Chapter 5

Tests of independence using the χ^2 distribution

5.1 Introduction

In the previous chapter, we used a χ^2 test to test hypotheses about the probability distribution of a random variable. We are now going to see that the test is much more versatile because it can be used to compare the probability distributions of more than one population, and that this is equivalent to a test of the *association* between two *categorical variables*.

Suppose we are interested in the probabilities that a graduate has a permanent job, temporary job, or is unemployed six months after graduation, and we want to compare these probabilities for female graduates with those for male graduates. As we now have *two* categorical variables (employment status and gender), it will be worth looking at different ways of presenting and storing categorical data.

5.2 Presenting categorical data

The following table shows the first few rows of a computer file of data for a sample of graduates in *raw form*, that is, with a row for each graduate. The first column merely identifies the graduate, the second is their employment status six months after graduation (0 = permanent job, 1 = temporary job, 2 = unemployed) and the third is their gender (0 = male, 1 = female).

Case number	Employment status	Gender
1	2	0
2	0	0
3	1	1
4	0	1
5	1	1
6	0	0
7	0	1
⋮	⋮	⋮

Most software will store data in this form, although sometimes the categories don't have to be coded as numbers and you can enter the category name (e.g. male/female or permanent/temporary/unemployed) directly.

As employment status and gender are categorical with three and two categories respectively, only 3×2 possible pairs of numbers can appear in the second and third columns. The data can therefore be consolidated into *frequency form* using a column for each categorical variable and another for counts or frequencies, as shown:

Employment status	Gender	Frequency
0	0	100
0	1	90
1	0	33
1	1	40
2	0	25
2	1	22

The most succinct way of representing categorical data is to arrange the frequencies in a *contingency table*. When there are only two categories, this is easy; each column represents a category of one categorical variable, and each row represents a category of the other categorical variable. The frequencies are written in the appropriate cell of the table. Notice that each cell of the contingency table corresponds to one of the frequencies of the table above.

	Permanent	Temporary	Unemployed
Male	100	33	25
Female	90	40	22

It is often useful to include the column and row totals, so we repeat the table with these below:

	Permanent	Temporary	Unemployed	Total
Male	100	33	25	158
Female	90	40	22	152
Total	190	73	47	310

5.3 The χ^2 test for independence

In a χ^2 test for independence, we ask: "Is there any evidence in the data of an *association* between two categorical variables"? As before, the *observed frequencies* are compared with *expected frequencies* calculated under a null hypothesis. In this case, the null hypothesis is that of independence between the variables.

5.3.1 Tests of independence: basic framework

As always, we now outline the basic framework for our hypothesis test. The steps to consider in a χ^2 test of independence are exactly the same as for the other hypothesis tests considered in this course so far, i.e.

1. **State the null hypothesis (H_0)**
   The null hypothesis is that there is no association between the two categorical variables, or that the two categorical variables are independent; i.e.

   H_0 : There is no association between the two categorical variables, or
   H_0 : The two categorical variables are independent.

2. **State the alternative hypothesis (H_1)**
   This is just the opposite statement to the null hypothesis:

   H_1 : There *is* an association between the two categorical variables, or
   H_1 : The two categorical variables are *not* independent.

3. **Calculate the test statistic**
   The test statistic is

$$X^2 = \sum \frac{(O-E)^2}{E},$$

   where O and E represent the *observed* and *expected* frequencies (respectively). This is most easily calculated by drawing up a table:

Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
$\vdots$	$\vdots$	$\vdots$

   and then summing the final column. Remember, the observed frequencies are the frequencies given in the table; *you* have to calculate the expected frequencies yourself! In the example concerning graduate employment status and gender, if (as the null hypothesis specifies) the two variables *are* independent, then, for example,

$$\Pr(\text{Female and Temporary}) = \Pr(\text{Female}) \times \Pr(\text{Temporary}).$$

Thus, the expected *frequency* for Female and Temporary is given by

$$\begin{aligned}
&\Pr(\text{Female and Temporary}) \times \text{overall sample size} \\
= \ & \frac{\text{row total for Female}}{\text{overall sample size}} \times \frac{\text{column total for Temporary}}{\text{overall sample size}} \times \text{overall sample size} \\
= \ & \frac{\text{row total for Female} \times \text{column total for Temporary}}{\text{overall sample size}} \\
= \ & \frac{152 \times 73}{310} \\
= \ & 35.794.
\end{aligned}$$

Generally, for an $r \times c$ contingency table with r rows and c columns, expected frequencies are calculated using

$$E = \frac{\text{row total} \times \text{column total}}{\text{overall sample size}}.$$

As with the χ^2 goodness–of–fit test, the χ^2 test for independence is only valid if all $E > 5$. If necessary, adjacent row or columns can be pooled to ensure this.

4. **Find the *p*–value for the test**
   A range for the p–value for a test of independence can be found in tabulated values from the χ^2 distribution (table 5.1). The degrees of freedom is found as

$$\nu = (\text{number of rows} - 1) \times (\text{number of columns} - 1).$$

5. **Reach a conclusion**
   This bit's the same as ever – use table 2.1 to interpret your p–value, and either retain (stick with) or reject the null hypothesis. Remember to always write a sentence in the context of the question, i.e. "the two categorical variables do not appear to be independent".

5.3.2 Example: Employment status and gender

We now perform a χ^2 test of independence to assess if there is any evidence of a relationship between employment status of graduates six months after graduation and gender.

Steps 1 and 2 *(hypotheses)*
Since we are testing for a relationship between employment status and gender, the null hypothesis for this test is

$$H_0 \ : \ \text{Employment status and gender are independent,}$$

i.e. there is no association between employment status and gender (i.e. it doesn't matter whether you're male or female, you have an equal chance of being in a permanent job, a temporary job, or unemployed six months after graduation!).

The alternative is simply

$$H_1 \ : \ \text{Employment status and gender are } \textit{not} \text{ independent,}$$

i.e. there *is* an association between employment status and gender (i.e. males and females do *not* have an equal chance of being in a permanent job, a temporary job, or unemployed six months after graduating!).

Step 3 *(calculating the test statistic)*
Recall that, for any χ^2 test, the test statistic is

$$X^2 \ = \ \sum \frac{(O-E)^2}{E},$$

where O and E are the observed and expected frequencies (respectively). We already have the observed frequencies:

	Permanent	Temporary	Unemployed	Total
Male	100	33	25	158
Female	90	40	22	152
Total	190	73	47	310

To calculate the test statistic, we also need the expected frequencies! Remember that, if the null hypothesis is true, then

$$E \ = \ \frac{\text{row total} \times \text{column total}}{\text{overall sample size}}.$$

Thus, for "cell 1" (Male and Permanent), we have

$$\begin{aligned} E_1 \ &= \ \frac{158 \times 190}{310} \\ &= \ \frac{30020}{310} \\ &= \ 96.839. \end{aligned}$$

For "cell 2" (Male and Temporary), we have

$$\begin{aligned} E_2 &= \frac{158 \times 73}{310} \\ &= \frac{11534}{310} \\ &= 37.206. \end{aligned}$$

Similarly, for "cell 3" (Male and Unemployed), we have

$$\begin{aligned} E_3 &= \frac{158 \times 47}{310} \\ &= \frac{7426}{310} \\ &= 23.955. \end{aligned}$$

The corresponding frequencies for the other cells are shown in the table below:

	Permanent	Temporary	Unemployed	Total
Male	96.839	37.206	23.955	158
Female	93.161	35.794	23.045	152
Total	190	73	47	310

Notice that none of the expected frequencies are ≤ 5, and so we can proceed straight to the calculation of the test statistic without having to pool any categories. Thus, we have

O	E	$\frac{(O-E)^2}{E}$
100	96.839	0.103
33	37.206	0.475
25	23.955	0.046
90	93.161	0.107
40	35.794	0.494
22	23.045	0.047
		1.272

and so our test statistic is $X^2 = 1.272$.

Step 4 *(finding the p–value for the test)*
As with goodness–of–fit tests, we refer to χ^2 tables (table 5.1) to obtain a range for our p–value. However, in tests for independence, the degrees of freedom is given by

$$\nu = (\text{number of rows} - 1) \times (\text{number of columns} - 1).$$

In our example we have two rows and three columns, and so

$$\begin{aligned} \nu &= (2-1) \times (3-1) \\ &= 1 \times 2 \\ &= 2. \end{aligned}$$

Notice that the degrees of freedom are *not* one less than the total number of cells in the table! Referring to table 5.1, we obtain the following critical values for 2 degrees of freedom:

Significance level	10%	5%	1%
Critical value	4.61	5.99	9.21

Our test statistic $X^2 = 1.272$ lies to the left of the first critical value; thus our p–value is bigger than 10%.

Step 5 *(form your conclusion)*
Since our p–value is larger than 10%, there is insufficient evidence to reject the null hypothesis. It appears that the two categorical variables (employment status and gender) *are* independent; i.e., we can conclude that employment status and gender are not associated, and so whether you are male or female, you have the same chance of being in a permanent job, a temporary job, or unemployed six months after graduation! (Of course, this conclusion is drawn from the information in this particular sample.)

5.3.3 Another example (Question A6 in the 2006 exam)

In a market research survey 200 people are shown a proposed design for the new Mini Cooper car, and they are asked if they like the design. The responses, broken down by age groups, are shown in the table below.

	Under 21	21–35	Over 35	**Total**
Liked design	24	38	70	132
Disliked design	35	17	16	68
Total	59	55	86	200

Is there any evidence to suggest that age is associated with attitude to the proposed design?

Solution

Steps 1 and 2 *(hypotheses)*

Step 3 *(test statistic)*

Recall that the test statistic is:

$$X^2 = \sum \frac{(O-E)^2}{E}.$$

We have the O's – these are just the **O**bserved frequencies. We need the "E"'s as well! Remember, these are given by:

$$E = \frac{\text{row total} \times \text{column total}}{\text{overall sample size}}.$$

For "cell 1" (Liked design and Under 21), we get:

For the other cells we get:

To calculate the test statistic, it helps to draw up a table:

O (Observed frequencies)	E (Expected frequencies)	$\frac{(O-E)^2}{E}$

Thus, our test statistic is:

$$X^2 = \frac{(O-E)^2}{E}$$

$$=$$

Step 4 *(p–value)*

Our degrees of freedom to use Table 5.1 is given by

$$\nu = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

$$=$$

$$=$$

Referring to Table 5.1, we thus get:

p–value	10%	5%	1%
Critical value			

Step 5 *(Conclusions)*

Exercises

1. Two groups of students were given I.Q. tests: group 1 consisted of 30 students classified by their peers as "drinkers", and group 2 consisted of 28 students classified by their peers as "non–drinkers". The number of students with a "below norm", "norm" and "above norm" I.Q. were counted, the results of which are summarised in the contingency table below:

	Below norm	Norm	Above norm	**Total**
Drinkers	12	10	8	30
Non–drinkers	9	10	9	28
Total	21	20	17	58

   Is there evidence of an association (or lack of independence) between the two factors?

2. A large retail chain is been taken to an industrial tribunal by one of its female staff for sexual discrimination in its employment strategy. Does the company have a discriminatory employment strategy?

	Male	Female	**Total**
Shelf Stackers	34	50	84
Checkout Staff	39	40	79
Buyers	12	8	20
Section Managers	10	1	11
Total	95	99	194

		p				
		50%	10%	5%	1%	0.1%
	1	0.45	2.17	3.84	6.63	10.83
	2	1.39	4.61	5.99	9.21	13.82
	3	2.37	6.25	7.82	11.34	16.27
	4	3.36	7.78	9.49	13.28	18.47
	5	4.34	9.24	11.07	15.09	20.52
	6	5.35	10.64	12.59	16.81	22.46
	7	6.35	12.02	14.07	18.48	24.32
ν	8	7.34	13.36	15.51	20.09	26.13
	9	8.34	14.68	16.92	21.67	27.88
	10	9.34	15.99	18.31	23.21	29.59
	12	11.34	18.55	21.03	26.22	32.91
	15	14.34	22.31	25.00	30.58	37.70
	20	19.34	28.41	31.41	37.57	45.32
	25	24.34	34.38	37.65	44.31	52.62
	30	29.34	40.26	43.77	50.89	59.70

Table 5.1: This table contains values of x for which $\Pr(X^2 > x) = p$, where X^2 has a χ^2–distribution with ν degrees of freedom