People's Democratic Republic of Algeria
University Med Khider of Biskra
Faculty of SNVSTU

Test of Independence

Dr. Ben Gherbal Hanane

Lecture: Data Analysis in Biosciences — Level: L3
Email: hanane.bengherbal@univ-biskra.dz

The goal is to measure two qualitative random variables in a population and determine whether these variables are independent, that is, whether knowing the value of one variable may influence the probability distribution of the other.

We say that there is independence between two random variables X and Y if

$$P((X = x) \cap (Y = y)) = P(X = x) \times P(Y = y).$$

Using this formulation of independence, we can write the hypotheses:

$$H_0 : X \text{ and } Y \text{ are independent.}$$

$$H_1 : X \text{ and } Y \text{ are dependent.}$$

We will use the Chi-square statistic to perform the independence test. To understand the steps of application of the test, we consider the following example.

Example

To target the customers of a new consumer product, a company conducts a survey of 321 people. The interest in the product is recorded as "no interest", "minor interest", or "major interest". The family situation (at least one dependent child: yes or no) is also recorded.

The goal is to verify whether interest in the product depends on family situation. The results are as follows:

Child	None	Minor	Major
Yes	10	12	3
No	7	38	9

A total of 79 people responded. We want to check whether there is a relationship between the two variables at the 5% significance level.

Solution

We test the hypotheses:

H_0 : independence between family situation and interest in the product.

H_1 : dependence between family situation and interest in the product.

The significance level is fixed at 5%, meaning that the probability of concluding dependence when the variables are actually independent is 5%.

The test consists of rejecting H_0 if

$$\chi^2 = \sum\sum \frac{(n_{ij} - t_{ij})^2}{t_{ij}} \geq \chi^2_{(m-1)(k-1),\alpha}.$$

We obtain the following contingency table:

	None	Minor	Major	Total
Yes	10	12	3	25
No	7	38	9	54
Total	17	50	12	79

Next, the table of theoretical (expected) frequencies:

t_{ij}	None	Minor	Major	$n_{i.}$
Yes	5.40	15.823	3.800	25
No	11.620	34.177	8.203	54
$n_{.j}$	17	50	12	79

with

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

the expected frequency for categories i and j under independence.

Conditions of Application

This approximate test is valid if:

1. $t_{ij} \geq 1$ for all i, j,
2. No more than 20% of the t_{ij} values are below 5.

There is one cell out of six with an expected value less than 5, which corresponds to:

$$\frac{1}{6} \times 100 = 16.667\% < 20\%.$$

So the test is valid.

Observed Statistic

$$\chi^2 = \sum\sum \frac{(n_{ij} - t_{ij})^2}{t_{ij}} = \frac{(10 - 5.40)^2}{5.40} + \ldots + \frac{(9 - 8.203)^2}{8.203} = 7.401.$$

From the Chi-square table:

$$\chi^2_{(m-1)(k-1);0.05} = \chi^2_{2;0.05} = 5.99,$$

since $(m-1)(k-1) = 2 \times 1 = 2$.

Because the observed value 7.401 is greater than the critical value 5.99, we reject H_0.

Conclusion

There is a significant relationship between family situation and interest in the product.

Exercise 5

We want to know whether the success R of a medical treatment is independent of the patient's blood pressure level T. A sample of 250 observations is given below:

$T \backslash R$	Failure	Success
Low	21	104
High	29	96

At the significance level $\alpha = 0.05$, does the success of the treatment depend on the blood pressure level?

Solution

Step 1: Contingency table with totals..

$T \backslash R$	Failure	Success	Row total
Low	21	104	125
High	29	96	125
Column total	50	200	250

Step 2: Hypotheses..

$$H_0 : R \text{ and } T \text{ are independent.}$$

$$H_1 : R \text{ and } T \text{ are dependent.}$$

Step 3: Expected frequencies under independence..

$$t_{ij} = \frac{(n_{i.})(n_{.j})}{n}$$

$$t_{11} = \frac{125 \times 50}{250} = 25, \qquad t_{12} = \frac{125 \times 200}{250} = 100,$$
$$t_{21} = 25, \qquad t_{22} = 100.$$

Expected table:

$T \backslash R$	Failure	Success
Low	25	100
High	25	100

All expected values are ≥ 5, so the chi-square test is valid.

Step 4: Chi-square statistic..

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

$$\chi^2 = \frac{(21-25)^2}{25} + \frac{(104-100)^2}{100} + \frac{(29-25)^2}{25} + \frac{(96-100)^2}{100}.$$

$$\chi^2 = \frac{16}{25} + \frac{16}{100} + \frac{16}{25} + \frac{16}{100} = 0.64 + 0.16 + 0.64 + 0.16 = 1.60.$$

Degrees of freedom:

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

Critical value at $\alpha = 0.05$:

$$\chi^2_{0.05,1} = 3.84.$$

Step 5: Decision.. Since

$$\chi^2_{\text{obs}} = 1.60 < 3.84,$$

we do **not** reject H_0.

Conclusion.. At the 5% significance level, there is no evidence that treatment success depends on blood pressure level. The variables appear to be independent.