11 CHI-SQUARED

Objectives

After studying this chapter you should

- be able to use the χ^2 distribution to test if a set of observations fits an appropriate model;
- know how to calculate the degrees of freedom;
- be able to apply the χ^2 model to contingency tables, including Yates' correction for the 2×2 tables.

11.0 Introduction

The chi-squared test is a particularly useful technique for testing whether observed data are representative of a particular distribution. It is widely used in biology, geography and psychology.

Activity 1 How random are your numbers?

Can you make up your own table of random numbers? Write down 100 numbers 'at random' (taking values from 0 to 9). Do this without the use of a calculator, computer or printed random number tables. Draw up a frequency table to see how many times you wrote down each number. (These will be called your **observed** frequencies.)

If your numbers really are random, roughly how many of each do you think there ought to be? (These are referred to as **expected** frequencies.)

What model are you using for this distribution of expected frequencies?

What assumptions must you make in order to use this model?

Do you think you were able to fulfil those assumptions when you wrote the numbers down?

Can you think of a way to test whether your numbers have a similar frequency distribution to what we would expect for true random numbers?

For analysing data of the sort used in Activity 1 where you are comparing observed with expected values, a chart as shown opposite is a useful way of writing down the data.

	Frequency	
Number	Observed, O_i	Expected, E_i
0		
1		
2		
3		
4		
.		
.		
.		

11.1 The chi-squared table

For your data in Activity 1, try looking at the differences $O_i - E_i$.

What happens if you total these?

Unfortunately the positive differences and negative differences always cancel each other out and you always have a zero total.

To overcome this problem the differences $O - E$ can be squared. So $\Sigma(O-E)^2$ could form the basis of your 'difference measure'. In this particular example however, each figure has an equal expected frequency, but this will not always be so (when you come to test other models in other situations). The importance assigned to a difference must be related to the size of the expected frequency. A difference of 10 must be more significant if the expected frequency is 20 than if it is 100.

One way of allowing for this is to divide each squared difference by the expected frequency for that category.

Here is an example worked out for you :

	Observed frequency	Expected frequency			
Number	O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	11	10	1	1	0.1
1	12	10	2	4	0.4
2	8	10	–2	4	0.4
3	14	10	4	16	1.6
4	7	10	–3	9	0.9
5	9	10	–1	1	0.1
6	9	10	–1	1	0.1
7	8	10	–2	4	0.4
8	14	10	4	16	1.6
9	8	10	–2	4	0.4
					6.0

For this set of 100 numbers $\sum \frac{(O-E)^2}{E} = 6.$

But what does this measure tell you?

How can you decide whether the observed frequencies are close to the expected frequencies or really quite different from them?

Firstly, consider what might happen if you tried to test some true random numbers from a random number table.

Would you actually get 10 for each number? The example worked out here did in fact use 100 random numbers from a table and not a fictitious set made up by someone taking part in the experiment.

Each time you take a sample of 100 random numbers you will get a slightly different distribution and it would certainly be surprising to find one with **all** the observed frequencies equal to 10. So, in fact, each different sample of 100 true random numbers will give a different value for $\sum \frac{(O-E)^2}{E}$.

The distribution of $\sum \frac{(O-E)^2}{E}$ is very close to the theoretical distribution known as χ^2 (or chi-squared). In fact, there is a family of χ^2 distributions, each with a different shape depending on the number of **degrees of freedom** denoted by υ (pronounced 'new').

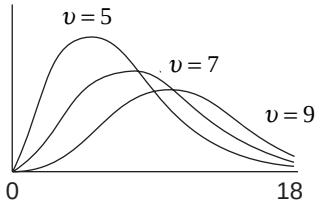
The distribution in this case is denoted by χ^2_υ.

For any χ^2 distribution, the number of degrees of freedom shows the number of independent free choices which can be made in allocating values to the expected frequencies. In these examples, there are ten expected frequencies (one for each of the numbers 0 to 9). However, as the total frequency must equal 100, only nine of the expected frequences can vary independently and the tenth one must take whatever value is required to fulfil that 'constraint'. To calculate the number of degrees of freedom

$$\upsilon = \text{number of classes or groups} - \text{number of constraints.}$$

Here there are ten classes and one constraint so

$$\upsilon = 10 - 1$$
$$= 9.$$



The shape of the χ^2_υ distribution is different for each value of υ and the function is very complicated. The mean of χ^2_υ is υ and the variance is 2υ. The distribution is positively skewed except for large values of υ for which it becomes approximately symmetrical.

Significance testing

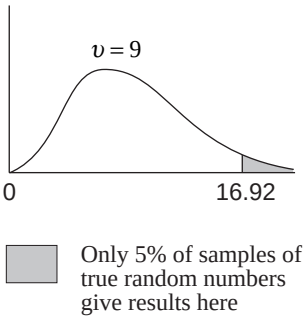
The set of random numbers shown in the table on page 204 generated a value of χ^2 equal to 6. You can see where this value comes in the χ^2 distribution with 9 degrees of freedom.

A high value of χ^2 implies a poor fit between the observed and expected frequencies, so the right hand end of the distribution is used for most hypothesis testing.

From χ^2 tables you find that only 5% of all samples of true random numbers will give a value of χ^2_9 greater than 16.92. This is called the **critical value** of χ^2 at 5%. If the **calculated value** of χ^2 from

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

is less than 16.92, it would support the view that the numbers are random. If not, you would expect that the numbers are not truly random.



Only 5% of samples of true random numbers give results here

What do you conclude from the example above, where $\chi^2 = 6$?

Activity 2

What happens when you test your made up 'random' numbers? Is their distribution close to what you would expect for true random numbers?

A summary of the critical values for χ^2 at 5% and 1% is given opposite for degrees of freedom $\upsilon = 1, 2, \ldots, 10$. (A more detailed table is given in the Appendix, Table 6, p261.)

Degree of freedom, υ	χ^2 5%	χ^2 1%
1	3.84	6.64
2	5.99	9.21
3	7.82	11.35
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21

Example

Nadir is testing an octahedral die to see if it is unbiased. The results are given in the table below.

Score	1	2	3	4	5	6	7	8
Frequency	7	10	11	9	12	10	14	7

Test the hypothesis that the die is fair.

Solution

Using χ^2, the number of degrees of freedom is $8-1=7$, so at the 5% significance level the critical value of χ^2 is 14.07. As before, a table of values is drawn up, the expected frequencies being based on a uniform distribution which gives

$$\text{frequency for each result} = \frac{1}{8}(7+10+11+9+12+10+14+7) = 10.$$

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
7	10	–3	9	0.9
10	10	0	0	0
11	10	1	1	0.1
9	10	–1	1	0.1
12	10	2	4	0.4
10	10	0	0	0
14	10	4	16	1.6
7	10	–3	9	0.9
				4.0

The calculated value of χ^2 is 4.0. This is well within the critical value, so Nadir could conclude that there is evidence to support the hypothesis that the die is fair.

Exercise 11A

1. Nicki made a tetrahedral die using card and then tested it to see whether it was fair. She got the following scores:

Score	1	2	3	4
Frequency	12	15	19	22

Does the die seem fair?

2. Joe has a die which has faces numbered from 1 to 6. He got the following scores:

Score	1	2	3	4	5	6
Frequency	17	20	29	20	18	16

He thinks that the die may be biased.

What do you think?

3. The table below shows the number of pupils absent on particular days in the week.

Day	M	Tu	W	Th	F
Number	125	88	85	94	108

Find the expected frequencies if it is assumed that the number of absentees is independent of the day of the week.

Test, at 5% level, whether the differences in observed and expected frequencies are significant.

11.2 Contingency tables

In many situations, individuals are classified according to two sets of attributes, and you may wish to investigate the dependency between these attributes. This is dealt with by using a contingency table and the χ^2 distribution.

2×2 contingency tables

The method of approach is illustrated in the example below.

Example

Some years ago a polytechnic decided to require all entrants to a science course to study a non-science subject for one year. In the first year all of the scheme entrants were given the choice of studying French or Russian. The numbers of students of each sex choosing each language are shown in the following table.

	French	Russian
Male	39	16
Female	21	14

Use a χ^2 test (including Yates' correction) at the 5% significance level to test whether choice of language is independent of sex.

Solution

The **observed** frequencies are given in the 2×2 contingency table.

	French	Russian	Total
Male	39	16	55
Female	21	14	35
Total	60	30	90

The null hypothesis is, as usual,

H_0 : there is no relationship between choice of language and sex

and so the alternative hypothesis is

H_1 : the choice of language is dependent on sex.

Assuming the null hypothesis, you need to calculate the expected frequency. For example,

$$P(\text{student is male}) = \frac{55}{90}$$

$$P(\text{student chooses French}) = \frac{60}{90}$$

Since these two events are independent under H_0,

$$P(\text{student is male and chooses French}) = \frac{55}{90} \times \frac{60}{90},$$

and, since there are 90 students,

$$\text{expected frequency (for male and French)} = \frac{55}{90} \times \frac{60}{90} \times 90$$

$$= \frac{55 \times 60}{90}$$
$$= 36.67.$$

There is no need to go through this procedure each time since it can be calculated directly from

$$\boxed{\text{Expected frequency} = \frac{(\text{row total})\ (\text{column total})}{(\text{grand total})}}$$

In fact, for a 2×2 table only one of these calculations is needed.

The row and column totals can be used to find the other expected values. For example,

$$\text{Expected frequency (for female and French)} = 60 - 36.67$$

$$= 23.33.$$

In this way, the table of expected frequency is as shown below.

	French	Russian	Total
Male	36.67	18.33	55
Female	23.33	11.67	35
Total	60	30	90

Since there is only one expected frequency needed in order to find the rest, the

$$\boxed{\text{degree of freedom, } \upsilon = 1}$$

But, for $\upsilon = 1$, you have to use Yates' continuity correction which evaluates

$$\chi^2_{calc} = \sum_{i=1}^{4} \frac{\left(|O_i - E_i| - 0.5\right)^2}{E_i}.$$

From tables, the critical χ^2 at 5% level is given by 3.84. Hence H_0 is rejected if $\chi^2_{calc} > 3.84$. Now

O_i	E_i	$\lvert O_i - E_i \rvert$	$\frac{\left(\lvert O_i - E_i \rvert - 0.5\right)^2}{E_i}$
39	36.67	2.33	0.091
16	18.33	2.33	0.183
21	23.33	2.33	0.144
14	11.67	2.33	0.287

$$\chi^2_{calc} = 0.705 < 3.84,$$

the critical χ^2 value. Hence you can conclude that there is no evidence to reject H_0; i.e. choice of subject and sex are independent.

Why are all the values in the $|O_i - E_i|$ column the same?

$h \times k$ contingency tables (h rows, k columns)

This is illustrated with an extension to the previous question, which also illustrates the convention that any entry with expected frequency of 5 or less should be eliminated by combining classes together.

Example

Following the example above, the choice of non-science subjects has now been widened and the current figures are as follows

	French	Poetry	Russian	Sculpture
Male	2	8	15	10
Female	10	17	21	37

Use a χ^2 test at the 5% significance level to test whether choice of subject is independent of sex. In applying the test you should combine French with another subject. Explain why this is necessary and the reasons for your choice of subject.

Solution

This is a 2×4 contingency table of **observed** values.

	French	Poetry	Russian	Sculpture	Total
Male	2	8	15	10	35
Female	10	17	21	37	85
Total	12	25	36	47	120

The **expected** frequency for 'male' and 'French' is

$$\frac{12 \times 35}{120} = 3.5.$$

Since this is less than 5, French should be combined with another subject, and the obvious choice is Russian since this is also a language.

Combining the French and Russian together gives

	Fr / Rus	Poetry	Sculpture	Total
Male	17	8	10	35
Female	31	17	37	85
Total	48	25	47	120

As before, H_0 : sex and choice of subject are independent

H_1 : sex and choice of subject are dependent.

The number of degrees of freedom is 2, since determining just 2 expected values will be sufficient to find the rest.

Note that, in general, for an $h \times k$ contingency table

$$\text{No. of degrees of freedom} = (h-1) \times (k-1)$$

(In the example above, $h = 2$, $k = 3$, giving the number of degrees of freedom as $(2-1) \times (3-1) = 1 \times 2 = 2.$) Thus, the critical χ^2 value is 5.99.

The expected frequency for 'male' and 'French and Russian' is

$$\frac{35 \times 48}{120} = 14.00$$

and for 'male' and 'poetry' is

$$\frac{35\times 25}{120} = 7.29\,.$$

The rest of the values can now be calculated from the row and column tables to give the following expected frequences

	Fr / Rus	Poetry	Sculpture	Total
Male	14.00	7.29	13.71	35
Female	34.00	17.71	33.29	85
Total	48	25	47	120

and the calculated χ^2 is given by

O_i	E_i	$\lvert O_i - E_i \rvert$	$\frac{(O_i - E_i)^2}{E_i}$
17	14.00	3.00	0.643
8	7.29	0.71	0.069
10	13.71	3.71	1.004
31	34.00	3.00	0.265
17	17.71	0.71	0.028
37	33.29	3.71	0.413

$$\chi^2_{calc} = 2.422 \; < \; 5.99$$

the critical value. So you conclude again that there is no dependence between sex and choice of subject.

11.3 Miscellaneous Exercises

1. During an investigation into visits to a Health Centre, interest is focused on the social class of those attending the surgery.

   The table below shows the number of patients attending the surgery together with the population of the whole area covered by the Health Centre, each categorised by social class.

Social Class	I	II	III	IV	V
Patients	28	63	188	173	48
Population	200	500	1600	1200	500

   Use a χ^2 test, at the 5% level of significance, to decide whether or not these results indicate that those attending the surgery are a representative sample of the whole area with respect to social class.

   As part of the same investigation, the following table was constructed showing the reason for the patients' visits to the surgery, again categorised by social class.

	Social Class				
Reason	I	II	III	IV	V
Minor physical	10	21	98	91	27
Major physical	7	17	49	40	15
Mental & other	11	25	41	42	6

   Is there significant evidence to conclude that the reason for the patients' visits to the surgery is independent of their social class?

   Use a 5% level of significance.

   Give an interpretation of your results. (AEB)

2. (a) In a survey on transport, electors from three different areas of a large city were asked whether they would prefer money to be spent on general road improvement or on improving public transport. The replies are shown in the following contingency table.

	Area		
	A	B	C
Road improvement preferred	78	46	24
Public transport preferred	22	34	36

Use a χ^2 test at the 1% significance level to test whether the proportion favouring expenditure on general road improvement is independent of the area.

(b) The same electors were also asked whether they had access to a private car for their personal use. The numbers who did were 70, 40 and 15 respectively in the areas A, B and C respectively and of these 61, 30 and 10 respectively favoured general road improvements.

Construct BUT DO NOT ANALYSE two contingency tables, one for those with access to private cars and one for those without such access.

Given that the value of $\sum \frac{(O-E)^2}{E}$ is 4.21 for the first of these tables and 4.88 for the second of these tables, test in each case, at the 5% significance level whether the proportion favouring general road improvements is independent of area.

(c) Examine your results in (a) and (b) and give an explanation of any apparent inconsistency. (AEB)

3. A hospital employs a number of visiting surgeons to undertake particular operations. If complications occur during or after the operation the patient has to be transferred to a larger hospital nearby where the required back up facilities are available.

A hospital administrator, worried by the effects of this on costs, examines the records of three surgeons. Surgeon A had 6 out of her last 47 patients transferred, surgeon B, 4 out of his last 72 patients and surgeon C, 14 out of his last 41.

Form the data into a 2×3 contingency table and test, at the 5% significance level, whether the proportion transferred is independent of the surgeon.

The administrator decides to offer as many operations as possible to surgeon B. Explainwhy and suggest what further information you would need before deciding whether the administrator's decision was based on valid evidence. (AEB)

4. A group of students studying A-level Statistics was set a paper, to be attempted under examination conditions, containing four questions requiring the use of the χ^2 distribution. The following table shows the type of question and the number of students who obtained good (14 or more out of 20) and bad (fewer than 14 out of 20) marks.

	Type of question			
	Contingency table	Binomial fit	Normal fit	Poisson fit
Good mark	25	12	12	11
Bad mark	4	11	3	12

(a) Test at the 5% significance level whether the mark obtained (by the students who attempted the question) is associated with the type of question.

(b) Under some circumstances it is necessary to combine classes in order to carry out a test. If it had been necessary to combine the binomial fit question with any other question, which question would you have combined it with and why?

(c) Given that a total of 30 students sat the paper, test, at the 5% significance level, whether the number of students attempting a particular question is associated with the type of question.

(d) Compare the difficulty and popularity of the different types of question in the light of your answers to (a) and (b). (AEB)

5. (a) The number of books borrowed from a library during a certain week were 518 on Monday, 431 on Tuesday, 485 on Wednesday, 443 on Thursday and 523 on Friday.

Is there any evidence that the number of books borrowed varies between the five days of the week? Use a 1% level of significance.

Interpret fully your conclusions.

(b) Analysis of the rate of turnover of employees by a personnel manager produced the following table showing the length of stay of 200 people who left the company for other employment.

	Length of employment (years)		
Grade	0-2	2-5	>5
Managerial	4	11	6
Skilled	32	28	21
Unskilled	25	23	50

Using a 1% level of significance, analyse this information and state fully the conclusions from your analysis. (AEB)