

Calculations in full

5.1. Chi-squared goodness of fit test

EXAMPLE 5.1. The distribution of holly leaf miners on *Ilex aquifolia*

As part of a student project, the number of holly leaf miners was recorded on a holly tree. The heights of the holly leaf miners within the tree were recorded. If the distribution was random you would expect equal numbers at each height.

Table 5.1. Contingency table for Example 5.1. The distribution of holly leaf miners on a single *Ilex aquifolia* tree

	Height on holly tree (m)			Total number of holly leaf miners
	0.00–1.99	2.00–3.99	4.00–5.99	
Observed number of holly leaf miners	131	38	2	171

Having organized the observed values in a contingency table and checked that the criteria for using this test are met (5.1.1.), you can now proceed with the test.

BOX 5.1. How to calculate a goodness of fit chi-squared test

THIS EXAMPLE

1. Hypotheses to be tested

H_0 : There is no difference between the numbers of holly leaf miners found at various heights (m) on the tree compared with those expected.

H_1 : There is a significant difference between the numbers of holly leaf miners found at various heights (m) on the tree compared with those expected.

2. How to work out expected values

We are expecting a random distribution of the holly leaf miners throughout the tree. This means that we should find equal numbers in each height category. To work out the expected numbers divide the total number of holly leaf miners (171) by the number of classes (3) i.e. $171/3 = 57$ (Table 5.2.).

3. How to work out $\chi^2_{\text{calculated}}$

$$\begin{aligned}\chi^2_{\text{calculated}} &= \frac{(131 - 57)^2}{57} + \frac{(38 - 57)^2}{57} + \frac{(2 - 57)^2}{57} \\ &= \frac{(74)^2}{57} + \frac{(-19)^2}{57} + \frac{(-55)^2}{57} \\ &= \frac{(5476)}{57} + \frac{(361)}{57} + \frac{(3025)}{57} \\ &= 96.07018 + 6.33333 + 53.07018 \\ &= 155.47368\end{aligned}$$

4. How to find χ^2_{critical}

The categories in this example are '0.00 – 1.99 m', '2.00 – 3.99 m', and '4.00 – 6.99 m', and therefore the degrees of freedom (v) = $3 - 1 = 2$.

When $p = 0.05$ and $v = 2$, then in the statistical table of critical values χ^2_{critical} is 5.99.

5. The rule

$\chi^2_{\text{calculated}}$ (155.47) is greater than χ^2_{critical} (5.99) at $p = 0.05$ and therefore we reject the null hypothesis.

In fact at $p = 0.001$ the χ^2_{critical} is 13.82. Therefore we can reject the null hypothesis at this higher level of significance.

6. What does this mean in real terms?

There is a significant difference ($\chi^2_{\text{calculated}}, p < 0.001$) between the numbers of holly leaf miners found at the various levels on the tree compared with those expected, such that the holly leaf miners are not found in equal numbers at all heights.

5.1.3. How to check if your data have a normal distribution using a goodness of fit chi-squared test

In Chapter 3 we discussed distributions, and in particular the normal distribution. Data with a normal distribution can be analysed using parametric statistics and it is therefore important to be able to check if your data are normally distributed. Section 3.8. and BOX 3.2. explained a number of ways to check your data to see if they are normally distributed. The best support for deciding whether your data are normally distributed is to calculate expected values using the Gaussian equation and then use a chi-squared goodness of fit test to check that your observed data are not

statistically significant from that predicted by the equation. To do this requires some extra steps to that described in BOX 5.1. and a change to how you work out the degrees of freedom, so we have included a worked example here. The observed values we will use come from Example 3.7. Length (mm) of two-spot ladybirds (*Adalia bipunctata*).

EXAMPLE 3.7. Length (mm) of two-spot ladybirds (*Adalia bipunctata*)

An investigator was interested in the length of two-spot ladybirds (*Adalia bipunctata*). In an observational investigation she measured the length (mm) of 50 ladybirds collected at random from a garden (Table 3.8).

Table 3.8. The length (mm) of 50 *Adalia bipunctata* sampled in a garden

Length of <i>Adalia bipunctata</i> (mm) (x)									
1	5	2	5	7	8	3	6	7	4
4	5	6	4	5	5	7	5	3	5
4	5	1	7	9	2	6	5	6	3
3	6	8	6	4	6	6	8	5	6
7	4	8	9	5	4	3	4	2	5

These data can be organized in a frequency table (Table 3.9.). The mid-point for each size class is required for this calculation, so we have included this in the table.

Table 3.9. Frequency table of length (mm) of two-spot *Adalia bipunctata* (ladybirds) showing how to calculate a mean for grouped data

Size classes for length (mm) of <i>Adalia bipunctata</i> (ladybird)									
	1.0–1.9	2.0–2.9	3.0–3.9	4.0–4.9	5.0–5.9	6.0–6.9	7.0–7.9	8.0–8.9	9.0–9.9
Mid-point of class (m)	1.45	2.45	3.45	4.45	5.45	6.45	7.45	8.45	9.45
Frequency (f)	2	3	5	8	12	9	5	4	2

Having confirmed as far as possible that the data meet the criteria for using a chi-squared goodness of fit test (5.1.3.), we calculate the expected numbers using the Gaussian equation, which is the mathematical

equation that describes the normal distribution. Where

$$y = \frac{1}{\sqrt{(2\pi s^2)}} e^{-b}$$

and

$$b = \frac{(x - \bar{x})^2}{2s^2}$$

The terms in this equation were explained in 3.2.1. and appendix c. The symbols e and π are constants and are 2.71828 and 3.14159 respectively. The \bar{x} (mean) and s^2 (variance) were calculated in BOX 3.2. and are $\bar{x} = 5.08$ mm, $s^2 = 3.74857$ mm². x is the mid-point for each class.

Therefore all the terms have known values apart from y . These y values can be calculated for each x in your sample. Don't be put off even though it looks complicated. Break the calculation down into smaller steps.

When $x = 1.45$ mm

The first part of the equation is

$$\frac{1}{\sqrt{(2\pi s^2)}} = \frac{1}{\sqrt{(2 \times 3.14 \times 3.74857)}} = \frac{1}{\sqrt{23.5529}} = \frac{1}{4.8531} = 0.20605$$

The second part of the equation involves the exponential term e . First work out the value for b :

$$b = \frac{(x - \bar{x})^2}{2s^2} = \frac{(1.45 - 5.08)^2}{2 \times 3.74857} = \frac{(-3.63)^2}{7.49714} = \frac{13.1769}{7.49714} = 1.75759$$

Use your calculator function buttons to find $e^{-1.75759} = 0.17246$

So when $x = 1.45$, $y = 0.20605 \times 0.17246 = 0.03554$.

This y value is a proportion and our final step is to calculate expected numbers from these proportions. Since the sample size for our ladybird example is 50 then the expected number for y , when $x = 1.45$ is $50 \times 0.03554 = 1.77678$

When $x = 2.45$ mm.

The first part of the equation is the same as before:

$$\frac{1}{\sqrt{(2\pi s^2)}} = 0.20605$$

$$b = \frac{(x - \bar{x})^2}{2s^2} = \frac{(2.45 - 5.08)^2}{2 \times 3.74857} = \frac{(-2.63)^2}{7.49714} = \frac{6.9169}{7.49714} = 0.92261$$

Use your calculator function buttons to find $e^{-0.92261} = 0.39748$.

So when $x = 2.45$, $y = 0.20605 \times 0.39748 = 0.08190$.

Therefore the expected number for y , when $x = 2.45$ is $50 \times 0.08190 = 4.09506$.

When $x = 3.45$ mm.

The first part of the equation is the same as before

$$\frac{1}{\sqrt{(2\pi s^2)}} = 0.20605$$

$$h = \frac{(x - \bar{x})^2}{2s^2} = \frac{(3.45 - 5.08)^2}{2 \times 3.74857} = \frac{(-1.63)^2}{7.49714} = \frac{2.6569}{7.49714} = 0.35439$$

Use your calculator function buttons to find $e^{-0.35439} = 0.70160$.

So when $x = 3.45$, $y = 0.20605 \times 0.70160 = 0.14457$.

This y value is a proportion and our final step is to calculate expected numbers from these proportions. Since the sample size for our ladybird example is 50, then the expected number for y , when $x = 3.45$, is $50 \times 0.14457 = 7.22826$.

When $x = 4.45$ mm.

The first part of the equation is the same as before:

$$\frac{1}{\sqrt{(2\pi s^2)}} = 0.20605$$

$$h = \frac{(x - \bar{x})^2}{2s^2} = \frac{(4.45 - 5.08)^2}{2 \times 3.74857} = \frac{(-0.63)^2}{7.49714} = \frac{0.3969}{7.49714} = 0.05294$$

Use your calculator function buttons to find $e^{-0.05294} = 0.94844$

So when $x = 4.45$, $y = 0.20605 \times 0.94844 = 0.19543$.

This y value is a proportion and our final step is to calculate expected numbers from these proportions. Since the sample size for our ladybird example is 50 then the expected number for y , when $x = 4.45$ is $50 \times 0.19543 = 9.7714$

When $x = 5.45$ mm.

The first part of the equation is the same as before:

$$\frac{1}{\sqrt{(2\pi s^2)}} = 0.20605$$

$$h = \frac{(x - \bar{x})^2}{2s^2} = \frac{(5.45 - 5.08)^2}{2 \times 3.74857} = \frac{(0.37)^2}{7.49714} = \frac{0.1369}{7.49714} = 0.01826$$

Use your calculator function buttons to find $e^{-0.01826} = 0.98191$

So when $x = 5.45$, $y = 0.20605 \times 0.98191 = 0.20232$.

This y value is a proportion and our final step is to calculate expected numbers from these proportions. Since the sample size for our ladybird

example is 50, then the expected number for y , when $x = 5.45$, is $50 \times 0.20232 = 10.11608$.

When $x = 6.45$ mm.

The first part of the equation is the same as before:

$$\frac{1}{\sqrt{(2\pi s^2)}} = 0.20605$$

$$b = \frac{(x - \bar{x})^2}{2s^2} = \frac{(6.45 - 5.08)^2}{2 \times 3.74857} = \frac{(1.37)^2}{7.49714} = \frac{1.8769}{7.49714} = 0.25035$$

Use your calculator function buttons to find $e^{-0.25035} = 0.77853$

So when $x = 6.45$, $y = 0.20605 \times 0.77853 = 0.16042$.

This y value is a proportion and our final step is to calculate expected numbers from these proportions. Since the sample size for our ladybird example is 50 then the expected number for y , when $x = 6.45$ is $50 \times 0.16042 = 8.0208$.

When $x = 7.45$ mm.

The first part of the equation is the same as before:

$$\frac{1}{\sqrt{(2\pi s^2)}} = 0.20605$$

$$b = \frac{(x - \bar{x})^2}{2s^2} = \frac{(7.45 - 5.08)^2}{2 \times 3.74857} = \frac{(2.37)^2}{7.49714} = \frac{5.6169}{7.49714} = 0.74921$$

Use your calculator function buttons to find $e^{-0.74921} = 0.47274$

So when $x = 7.45$, $y = 0.20605 \times 0.47274 = 0.09741$

This y value is a proportion and our final step is to calculate expected numbers from these proportions. Since the sample size for our ladybird example is 50, then the expected number for y , when $x = 7.45$ is $50 \times 0.09741 = 4.87042$.

When $x = 8.45$ mm.

The first part of the equation is the same as before:

$$\frac{1}{\sqrt{(2\pi s^2)}} = 0.20605$$

$$b = \frac{(x - \bar{x})^2}{2s^2} = \frac{(8.45 - 5.08)^2}{2 \times 3.74857} = \frac{(3.37)^2}{7.49714} = \frac{11.3569}{7.49714} = 1.51483$$

Use your calculator function buttons to find $e^{-1.51483} = 0.21985$

So when $x = 8.45$, $y = 0.20605 \times 0.21985 = 0.0453$.

This y value is a proportion and our final step is to calculate expected numbers from these proportions. Since the sample size for our ladybird example is 50 then the expected number for y , when $x = 8.45$, is $50 \times 0.0453 = 2.265$.

When $x = 9.45$ mm.

The first part of the equation is the same as before:

$$\frac{1}{\sqrt{(2\pi s^2)}} = 0.20605$$

$$h = \frac{(x - \bar{x})^2}{2s^2} = \frac{(9.45 - 5.08)^2}{2 \times 3.74857} = \frac{(4.37)^2}{7.49714} = \frac{19.0969}{7.49714} = 2.54722$$

Use your calculator function buttons to find $e^{-2.54722} = 0.0783$

So when $x = 9.45$, $y = 0.20605 \times 0.0783 = 0.01613$.

This y value is a proportion and our final step is to calculate expected numbers from these proportions. Since the sample size for our ladybird example is 50, then the expected number for y , when $x = 9.45$, is $50 \times 0.01613 = 0.80667$.

These are your expected values. It is clear that 5/9 of the expected values are less than 5. To overcome this (5.7.) we will add together the observed values ($2 + 3 = 5$), expected values ($1.77678 + 4.09506 = 5.87185$) for the two lower classes, and the observed ($4 + 2 = 6$) and expected numbers ($2.265 + 0.80667 = 3.07167$) for the two upper classes. These values will be used in the chi-squared test. You can now carry on with the chi-squared goodness of fit test (BOX 5.2.).

BOX 5.2. To check if your data are normally distributed using a goodness of fit chi-squared test

THIS EXAMPLE

1. Hypotheses to be tested

H₀: There is no difference between the observed lengths of ladybirds compared with that expected if the data are normally distributed.

H₁: There is a difference between the observed lengths of ladybirds compared with that expected if the data are normally distributed. The data are not normally distributed.

2. How to work out expected values

Scroll up to see how the expected values are calculated and how the small expected values are dealt with.

3. How to work out $\chi^2_{\text{calculated}}$

$$\begin{aligned}\chi^2_{\text{calculated}} &= \frac{(5 - 5.87185)^2}{5.87185} + \frac{(5 - 7.22826)^2}{7.22826} + \frac{(8 - 9.7714)^2}{9.7714} \\ &\quad + \frac{(12 - 10.11608)^2}{10.11608} + \frac{(9 - 8.0208)^2}{8.0208} + \frac{(5 - 4.87042)^2}{4.87042} \\ &\quad + \frac{(6 - 3.07167)^2}{3.07167} \\ &= \frac{(0.87185)^2}{1.77678} + \frac{(2.22826)^2}{7.22826} + \frac{(-1.7714)^2}{9.7714} + \frac{(1.88392)^2}{10.11608} \\ &\quad + \frac{(0.9791)^2}{8.0208} + \frac{(0.12958)^2}{4.87042} + \frac{(2.92833)^2}{3.07167} \\ &= \frac{0.76012}{5.87185} + \frac{4.96514}{7.22826} + \frac{3.13786}{9.7714} + \frac{3.54908}{10.11608} + \frac{0.95883}{8.0208} \\ &\quad + \frac{0.01679}{4.87042} + \frac{8.57512}{3.07167} \\ &= 0.12945 + 0.68691 + 0.32113 + 0.35084 + 0.11954 \\ &\quad + 0.00345 + 2.79168 \\ &= 4.40299\end{aligned}$$

4. How to find χ^2_{critical}

For the calculation we combined the lower two and top two classes, so there are seven categories in this example, so $v = 7 - 2 = 5$.

When $p = 0.05$, the degrees of freedom (v) = 5, then χ^2_{critical} is 11.07.

5. The rule

$\chi^2_{\text{calculated}}$ (4.40) is less than χ^2_{critical} (11.07) at $p = 0.05$ and therefore we do not reject the null hypothesis.

6. What does this mean in real terms?

There is no significant difference ($\chi^2_{\text{calculated}} = 4.40$, $p = 0.05$) between the observed lengths of ladybirds (mm) compared with that expected if the data are normally distributed. The data can be said to be normally distributed.

5.2. Heterogeneity in a goodness of fit test

There are some circumstances where you have several samples and wish to know if they can be pooled. For example, you may have carried out an investigation into the genetic inheritance of flower colour where you had several pairs of parent plants. Crossing within these pairs would produce F1 offspring. If you kept the seed from each cross separate from the others, grew

these F1 plants up and then crossed these you could collect the results from several different lineages. These separate lineages are known as accessions.

EXAMPLE 5.2. The genetics of tepal colour in *Allium schoenoprasum*

Crosses were carried out in three pairs of plants of *Allium schoenoprasum*. In each cross one parent had purple flowers and the other was white flowered. Seeds from each cross were collected and grown on. The offspring from each cross were kept as separate accessions. These F1 plants were all purple flowering and thought to be heterozygous for a single gene that controlled the pigmentation in the flowers and purple was believed to be the dominant allele. These F1 plants were then crossed amongst themselves within an accession and the seed from these crosses grown up and the flower colours recorded (Table 5.4.). If the genetic theory is correct then the F2 plants should occur in the ratio of three purple flowering plants:one white flowering plant.

Table 5.4. Flower colour exhibited by F2 *Allium schoenoprasum* in four accessions

	Flower colour in the F2 generation		Total number of plants flowering in the F2
	Purple	White	
Accession 1	127	41	168
Accession 2	123	39	162
Accession 3	107	53	160
Accession 4	130	42	172
Total	487	175	662

In examining Table 5.4. it appears that the total values (487 purple and 175 white) are reasonably close to a 3:1 ratio. When a goodness of fit chi-squared test is carried out on just these total values there is no significant difference ($\chi^2 = 0.73$, $p < 0.05$) between the observed values and that expected for a three purple:one white ratio (Table 5.5.). But should you combine the data from your different samples in this way? You can get some idea by carrying out a goodness of fit chi-squared test on each accession separately (Table 5.5.). These analyses confirm that the results from accessions 1, 2, and 4 also conform to a 3:1 ratio at $p = 0.05$. Accession 3, however, is different and the test there indicates that the observed values depart significantly from the predicted 3:1 ratio. Does this difference between accession 3 and accession 1, 2, and 4 mean that the data should not be pooled? In these circumstances you may use the chi-squared test to see if the data are statistically heterogeneous (different).

Table 5.5. Contingency table with expected values for a chi-squared test for heterogeneity for Example 5.2. Flower colour exhibited by F2 *Allium schoenoprasum* in four accessions

		Flower colour in the F2 generation		Total number of plants flowering in the F2	χ^2 from the goodness of fit test for each accession and for the total values
		Purple	White		
Accession 1	Observed	127	41	168	0.03175 (NS)
	Expected	126	42		
Accession 2	Observed	123	39	162	0.07407 (NS)
	Expected	121.5	40.5		
Accession 3	Observed	107	53	160	5.63333 $0.05 > p > 0.01$
	Expected	120	40		
Accession 4	Observed	130	42	172	0.03101 (NS)
	Expected	129	43		
Total	Observed	487	175	662	0.72709 (NS)

In this example when the criteria for using the chi-squared test for heterogeneity are checked it is clear that the criteria are met (5.2.1.). You now proceed with the chi-squared test for heterogeneity.

BOX 5.3. How to calculate a chi-squared test for heterogeneity

THIS EXAMPLE

1. Hypotheses to be tested

H_0 : There is no significant heterogeneity between the accessions of F2 *Allium schoenoprasum*.

H_1 : There is significant heterogeneity between the F2 accessions of *Allium schoenoprasum*.

2. How to work out expected values.

We are expecting three purple flowering plants:one white flowering plant. Since the grand total number of plants examined was 662 then $\frac{3}{4}$ should be purple flowering, i.e 496.5, and $\frac{1}{4}$ plants should be white flowering, i.e. 165.5. This can be extended to each accession (Table 5.5.).

3. How to work out $\chi^2_{\text{calculated}}$

1.

$$\begin{aligned}\text{Accession 1. } \chi^2 &= \frac{(127 - 126)^2}{126} + \frac{(41 - 42)^2}{42} = \frac{(1)^2}{126} + \frac{(1)^2}{42} = 1/126 + 1/42 \\ &= 0.00794 + 0.02381 \\ &= 0.03175\end{aligned}$$

$$\begin{aligned}\text{Accession 2. } \chi^2 &= \frac{(123 - 121.5)^2}{121.5} + \frac{(39 - 40.5)^2}{40.5} = \frac{(1.5)^2}{121.5} + \frac{(1.5)^2}{40.5} \\ &= \frac{2.25}{121.5} + \frac{2.25}{40.5} = 0.01852 + 0.05556 \\ &= 0.07407\end{aligned}$$

$$\begin{aligned}\text{Accession 3. } \chi^2 &= \frac{(107 - 120)^2}{120} + \frac{(53 - 40)^2}{40} = \frac{(-13)^2}{120} + \frac{(13)^2}{40} \\ &= \frac{169}{120} + \frac{169}{40} = 1.40833 + 4.225 \\ &= 5.63333\end{aligned}$$

$$\begin{aligned}\text{Accession 4. } \chi^2 &= \frac{(130 - 129)^2}{129} + \frac{(42 - 43)^2}{43} = \frac{(1)^2}{129} + \frac{(-1)^2}{43} \\ &= 0.00775 + 0.02326 \\ &= 0.03101\end{aligned}$$

Total values

$$\begin{aligned}\chi^2 &= \frac{(487 - 496.5)^2}{496.5} + \frac{(175 - 165.5)^2}{165.5} = \frac{(-9.5)^2}{496.5} + \frac{(9.5)^2}{165.5} \\ &= \frac{90.25}{496.5} + \frac{90.25}{165.5} = 0.18177 + 0.54532 = 0.72709\end{aligned}$$

ii. 'Summed' chi-squared

In our current example this 'summed' value is:

$$0.03175 + 0.07407 + 5.63333 + 0.03101 = 5.77016$$

iii. 'Deviation' chi-squared = 0.72709

$$\text{iv. } \chi^2_{\text{calculated}} = 5.77016 - 0.72709 = 5.04307$$

4. How to find χ^2_{critical}

- i. There are two categories (purple and white) so $v = 2 - 1 = 1$
- ii. There are four rows of observed values the accessions, therefore $v = 4$.
- iii. For this example the degrees of freedom $(v) = 4 - 1 = 3$.

The critical value to test for heterogeneity in the data is found in the chi-squared table at $v = 3$, $p = 0.05$ and is $\chi^2_{\text{critical}} = 7.81$.

5. *The rule*

$\chi^2_{\text{calculated}}$ (5.04) is less than χ^2_{critical} (7.81) at $p = 0.05$ and therefore we do not reject the null hypothesis.

6. *What does this mean in real terms?*

There is no statistically significant heterogeneity ($p = 0.05$) between the accessions of *Allium schoenoprasum* and it is therefore reasonable to sum the data across all accessions and use a goodness of fit chi-squared test on the totals.

5.3. Chi-squared test for association

A chi-squared test for association is used when you have categorical data for two variables and you wish to examine the possibility of an association between these two variables. The term ‘association’ has a very specific meaning. Before proceeding with this test you should read the first paragraph of Chapter 6 to ensure you understand how this term is being used.

EXAMPLE 5.3. Shell colour in *Cepea nemoralis* in coastal and hedgerow habitats

An investigation was carried out into the frequency of banding and colour patterns in snail shells and habitat. Four patterns were observed in the two populations studied: pink with bands, pink with no bands, yellow with bands, and yellow with no bands. These results are shown in Table 5.6. The association that is under investigation is therefore between the variables ‘shell pattern’ and ‘habitat’.

Table 5.6. Shell colour in *Cepea nemoralis* in coastal and hedgerow habitats

Habitat	Shell pattern and colour in <i>Cepea nemoralis</i>				Total
	Banded Yellow	No bands Yellow	Banded Pink	No bands Pink	
Coastal Observed	10	19	5	16	50
Hedgerow Observed	17	8	19	11	55
Total	27	27	24	27	105

At this point you will have arranged your data in a contingency table and checked the criteria for using this test (5.3.1.). You can now proceed to test for an association between the two variables.

BOX 5.4. How to calculate an $r \times c$ chi-squared test for association

THIS EXAMPLE

1. Hypotheses to be tested

H_0 : There is no association between the distribution of shell patterns observed and the habitat (coastal and hedgerow) of *Cepea nemoralis*.

H_1 : There is a significant association between the distribution of shell patterns and the habitat (coastal and hedgerow) of *Cepea nemoralis*.

2. How to work out expected values.

Look first at the expected value in row 1, column 1 in Table 5.3. To calculate this expected value = the column total for banded yellow (27) \div grand total (105) \times row total for coastal snails (50) = 12.85714.

This calculation is repeated for each row \times column combination as shown in Table 5.7.

Table 5.7. Contingency table with expected values for a chi-squared test for association using data from Example 5.3. Shell colour in *Cepea nemoralis* in coastal and hedgerow habitats

Habitat	Shell pattern and colour in <i>Cepea nemoralis</i>				Total
	Banded Yellow	No bands Yellow	Banded Pink	No bands Pink	
Coastal Observed	10	19	5	16	50
Coastal Expected	$27/105 \times 50 = 12.857143$	$27/105 \times 50 = 12.857143$	$24/105 \times 50 = 11.428571$	$27/105 \times 50 = 12.857143$	
Hedgerow Observed	17	8	19	11	55
Hedgerow Expected	$27/105 \times 55 = 14.142057$	$27/105 \times 55 = 14.142057$	$24/105 \times 55 = 12.571429$	$27/105 \times 55 = 14.142057$	
Total	27	27	24	27	105

3. How to work out $\chi^2_{\text{calculated}}$

$$\chi^2 = \frac{(10 - 12.85714)^2}{12.85714} + \frac{(19 - 12.85714)^2}{12.85714} + \frac{(24 - 11.42857)^2}{11.42857} + \frac{(16 - 12.85714)^2}{12.85714} + \frac{(17 - 14.14206)^2}{14.14206} + \frac{(8 - 14.14206)^2}{14.14206} + \frac{(19 - 12.57143)^2}{12.57143} + \frac{(11 - 14.14206)^2}{14.14206}$$

$$\begin{aligned}
&= \frac{(-2.85714)^2}{12.85714} + \frac{(6.14286)^2}{12.85714} + \frac{(12.57143)^2}{11.42857} + \frac{(3.14286)^2}{12.85714} + \frac{(2.85743)^2}{14.14206} \\
&\quad + \frac{(-6.14206)^2}{14.14206} + \frac{(6.42857)^2}{12.57143} + \frac{(-3.14206)^2}{14.14206} \\
&= \frac{8.16326}{12.85714} + \frac{37.73469}{12.85714} + \frac{158.04083}{11.42857} + \frac{9.87755}{12.85714} + \frac{8.16491}{14.14206} + \frac{37.72486}{14.14206} \\
&\quad + \frac{41.32653}{12.57143} + \frac{9.87252}{14.14206} \\
&= 0.63492 + 2.93492 + 13.82857 + 0.76825 + 0.57735 + 2.66757 \\
&\quad + 3.28734 + 0.6980
\end{aligned}$$

$$\chi^2_{\text{calculated}} = 15.18472$$

4. How to find χ^2_{critical}

In our snails example there are two rows (coastal and hedgerow). Do not include your 'expected' rows these are part of your calculation. There are four columns (banded yellow, no bands yellow, banded pink, no bands pink).

Therefore the degrees of freedom (v) = $(2 - 1) \times (4 - 1) = 1 \times 3 = 3$

The critical value to test for an association between the snail shell patterns and habitat is found in the chi-squared table where $v=3$, $p=0.05$, and is $\chi^2 = 7.81$.

5. The rule

$\chi^2_{\text{calculated}}$ (15.18) is greater than χ^2_{critical} (7.81) at $p = 0.05$ and therefore we reject the null hypothesis.

6. What does this mean in real terms?

There is a significant association ($\chi^2_{\text{calculated}} = 15.18$, $p = 0.05$) between the distribution of shell patterns and habitat (coastal and hedgerow) of *Cepea nemoralis*.

In fact at $p = 0.01$, $\chi^2_{\text{critical}} = 11.34$ and at $p = 0.001$ $\chi^2_{\text{critical}} = 16.27$. Therefore you may reject the null hypothesis at $p = 0.01$ but not at $p = 0.001$. This can be written as: There is a highly significant association ($\chi^2_{\text{calculated}} = 15.18$, $0.01 > p > 0.001$) between the distribution of shell patterns and habitat (coastal and hedgerow) of *Cepea nemoralis*.

5.4. Chi-squared with one degree of freedom

In any chi-squared test when there is only one degree of freedom the $\chi^2_{\text{calculated}}$ value is too high if calculated in the ways described earlier in this chapter. These calculations therefore have to be modified using a Yates'

correction. The Yates' correction reduces the $\chi^2_{\text{calculated}}$ value by subtracting 0.5 from the numerator.

5.4.1 Chi-squared goodness of fit test when there is one degree of freedom

In the earlier examples relating to a goodness of fit test we considered the distribution of holly leaf miners on a single holly tree (BOX 5.1.) and we checked to see if the distribution of a set of data is normal (BOX 5.2.). Another application of a goodness of fit test is when you have a genetic theory that is tested by carrying out a controlled cross. The observed results are then compared with those predicted by your theory. We used an example like this in 5.2., where there were several accessions. Here we only have a single accession and one degree of freedom.

EXAMPLE 5.4. The genetics of tepal colour in *Allium schoenoprasum*

A cross was carried out between two plants of *Allium schoenoprasum*. As in Example 5.2., the flower colours were purple and white respectively. The genetic model proposed for the inheritance of the tepal colour means that the F2 plants should occur in the ratio three purple flowering plants:one white flowering plant.

A contingency table for the data relating to this example is shown in Table 5.9. If you analysed this data using a chi-squared goodness of fit test then the degrees of freedom will be the number of categories $- 1 = 2 - 1 = 1$, therefore you should use the Yates' correction as show in BOX 5.5. The criteria for using this modified version of the chi-squared goodness of fit test are the same as described in 5.1.1., apart from the number of categories. For this modified test you will have one variable with only two categories.

Table 5.9. Flower colour in the F2 of a single *Allium schoenoprasum* cross

	Flower colour in the F2		Total number of plants flowering in the F2
	Purple	White	
Numbers of plants	131	37	168

BOX 5.5. How to calculate a goodness of fit chi-squared test when there is one degree of freedom

THIS EXAMPLE

1. Hypotheses to be tested

H₀: There is no difference between the observed numbers of plants with purple or white flowers in the F₂ of *Allium schoenoprasum* plants and that expected from the *a priori* prediction of three purple flowering plants:one white flowering plant.

H₁: There is a significant difference between the observed numbers of plants with purple or white flowers in the F₂ of *Allium schoenoprasum* plants and that expected from the genetic prediction of three purple flowering plants:one white flowering plant.

2. How to work out expected values

We are expecting three purple flowering plants:one white flowering plant. Since the total number of plants examined was 168 then $\frac{3}{4}$ should be purple flowering i.e 126, and $\frac{1}{4}$ plants should be white flowering i.e. 42.

3. How to work out $\chi^2_{\text{calculated}}$

$$\begin{aligned}\chi^2_{\text{calculated}} &= \frac{(|131 - 126| - 0.5)^2}{126} + \frac{(|37 - 42| - 0.5)^2}{42} = \frac{(|5| - 0.5)^2}{126} \\ &\quad + \frac{(|-5| - 0.5)^2}{42} \\ &= \frac{(5 - 0.5)^2}{126} + \frac{(5 - 0.5)^2}{42} = \frac{(4.5)^2}{126} + \frac{(4.5)^2}{42} = \frac{20.25}{126} + \frac{20.25}{42} \\ &= 0.16071 + 0.48214 = 0.64285\end{aligned}$$

4. How to find χ^2_{critical}

The categories in this example are 'purple' and 'white' and therefore $v = 2 - 1 = 1$.

When $p = 0.05$, $v = 1$, then χ^2_{critical} is 3.84

5. The rule

$\chi^2_{\text{calculated}}$ (0.64) is less than χ^2_{critical} (3.84) at $p = 0.05$ and therefore we do not reject the null hypothesis.

6. What does this mean in real terms?

The flower colours in the F₂ generation of *Allium schoenoprasum* plants do not differ significantly ($\chi^2_{\text{calculated}}$ 0.64, $p = 0.05$) from the predicted ratio of three purple:one white. This indicates that the genetic model is probably correct.

5.4.2. Chi-squared test for association when there is only one degree of freedom

Some experiments are designed to test for an association between two variables both of which have discrete categories. In our first example (Example 5.3.), we carried out a test of association between the numbers of snails with particular shell patterns and their habitat. But there are many occasions when our two variables will each only have two categories. This is often referred to as a 2×2 chi-squared test for association.

EXAMPLE 5.5. Frequency of *Cepea nemoralis* and *Cepea hortensis* in a hedge and wood

A survey of two species of snail was carried out at two habitats: a wood and hedgerow. The numbers of snails at each location was recorded (Table 5.11.)

Table 5.11. The distribution of *Cepea nemoralis* and *Cepea hortensis* in a woodland and a hedgerow

	Species of snail		Total
	<i>C. nemoralis</i>	<i>C. hortensis</i>	
Hedgerow Observed	89	59	148
Woodland Observed	16	8	24
Total	105	67	172

In an $r \times c$ chi-squared test for association the degrees of freedom are $(r - 1)(c - 1)$ (BOX 5.4.). For this example the degrees of freedom will therefore be $(2 - 1)(2 - 1) = 1$, so a Yates' correction needs to be applied. The criteria for using this modified version of the chi-squared test for association are the same as described in 5.3.1., apart from the number of categories. For this modified test you will have two variables each with only two categories.

Having organized your observed values in a contingency table and checked that the criteria for using this test are met, you can now proceed with the modified test.

BOX 5.6. How to calculate a 2 × 2 chi-squared test for association**THIS EXAMPLE****1. Hypotheses to be tested**

H₀: There is no association between the distribution of the two snail species (*Cepea nemoralis* and *C. hortensis*) and habitat (hedgerow and woodland).

H₁: There is a significant association between the distribution of the two snail species (*C. nemoralis* and *C. hortensis*) and habitat (hedgerow and woodland).

2. How to work out expected values

Look first at the expected value in row 1, column 1 in Table 5.12. To calculate this expected value take the column total for *C. nemoralis* (105) ÷ grand total (172) × row total for hedgerow snails (148) = 90.34884

This calculation is repeated for each row × column combination as shown in Table 5.12.

Table 5.12. Contingency table with expected values calculated for a 2 × 2 chi-squared test for association using Yates's correction for the Example 5.5. The distribution of *Cepea nemoralis* and *Cepea hortensis* in two habitats

	Species of snail		Total
	<i>C. nemoralis</i>	<i>C. hortensis</i>	
Hedgerow Observed	89	59	148
Hedgerow Expected	90.34884	57.65116	
Woodland Observed	16	8	24
Woodland Expected	14.65116	9.34884	
Total	105	67	172

3. How to work out $\chi^2_{\text{calculated}}$

$$\begin{aligned} \chi^2_{\text{calculated}} &= \frac{(|89 - 90.34884| - 0.5)^2}{90.34884} + \frac{(|59 - 57.65116| - 0.5)^2}{57.65116} \\ &+ \frac{(|16 - 14.65116| - 0.5)^2}{14.65116} + \frac{(|8 - 9.34884| - 0.5)^2}{9.34884} \\ &= \frac{(|-1.34884| - 0.5)^2}{90.34884} + \frac{(|1.34884| - 0.5)^2}{57.65116} + \frac{(|1.34884| - 0.5)^2}{14.65116} \\ &+ \frac{(|-1.34884| - 0.5)^2}{9.34884} \end{aligned}$$

$$\begin{aligned}
&= \frac{(1.34884 - 0.5)^2}{90.34884} + \frac{(1.34884 - 0.5)^2}{57.65116} + \frac{(1.34884 - 0.5)^2}{14.65116} \\
&\quad + \frac{(1.34884 - 0.5)^2}{9.34884} \\
&= \frac{(0.84884)^2}{90.34884} + \frac{(0.84884)^2}{57.65116} + \frac{(0.84884)^2}{14.65116} + \frac{(0.84884)^2}{9.34884} \\
&= \frac{0.72053}{90.34884} + \frac{0.72053}{57.65116} + \frac{0.72053}{14.65116} + \frac{0.72053}{9.34884} \\
&= 0.00798 + 0.0125 + 0.04918 + 0.07707 = 0.14672
\end{aligned}$$

4. How to find χ^2_{critical}

In our snails example there are two rows (hedgerow and woodland). Do not include your 'expected' rows these are part of your calculation. There are two columns (*C. nemoralis* and *C. hortensis*)

Therefore $v = (2 - 1) \times (2 - 1) = 1 \times 1 = 1$

The critical value to test for an association between the snail species and habitat is found in the chi-squared table where $v = 1$, $p = 0.05$, and is $\chi^2_{\text{critical}} = 3.84$.

5. The rule

$\chi^2_{\text{calculated}}$ (0.15) is less than χ^2_{critical} (3.84) at $p = 0.05$ and therefore we do not reject the null hypothesis.

6. What does this mean in real terms?

There is no significant association ($\chi^2_{\text{calculated}}$ 0.15, $p = 0.05$) between the distribution of snail species and the two habitats.

5.5. G tests

5.5.1. G goodness of fit test

BOX 5.7. How to calculate a goodness of fit G test

The full calculation is included in BOX 5.7. in the book.

5.5.2. An $r \times c$ G test for association

Here we are using the data from Example 5.3. where the association between habitat and distribution of snail shell patterns is examined. The observed data are found in Table 5.6. The criteria for using this test are given in 5.3.1. The process of calculating the $G_{\text{calculated}}$ statistic differs critically from the chi-squared. Again the Williams' correction has been used. Where steps are similar to those in BOX 5.4. this is indicated.

BOX 5.8. How to calculate an $r \times c$ G test for association**THIS EXAMPLE****1. Hypotheses to be tested**

BOX 5.4.

2. How to work out expected values

Unlike all other calculations in this chapter there are no expected values used in this G test.

3. How to work out $G_{\text{calculated}}$

i. Use the observed values from Table 5.6.

Table 5.6. Shell colour in *Cepea nemoralis* in coastal and hedgerow habitats

Shell pattern and colour in <i>C. nemoralis</i>					
Habitat	Banded yellow	No bands yellow	Banded pink	No bands pink	Total
Coastal Observed	10	19	5	16	50
Hedgerow Observed	17	8	19	11	55
Total	27	27	24	27	105

ii. First take the natural log of an observed value. Then multiply this by the same observed value.

$$10 \times \ln 10 = 10 \times 2.30258 = 23.02585$$

Repeat this for each observed value

$$19 \times \ln 19 = 19 \times 2.94444 = 55.94434$$

$$5 \times \ln 5 = 5 \times 1.60944 = 8.04719$$

$$16 \times \ln 16 = 16 \times 2.77259 = 44.36142$$

$$17 \times \ln 17 = 17 \times 2.83321 = 48.16463$$

$$8 \times \ln 8 = 8 \times 2.07944 = 16.63553$$

$$19 \times \ln 19 = 19 \times 2.94444 = 55.94434$$

$$11 \times \ln 11 = 11 \times 2.39789 = 26.37685$$

and add all these values together = 278.50015

iii. The grand total is 105. So this step is $105 \times \ln 105 = 105 \times 4.65396 = 488.66584$

iv. For the row totals

$$50 \times \ln 50 = 50 \times 3.91202 = 195.60115$$

$$55 \times \ln 55 = 55 \times 4.0073 = 220.40333$$

For the column totals

$$27 \times \ln 27 = 27 \times 3.29584 = 88.9876$$

$$27 \times \ln 27 = 27 \times 3.29584 = 88.9876$$

$$24 \times \ln 24 = 24 \times 3.17805 = 76.2733$$

$$27 \times \ln 27 = 27 \times 3.29584 = 88.9876$$

The total for rows and columns in this example = 759.24055

v. In this example this step will be $278.50015 + 488.66584 - 759.24055 = 7.92544$

vi. $G = 2 \times 7.92544 = 15.85088$

4. Williams' correction

i. $1/50 + 1/55 = 0.03818$

$0.03818 \times 105 = 4.00909$

$4.00909 - 1 = 3.00909$

ii. $1/27 + 1/27 + 1/24 + 1/27 = 0.15277$

$0.15277 \times 105 = 16.04167$

$16.04167 - 1 = 15.04167$

iii. (i) \times (ii) = 3.00909×15.04167
= 45.26173

iv. $6 \times 105 \times (2 - 1) \times (4 - 1) = 6 \times 105 \times 1 \times 3 = 1890$

v. $W = 1 + \frac{45.26173}{1890} = 1 + 0.02395 = 1.02395$

$G_{\text{calculated}} = \frac{15.85088}{1.02395} = 15.48016$

5. How to find G_{critical}

When $\nu = 3$ and $p = 0.05$, $G_{\text{critical}} = 7.81$.

6. The rule

$G_{\text{calculated}}$ (15.48) is greater than G_{critical} (7.81) at $p = 0.05$ and therefore we reject the null hypothesis.

7. What does this mean in real terms?

BOX 5.4.