Probability and Statistics

Chi-squared Test

HCMC University of Technology

Dung Nguyen

Outline I

1. Testing for Goodness of Fit

2. Contingency Table Tests

1 Testing for Goodness of Fit

Roll a die 99 times

Outcome	1	2	3	4	5	6
Observed frequency	15	12	15	14	20	23

Is this die fair at $\alpha = 0.01$?

Let X be the outcome of rolling the die and $p_i = P(X = i)$

We want to test the following hypotheses

$$\begin{aligned} H_0 : p_1 = 1/6, \quad p_2 = 1/6, \quad p_3 = 1/6, \\ p_4 = 1/6, \quad p_5 = 1/6, \quad p_6 = 1/6 \\ \text{vs.} \quad H_1 : \exists i : p_i \neq 1/6. \end{aligned}$$

Assume that H_0 is true.

Outcome	1	2	3	4	5	6
Observed frequency	15	12	15	14	20	23
$P(X = i)$	1/6	1/6	1/6	1/6	1/6	1/6
Expected frequency	16.5	16.5	16.5	16.5	16.5	16.5

Then

$$\begin{aligned} X^2 = & \frac{(15-16.5)^2}{16.5} + \frac{(12-16.5)^2}{16.5} + \frac{(15-16.5)^2}{16.5} \\ & + \frac{(14-16.5)^2}{16.5} + \frac{(20-16.5)^2}{16.5} + \frac{(23-16.5)^2}{16.5} = 5.18. \end{aligned}$$

$$X^2 = \frac{(15-16.5)^2}{16.5} + \frac{(12-16.5)^2}{16.5} + \frac{(15-16.5)^2}{16.5} + \frac{(14-16.5)^2}{16.5} + \frac{(20-16.5)^2}{16.5} + \frac{(23-16.5)^2}{16.5} = 5.18$$

$$X^2 = \frac{15^2}{16.5} + \frac{12^2}{16.5} + \frac{15^2}{16.5} + \frac{14^2}{16.5} + \frac{20^2}{16.5} + \frac{23^2}{16.5} - 99 = 5.18$$

The degree of freedom: $df = 6 - 1 = 5$. The threshold value: χ

Goodness of fit tests

Pearson's article (1900): establish the asymptotic chi-square distribution for a goodness of fit statistic for the multinomial distribution.

- The test is based on the chi-square distribution.
- A sample of size n from a population whose probability distribution is unknown.
- O_j = the observed frequency in the j-th class interval.
- E_j = the expected frequency in the j-th class interval.

test statistic

The test statistic is

$$\chi^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}$$

or

$$\chi^2 = \sum_{j=1}^{m} \frac{(\nu_j - np_j)^2}{np_j}$$

Also

$$\chi^2 = \sum_{j=1}^{m} \frac{O_j^2}{E_j} - n$$

or

$$\chi^2 = \sum_{j=1}^{m} \frac{\nu_j^2}{np_j} - n$$

Example 1 – Printed Circuit Boards

The number of defects in printed circuit boards is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ printed boards has been collected, and the following number of defects observed.

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

Question 1: Does the data set follow Poisson distribution with $\lambda = 0.5$?

Question 2: Does the data set follow a Poisson distribution? (λ is unknown)

Solution

Question 1: Does the data set follow Poisson distribution with $\lambda = 0.5$?

Y: the number of defects on a circuit board.

$H_0 : Y \sim \text{Poisson}(0.5)$.

First of all, we assume that H_0 is true. Then

$$P(Y = k) = e^{-0.5}\frac{0.5^k}{k!}$$

$$P(Y = 0) = e^{-0.5}\frac{0.5^0}{0!} = 0.6065$$

$$P(Y = 1) = e^{-0.5}\frac{0.5^1}{1!} = 0.3033$$

$$P(Y = 2) = e^{-0.5}\frac{0.5^2}{2!} = 0.0758$$

$$P(Y = 3) = e^{-0.5}\frac{0.5^3}{3!} = 0.0126$$

$$P(Y \geq 3) = 1 - (0.6065 + 0.3033 + 0.0758) = 0.0144.$$

Solution

Number of Defects	Observed Frequency	Probability	Expected Frequency
0	32	0.6065	36.3918
1	15	0.3033	18.1959
2	9	0.0758	4.5490
≥ 3	4	0.0144	0.8633
Total	60	1	60

$$\chi^2 = \frac{(32-36.3918)^2}{36.3918} + \frac{(15-18.1959)^2}{18.1959} + \frac{(9-4.5490)^2}{4.5490} + \frac{(4-0.8633)^2}{0.8633} = 16.8442$$

$$= \frac{32^2}{36.3918} + \frac{15^2}{18.1959} + \frac{9^2}{4.5490} + \frac{4^2}{0.8633} - 60 = 16.8442$$

$$\chi^2_{0.05,3} = 7.81$$

Solution

Question 2: Does the data set follow a Poisson distribution? (λ is unknown)

$$\lambda = \frac{32 \times 0 + 15 \times 1 + 9 \times 2 + 4 \times 3}{32 + 15 + 9 + 4} = 0.75$$

# of Defects	Observed freq.	Prob.	Expected freq.
0	32	0.4724	28.3420
1	15	0.3543	21.2565
2	9	0.1329	7.9711
≥ 3	4	0.0405	2.4303
Total	60	1	60

Solution

# of Defects	Observed freq.	Prob.	Expected freq.
0	32	0.4724	28.3420
1	15	0.3543	21.2565
2	9	0.1329	7.9711
≥ 3	4	0.0405	2.4303
Total	60	1	60

$$\chi^2 = \frac{(32-28.32)^2}{28.32} + \frac{(15-21.24)^2}{21.24} + \frac{(9-7.98)^2}{7.98} + \frac{(4-2.46)^2}{2.46} = 3.4602$$

$$= \frac{32^2}{28.32} + \frac{15^2}{21.24} + \frac{9^2}{7.98} + \frac{4^2}{2.46} - 60 = 3.4602$$

$$\chi^2_{0.05,2} = 5.99$$

Solution

Number of Defects	Observed Frequency	Probability	Expected Frequency
0	32	0.472	28.32
1	15	0.354	21.24
≥ 2	13	0.174	10.44
Total	60	1	

$$\begin{aligned}\chi^2 &= \frac{(32-28.32)^2}{28.32} + \frac{(15-21.24)^2}{21.24} + \frac{(13-10.44)^2}{10.44} \\ &= \frac{32^2}{28.32} + \frac{15^2}{21.24} + \frac{13^2}{10.44} - 60 = 2.94\end{aligned}$$

$\chi^2_{0.05,1} = 3.$

Question: Does the data set follow a normal distribution?

2 Contingency Table Tests

Contingency Table

- The n elements of a sample from a population may be classified according to two different criteria.
- Question: Are the two methods of classification statistically independent?

	Method 2				Total
Method 1	N_{11}	N_{12}	$\dots$	N_{1J}	$n_{1.}$
	N_{21}	N_{22}	$\dots$	N_{2J}	$n_{2.}$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	N_{I1}	N_{I2}	$\dots$	N_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.J}$	n

Example 2 – Job Classification

A company has to choose among three health insurance plans. Management wishes to know whether the preference for plans is independent of job classification and wants to use $\alpha = 0.05$.

The opinions of a random sample of 500 employees are shown

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	40	60	60	160
Totals	200	200	100	500

Test procedure I

- Goal: Test the hypothesis that the row-and-column methods of classification are independent.
- Reject this hypothesis = there is some interaction between the two criteria of classification.
- The exact test procedures are difficult to obtain, but an approximate test statistic is valid for large n.

Example 3 – Job Classification

A company has to choose among three health insurance plans. Management wishes to know whether the preference for plans is independent of job classification and wants to use $\alpha = 0.05$.

The opinions of a random sample of 500 employees are shown

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	40	60	60	160
Totals	200	200	100	500

Solution

H_0: Job Classification and Choice of health Insurance Plan are independent.
First we assume that H_0 is true.

Job Classification	Health Insurance Plan 1	2	3	Totals
Salaried workers	x_1	x_2	x_3	340
Hourly workers	x_4	x_5	x_6	160
Totals	200	200	100	500

$$\begin{cases} x_1 + x_2 + x_3 = 340 \\ x_4 + x_5 + x_6 = 160 \\ x_1 + x_4 = 200 \\ x_2 + x_5 = 200 \\ x_3 + x_6 = 100 \end{cases} \Longrightarrow A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$\text{rank}(A) = 4 \Longrightarrow$ the degree of freedom = the # of independent variables = $6 - 4 = 2$.

Wait a minute ...

Wait a minute ...

Do we have to compute the rank of a matrix to determine the df?

Wait a minute ...

Do we have to compute the rank of a matrix to determine the df?

$$df = (2-1)(3-1) = 2$$

Solution

Expected values

Job Class.	Health Insurance Plan 1	2	3	Tot.
Salaried	$\frac{200\times 340}{500}=136$	$\frac{200\times 340}{500}=136$	$\frac{100\times 340}{500}=68$	340
Hourly	$\frac{200\times 160}{500}=64$	$\frac{200\times 160}{500}=64$	$\frac{100\times 160}{500}=32$	160
Totals	200	200	100	500

Job Classification	Health Insurance Plan 1	2	3	Totals
Salaried workers	160, 136	140, 136	40, 68	340
Hourly workers	40, 64	60, 64	60, 32	160
Totals	200	200	100	500

Solution

Job Classification	Health Insurance Plan 1	2	3	Totals
Salaried workers	160, 136	140, 136	40, 68	340
Hourly workers	40, 64	60, 64	60, 32	160
Totals	200	200	100	500

$$\chi^2 = \frac{(160-136)^2}{136} + \cdots + \frac{(60-32)^2}{32} = 49.6$$

$$\chi^2 = 500\left[\frac{160^2}{200 \times 340} + \cdots + \frac{60^2}{100 \times 160} - 1\right] = 49.6$$

$$\chi^2_{0.05,2} = 5.99$$

Test procedure I

- Goal: Test the hypothesis that the row-and-column methods of classification are independent.
- Reject this hypothesis = there is some interaction between the two criteria of classification.
- The exact test procedures: difficult to obtain
- An approximate test statistic: Chi-squared test for large n.

Test procedure II

- Let p_{ij} be the probability that a randomly selected element falls in the ij-th cell.
- If the two classifications are independent then $p_{ij} = u_i v_j$, where
  - u_i is the probability that a randomly selected element falls in row class i and
  - v_j is the probability that a randomly selected element falls in column class j.
- Assuming independence, the estimators of u_i and v_j are

$$u_i = \frac{n_{i.}}{n} \quad \text{and} \quad v_i = \frac{n_{.j}}{n}$$

- The expected frequency of each cell is

$$E_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Statistic for Contingency Table Test

For large n, the statistic

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

has an approximate chi-square distribution with $(I-1)(J-1)$ degrees of freedom if the null hypothesis is true.

We should reject the null hypothesis if the value of the test statistic χ^2 is too large.

Computational formula:

$$\chi^2 = n\left[\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right]$$

Example

Grades in a statistics course and an operation research course take simultaneously were as follows for a group of students

Statistics Grade	Operation Research Grade			
	A	B	C	Others
A	25	6	17	13
B	17	16	15	6
C	18	4	18	10
Others	10	8	11	20

Are the grades in two courses related? Use $\alpha = 0.01$ in reaching your conclusion. What is the $\mathbf{p_v}$?