

# Chapter Two

---

## $\chi^2$ tests

---

The  $\chi^2$  test for independence is one of the most used and abused statistical tests. The  $\chi^2$  distribution is also a wonderfully useful distribution which turns up in many other areas. This chapter will consider three of the more common *hypothesis tests* based directly on the  $\chi^2$  distribution having a test statistic set out by Fisher (1904). Given its ubiquity, we will consider the  $\chi^2$  test for independence in the greatest detail. Because the  $\chi^2$  test for homogeneity is so closely related, we will briefly consider this test. Finally, whilst there are many ways of testing goodness of fit, we will consider the use of the  $\chi^2$  test for goodness of fit as it is based on a very similar test statistic to the other tests.

In common with many other statistical hypothesis tests, in each of these three tests we need to be very clear about:

- a. the precise nature of the null hypothesis  $H_0$  envisaged,
- b. the degrees of freedom  $\nu$  which apply to the test statistic,
- c. the distribution of the test statistic under  $H_0$  for a given number of degrees of freedom  $\nu$ .

All tests considered in this chapter have the property that the test statistic under  $H_0$  follows a  $\chi^2$  distribution with  $\nu$  degrees of freedom. In the case of the test for independence and the test for homogeneity, it turns out that our calculations for the test statistic are identical but the assumptions underlying them are quite different. In discussing these particular  $\chi^2$  tests, we need to attend to two small but important points:

- (i) the so-called Yates correction for  $2 \times 2$  contingency tables and
- (ii) dealing with cells where the expected count is less than 5.

Both of these points are important in limiting the potential for false positive results.

As covered in this chapter, the  $\chi^2$  test is a useful exploratory tool. There are a number of extensions to the tests we examine here, and in practice one may wish to follow up any interesting tables by other techniques such as log-linear modelling or correspondence analysis. This chapter will conclude with some pointers to further reading on the analysis of contingency tables.

## 2.1 Contingency Tables

In order to consider the  $\chi^2$  test for independence we will need to formalise some ideas about contingency tables. Essentially, we have collected a sample of data on one group of individuals, but have cross-classified these according to two categorical variables. These may either be truly categorical variables (site of tumour, type of tumour), or they could be “discretised” continuous variables (e.g. volume of tumour broken into a number of classes).

We’re going to look at a basic contingency table. The notation in this section follow ?). Contingency tables are formed when we cross-tabulate our data with reference to two qualitative variables; these are called pivot tables in Excel. To explain the notation we will use with respect to contingency tables we set out a prototypical blank contingency table below:

		Columns Variable 2				
		1	2	...	$c$	Total
Rows (variable 1)	1	$O_{11}$	$O_{12}$	...	$O_{1c}$	$O_{1.}$
	2	$O_{21}$	$O_{22}$	...	$O_{2c}$	$O_{2.}$
				...		
	r	$O_{r1}$	$O_{r2}$	...	$O_{rc}$	$O_{r.}$
Total		$O_{.1}$	$O_{.2}$	...	$O_{.c}$	$O_{..} = N$

To simplify things later, we use the “dot” notation to indicate summation. So, the row totals (called the marginal totals because of the way we may think of these being recorded in the margins of the table) are given by:

$$\begin{aligned} O_{i.} &= O_{i1} + O_{i2} + \dots + O_{ic} \\ &= \sum_{j=1}^c O_{ij}. \end{aligned}$$

In a similar manner, the (marginal) column totals can be denoted by:

$$\begin{aligned} O_{.j} &= O_{1j} + O_{2j} + \dots + O_{rj} \\ &= \sum_{i=1}^r O_{ij} \end{aligned}$$

and we could give the grand total as the sum of all cells, or the sum of the row totals or the sum of the column totals:

$$\begin{aligned} O_{..} &= \sum_{i=1}^r \sum_{j=1}^c O_{ij} \\ &= \sum_{i=1}^r O_{i.} = \sum_{j=1}^c O_{.j} = N, \end{aligned}$$

noting that we use  $N$  to indicate the grand total.

## Practical Illustration

To fix ideas, we start with a simple example from ?) where we have data on deaths from Tuberculosis (TB). We produce a contingency table showing the gender of the person dying from TB, and the type of TB causing death:

	Males	Females	Total
TB of respiratory system	3534	1319	4853
Other forms of TB	270	252	522
All forms of TB	3804	1571	5375

and we can see for these data that  $r = c = 2$  (i.e. we have a  $2 \times 2$  table), and for example that  $O_{11} = 3534$  and  $O_{1.} = 4853$ .

## 2.2 $\chi^2$ test for independence

The Null Hypothesis  $H_0$  in the  $\chi^2$  test for independence

Degrees of Freedom

Test statistic for independence in contingency tables

Yates continuity test for  $2 \times 2$  tables

## 2.3 The $\chi^2$ test for homogeneity

## 2.4 $\chi^2$ tests for Goodness of fit

Fitting a distribution to a sample of data

Goodness of fit for a sample of discrete data

Goodness of fit for a sample of continuous data

2.5 Further reading

2.6 Some notes on the  $\chi^2$  distribution  
(optional)

2.7 Summary