Goodness of fit tests

The 'goodness of fit' describes how well a statistical model or distribution fits a set of data.

Fitting a distribution

Let a sample of size n be presented as follows:

x	Observed Frequency
x_1	O_1
x_2	O_2
$\vdots$	$\vdots$
x_k	O_k

We want to see whether the following distribution fits the above data:

x	Probability
x_1	p_1
x_2	p_2
$\vdots$	$\vdots$
x_k	p_k

H_0: The distribution fits the data.

H_1: The distribution does not fit the data.

If H_0 is true, then the 'expected frequencies' are

$E_i = np_i, \qquad i = 1, 2, \cdots, k.$

We check whether the observed frequencies and the expected frequencies are close by calculating the following test-statistic

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

which follows chi-square distribution with $k - 1$ d.f. under the null hypothesis. The d.f. is one less because of the restriction $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n$.

Example

The data below present the outcomes of a dice that has been thrown 1000 times. Is the dice fair?

x	Frequency
1	140
2	180
3	150
4	180
5	160
6	190

Solution

x	Observed Frequency	Probability	Expected Frequency
1	140	1/6	166.7
2	180	1/6	166.7
3	150	1/6	166.7
4	180	1/6	166.7
5	160	1/6	166.7
6	190	1/6	166.7

$\chi^2 = 11.60$

Always upper tail test

$\chi^2_{0.95,5} = 11.07$

We reject the null and conclude that the dice is not fair.

Goodness of fit when some parameters are unspecified

Example

Suppose the weekly number of accidents over a 30-week period is as follows:

8, 0, 0, 1, 3, 4, 0, 2, 12, 5, 1, 8, 0, 2, 0, 1, 9, 3, 4, 5, 3, 3, 4, 7, 2, 0, 1, 2, 1, 2

Test the hypothesis that the number of accidents in a week has a Poisson distribution.

Solution

Here, the parameter of Poisson distribution, λ, is not specified. We have to estimate it from the data.

$\hat{\lambda} = \bar{X} = 3.1$

x	Observed Frequency	Probability	Expected Frequency
0	6	0.045	1.35
1	5	0.140	4.19
2	5	0.216	6.49
3	4	0.224	6.71
≥ 4	10	0.375	11.25

$\chi^2 = 17.72$

$\chi^2_{0.95,3} = 7.81$

We reject the null and conclude that the data do not follow Poisson distribution.

Here, the d.f. is $5 - 2 = 3$ instead of $5 - 1 = 4$, because estimation of one parameter leads to loss of one d.f.

Test of independence in contingency tables

A contingency table is a bivariate frequency table also known as cross-tab. When we want to check the association between two categorical variables, we construct a contingency table. We may then perform a chi-square test of independence.

Example

The following table shows 100 persons classified according to gender and colorblindness. Is there association between gender and colorblindness?

H_0: There is no association between gender and colorblindness.

H_1: There is association between gender and colorblindness.

Level of significance = 0.05

Gender	Colorblind		Total
	Y	N	
M	10 (8)	30 (32)	40
F	10 (12)	50 (48)	60
Total	20	80	100

The table shows observed frequencies. The values in the parentheses are expected frequencies (expected when there is no association). When there are r rows and c columns in the contingency table, the test-statistic is

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

which follows chi-square distribution with $(r-1)(c-1)$ d.f. under the null hypothesis.

What is your conclusion based on the table above?