedom catering for situation above will have $m - p - 1$ where p is the number of parameter to be estimated and m is the number of summations to be made

Example

The number of errors in a money script with 440 pages are shown below

number of errors	frequency
0	18
1	53
2	103
3	107
4	82
5	46
6	18
7	10
8	2
9	1

Test at 0.05 level of significance that the number of errors follow a poison distribution

solution

can the data be modelled by a poison distribution

H_0 number of errors follow a poison distribution

H_1 number of errors does not follow a poison distribution

$\alpha = 0.05$

from sample data the mean number of errors

$$\hat{\lambda} = \frac{1341}{440} = 3.04$$

$$\hat{\lambda} = 3$$

$$E_i = np_i$$

where p_i is the probability of having i errors and n- total frequency.

$$p_i = \frac{e^{-\lambda}\lambda^i}{i!}, i = 0, 1, 2, ...9$$

with $\lambda = 3$

error	O_i	p_i	e_i
0	18	0.0498	21.9
1	53	0.1494	65.7
2	103	0.2240	98.6
3	107	0.2240	98.6
4	82	0.680	73.9
5	46	0.1008	44.4
6	18	0.0504	22.2
7	10	0.0216	9.5
8	2	0.0082	3.6
9	1	0.0380	1.2

critical value $\chi^2 > \chi^2_{0.05}$

$$m = 8$$

because the last 3 is combined to make 1 row

$$\chi^2 > \chi^2_{0.05,8-1-1} = \chi^2_{0.05,6} = 12.59$$

$$\chi^2 = \frac{(18-21.9)^2}{21.9} + \frac{(53-65.7)^2}{65.7} +$$

$$\frac{(103-98.6)^2}{98.6} + \frac{(107-98.6)^2}{98.6} + \frac{(82-73.9)^2}{73.9} +$$

$$\frac{(46-44.4)^2}{44.4} + \frac{(18-22.2)^2}{22.2} + \frac{(13-14.3)^2}{14.3}$$

$$\chi^2 = 5.28$$

since $5.28 < 12.59$

we fail to reject the null hypothesis. a number of errors follow a poison distribution.

Example

A botanist claims that the number of seeds germinating in a group of 4 seeds follows a binomial distribution with $p = 0.6$. From 100 experiments, the following data was collected:

germinate (x)	0	1	2	3	4
frequency	2	10	28	40	20

Test the hypothesis that the germination follows the claimed binomial distribution at the 5% significance level. Use the Chi-Square goodness-of-fit test.

- Number of trials per experiment: $n = 4$
- Probability of success (germination): $p = 0.6$
- Number of experiments: $N = 100$

Observed Frequencies:

Germinated x	0	1	2	3	4
Observed O_i	2	10	28	40	20

The binomial probability formula is:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Calculate probabilities:

$$P(0) = \binom{4}{0}(0.6)^0(0.4)^4 = 1 \cdot 1 \cdot 0.0256 = 0.0256$$

$$P(1) = \binom{4}{1}(0.6)^1(0.4)^3 = 4 \cdot 0.6 \cdot 0.064 = 0.1536$$

$$P(2) = \binom{4}{2}(0.6)^2(0.4)^2 = 6 \cdot 0.36 \cdot 0.16 = 0.3456$$

$$P(3) = \binom{4}{3}(0.6)^3(0.4)^1 = 4 \cdot 0.216 \cdot 0.4 = 0.3456$$

$$P(4) = \binom{4}{4}(0.6)^4(0.4)^0 = 1 \cdot 0.1296 \cdot 1 = 0.1296$$

Multiply each probability by $N = 100$ to get expected frequencies:

x	$P(X = x)$	$E_i = 100 \cdot P(X = x)$
0	0.0256	2.56
1	0.1536	15.36
2	0.3456	34.56
3	0.3456	34.56
4	0.1296	12.96

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

x	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
0	2	2.56	$\frac{(2-2.56)^2}{2.56} = 0.1225$
1	10	15.36	$\frac{(10-15.36)^2}{15.36} = 1.8701$
2	28	34.56	$\frac{(28-34.56)^2}{34.56} = 1.2460$
3	40	34.56	$\frac{(40-34.56)^2}{34.56} = 0.8583$
4	20	12.96	$\frac{(20-12.96)^2}{12.96} = 3.8037$

$$\chi^2 = 0.1225 + 1.8701 + 1.2460 + 0.8583 + 3.8037 = \boxed{7.9006}$$

- Degrees of freedom:
  df = number of categories $-1-$ parameters estimated $= 5-1-0 = 4$
- Significance level: $\alpha = 0.05$
- Critical value $\chi^2_{0.05,4} = 9.488$

Since

$$\chi^2_{\text{calculated}} = 7.9006 < \chi^2_{\text{critical}} = 9.488$$

we **fail to reject the null hypothesis**.
There is no significant difference between the observed and expected frequencies. The binomial distribution with $p = 0.6$ is a good fit for the data at the 5% significance level.

Example

A manufacturer of light bulbs would like to determine whether the life of he bulb is normally distributed. a random sample of 100 bulbs is selected and tested, their life times to failure being measured in hours are given below

Range	frequency
< 850	3
850-900	4
900-950	8
950-1000	12
1000-1050	16
1050-1100	28
1100-1150	24
>1150	5

Test at 0.05 level of significance that the life of he bulb are normally distributed.

We are testing if the lifetimes of bulbs are normally distributed. The sample size is $n = 100$. Observed frequencies:
We approximate class midpoints as:

$$\text{Midpoints: } [825, 875, 925, 975, 1025, 1075, 1125, 1175]$$

Using the formula for the mean:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{104450}{100} = 1044.5 \text{ hours}$$

Using the formula for standard deviation:

$$s = \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i}} \approx 83.93 \text{ hours}$$

We assume a normal distribution with $\mu = 1044.5$, $\sigma = 83.93$. For each class, we calculate:

$$E_i = 100 \cdot P(X < \mathit{limit})$$

for example

$$z = \frac{850 - 1044.5}{83.93} = \frac{-194.5}{83.93} = -2.317$$

$$\phi(-2.317) = 0.0102$$

$$z = \frac{900 - 1044.5}{83.93} = -1.722$$

$$\phi(-1.722) = 0.0425$$

$$z = \frac{950 - 1044.5}{83.93} = -1.126$$

$$\phi(-1.126) = 0.1301$$

$$z = \frac{1000 - 1044.5}{83.93} = -0.530$$

$$\phi(-0.530) = 0.2981$$

$$z = \frac{1050 - 1044.5}{83.93} = 0.066$$

$$\phi(0.066) = 0.5263$$

$$z = \frac{1100 - 1044.5}{83.93} = 0.661$$

$$\phi(0.661) = 0.7457$$

$$z = \frac{1150 - 1044.5}{83.93} = 1.257$$

$$\phi(1.257) = 0.8956$$

$$P(x > 1150) = \phi(1.257) = 0.1044$$

Class (x)	Observed O_i	$\phi(z_b) - \phi(z_a)$	Expected E_i
$<$850	3	0.0102	1.02
850–900	4	0.0323	3.23
900–950	8	0.0876	8.76
950–1000	12	0.1680	16.80
1000–1050	16	0.2282	22.82
1050–1100	28	0.2194	21.94
1100–1150	24	0.1499	14.99
$>$1150	5	0.1044	10.44

The test statistic is:

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

Computing each term and summing:

$$\chi^2 \approx 17.43$$

$$\text{df} = k - p - 1 = 8 - 2 - 1 = 5$$

Where:

- $k = 8$: number of classes
- $p = 2$: parameters estimated (mean and standard deviation)

At $\alpha = 0.05$, the critical value is:

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

Since:

$$\chi^2 = 17.43 > 11.07 \quad \Rightarrow \quad \text{Reject } H_0$$

Or using the p-value:

$$p = 0.0038 < 0.05 \quad \Rightarrow \quad \text{Reject } H_0$$

There is sufficient evidence at the 5% significance level to reject the hypothesis that the lifetimes of the bulbs follow a normal distribution.

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: Chi-square test