Chi-squared test for independence of qualitative variables

The data for these tests are contingency tables showing the relationship between 2 qualitative variables. For example, suppose we have the following information regarding hair and eye colour

	Red hair	Blonde hair	Dark hair	$\sum$
Blue eyes	30	90	30	150
Brown eyes	20	60	70	150
$\sum$	50	150	100	300

i.e. 30 people have red hair and brown eyes.

Test for independence

Let $n_{i,j}$ be the entry in the i-th row and j-th column of the contingency table. We wish to choose between the hypotheses

$$\begin{aligned} H_0 &: \text{ hair colour and eye colour are independent.} \\ H_A &: \text{ hair and eye colour are dependent.} \end{aligned}$$

Row and column sums

The number of people in the sample with blue eyes is the sum of the entries in the first row (150).

The number of people in the sample with brown eyes is the sum of the entries in the second row (150).

The sum of all the entries is the number of individuals in the sample (300).

Probabilities of given events

The probability that an individual in the sample has blue eyes, P(blue), is the number of people with blue eyes divided by the number of people in the sample.

$$\text{i.e. } P(\text{blue}) = \frac{150}{300} = \frac{1}{2}$$

Similarly,

$$P(\text{brown}) = \frac{150}{300} = \frac{1}{2}$$

Probabilities of given events

The number of individuals with red hair is the sum of the number of entries in the first column (50). Arguing as above the probability the an individual in the sample has red hair is

$$\mathrm{P}(\text{red}) = \tfrac{50}{300} = \tfrac{1}{6}$$

In a similar way,

$$\mathrm{P}(\text{blond}) = \tfrac{150}{300} = \tfrac{1}{2}$$
$$\mathrm{P}(\text{dark}) = \tfrac{100}{300} = \tfrac{1}{3}$$

Probabilities under the hypothesis of independence

If the traits are independent, then the probability that an individual has a given hair colour and given eye colour is the product of the two corresponding probabilities e.g.

$$\mathrm{P(blond\ hair,\ blue\ eyes)} = \mathrm{P(blond\ hair)P(blue\ eyes)}$$

In order to test whether two traits are independent, we need to calculate what we would expect to observe if the traits were independent.

The following calculations allow us to calculate what we expect to see under the null hypothesis of independence.

Probabilities under the hypothesis of independence

In general, let $n_{i,\bullet}$ be the sum of the entries in the i-th row and $n_{\bullet,j}$ be the sum of the entries in the j-th column.

Let n be the total number of observations (the sum of the entries in the cells).

Hence, here $n_{1,\bullet} = n_{2,\bullet} = 150$ (the sums of entries in first and second rows, respectively).

Also, $n_{\bullet,1} = 50, n_{\bullet,2} = 150, n_{\bullet,3} = 100$ (sum of entries in columns 1,2 and 3, respectively).

n=300 (i.e. there are 300 observations in total)

Probabilities under the hypothesis of independence

The probability of an entry being in the i-th row is $p_{i,\bullet} = \frac{n_{i,\bullet}}{n}$.

The probability of an entry being in the j-th column is $p_{\bullet,j} = \frac{n_{\bullet,j}}{n}$.

If the two traits considered are independent, then the probability of an entry being in the cell in the i-th row and j-th column is $p_{i,j}$, where

$$p_{i,j} = p_{i,\bullet} p_{\bullet,j} = \frac{n_{i,\bullet}}{n} \frac{n_{\bullet,j}}{n} = \frac{n_{i,\bullet} n_{\bullet,j}}{n^2}$$

Expected number of observations in a cell under H_0

Under the null hypothesis, we expect $e_{i,j}$ observations in cell (i,j), where

$$e_{i,j} = np_{i,j}$$

Hence, we can calculate the expected number of observations in each cell

$$e_{i,j} = np_{i,j} = \frac{n_{i,\bullet} n_{\bullet,j}}{n},$$

i.e. the expected number is the row sum times the column sum divided by the total number of observations.

Table of expected values

In the example considered the calculation is as follows:

	Red hair	Blond hair	Dark hair	$\sum$
Blue eyes	$\frac{150\times 50}{300} = 25$	$\frac{150\times 150}{300} = 75$	$\frac{150\times 100}{300} = 50$	150
Brown eyes	$\frac{150\times 50}{300} = 25$	$\frac{150\times 150}{300} = 75$	$\frac{150\times 100}{300} = 50$	150
$\sum$	50	150	100	300

Comparison of observations and expectations (expectations in brackets)

	Red hair	Blond hair	Dark hair	$\sum$
Blue eyes	30 (25)	90 (75)	30 (50)	150
Brown eyes	20 (25)	60 (75)	70 (50)	150
$\sum$	50	150	100	300

It should be noted that the sum of the expectations in a row is equal to the sum of the observations. An analogous result holds for the columns.

The test statistic

The test statistic is

$$T = \sum_{i,j} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}},$$

where the summation is carried out over all the cells of the contingency table.

The realisation of this statistic is labelled t. This is a measure of the distance of our observations from those we expect under H_0.

It should be noted that if the null hypothesis is true, then $n_{i,j}$ and $e_{i,j}$ are likely to be similar, hence the realisation t will tend to be close to 0 (by definition this realisation is non-negative). Large values of t indicate that the traits are dependent.

Calculation of the realisation of the test statistic

In this case,

$$t = \frac{(30-25)^2}{25} + \frac{(90-75)^2}{75} + \frac{(30-50)^2}{50} + \dots$$
$$+ \frac{(20-25)^2}{25} + \frac{(60-75)^2}{75} + \frac{(70-50)^2}{50} = 24.$$

Distribution of the test statistic under H_0

Given the traits are independent, the test statistic has an approximate chi-squared distribution with $(r-1)\times(c-1)$ degrees of freedom, where r and c are the numbers of rows and columns, respectively.

It should be noted that this approximation is **reasonable if at least 5 observations are expected in each cell under H_0.**

Making conclusions

Since large values of t indicate that the traits are dependent, we reject the null hypothesis of independence at a significance level of α if

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

where $\chi^2_{(r-1)(c-1),\alpha}$ is the critical value for the appropriate chi-squared distribution. These values can be read from Table 8.

Making conclusions

In this case at a significance level of 0.1%

$$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{2,0.001} = 13.815.$$

Since $t = 24 > \chi^2_{2,0.001} = 13.815$, we reject the null hypothesis of independence.

Since we rejected at a significance level of 0.1%, we have very strong evidence that eye and hair colour are dependent.

Describing the nature of an association

In order to see the nature of the association between hair and eye colour, we should compare the observed and expected values.

Comparing the table of observed and expected values, it can be seen that dark hair is associated with brown eyes and blond hair with blue eyes (in both cases there are more observations than expected under the null hypothesis).

Simpsons Paradox

The nature of an association between two traits may change and even reverse direction when the data from several groups is combined into a single group.

For example, consider the following data regarding admissions to a university.

The Physics department accepted 60 applications from 100 males and 20 applications from 30 females.

The Fine Arts department accepted 10 applications from 100 males and 30 applications from 170 females.

Simpson's paradox

Hence, the Physics department accepted 60% of applications from males and 66.7% of applications from females. The Fine Arts department accepted 10% of applications from males and 17.6% of applications from females.

Hence, in both departments females are slightly more successful.

Simpson's paradox

Combining the departments, we obtain the following contingency table

	Accepted	Rejected	$\sum$
Male	70	130	200
Female	50	150	200
$\sum$	120	280	400

Combining the results females are less successful.

We now test the null hypothesis that acceptance is independent of sex. The alternative is that acceptance is associated with sex (i.e. the likelihood of acceptance depends on sex).

Simpson's paradox

The table of expectations is as follows

	Accepted	Rejected	$\sum$
Male	$\frac{200\times120}{400} = 60$	200-60=140	200
Female	120-60=60	280-140=140	200
$\sum$	120	280	400

Comparing the table of observations and expectations

	Accepted	Rejected	$\sum$
Male	70 (60)	130 (140)	200
Female	50 (60)	150 (140)	200
$\sum$	120	280	400

Simpson's paradox

The test statistic is

$$T = \sum_{i,j} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}}$$

The realisation is given by

$$t = \frac{(70-60)^2}{60} + \frac{(130-140)^2}{140} + \frac{(50-60)^2}{60} + \frac{(150-140)^2}{140} \approx 4.76$$

Simpson's paradox

Testing the null hypothesis at the 5% level, the critical value is

$$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{1,0.05} = 3.841.$$

Since $t > \chi^2_{1,0.05} = 3.841$, we reject the null hypothesis.

We conclude that acceptance is associated with sex. Looking at the table which compares the observed and expected values, females are less likely to be accepted than expected.

Hence, using this data we might conclude that there is evidence that females are discriminated against.

Simpson's paradox

However, in both departments a female is more likely to be accepted.

On average, females are less likely to be accepted since they are more likely to apply to the fine arts department and the fine arts department rejects a higher proportion of applications.

Lurking variables

In this case department is what we call a hidden or lurking variable.

A hidden or lurking variable is a variable which may influence the observed results, but is not considered in the analysis.

For example, women's salaries are significantly lower on average that men's salaries.

1. Can this be used as evidence that women are discriminated against?
2. What are the lurking variables which may explain such a difference?
3. How should we test whether females are discriminated against with regard to salary?