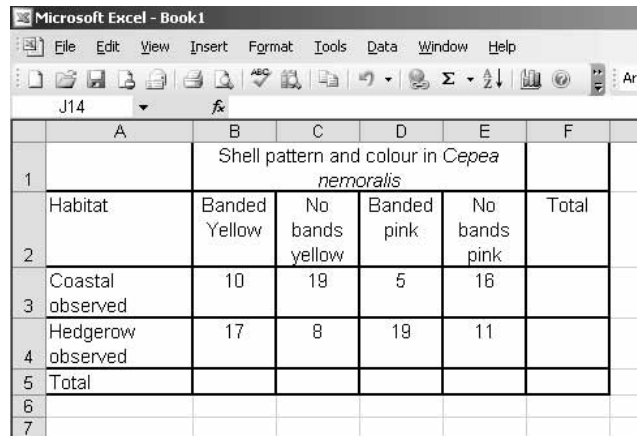


5.5.2. An $r \times c$ G test for association

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

BOX 5.8. How to calculate an $r \times c$ G test for association

Step 1. Enter your data into a spreadsheet. There is no need to leave spaces for the expected values, but make sure that there is room for the row and column totals.



The screenshot shows a Microsoft Excel spreadsheet with the following data:

	A	B	C	D	E	F
1		Shell pattern and colour in <i>Cepea nemoralis</i>				
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
3	Coastal observed	10	19	5	16	
4	Hedgerow observed	17	8	19	11	
5	Total					
6						
7						

Step 2. Calculate the totals.

First calculate the row totals. Into cell f3, type the formula ‘=sum(b3:e3)’, the click on the green tick, or press ‘return’.

	A	B	C	D	E	F
1		Shell pattern and colour in <i>Cepea nemoralis</i>				
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
3	Coastal observed	10	19	5	16	50
4	Hedgerow observed	17	8	19	11	
5	Total					
6						
7						

Drag this formula down as far as cell f5. Hover the cursor over the bottom right-hand corner of cell f3, hold down the left mouse button (the cursor should change from an open horizontal-vertical cross into an addition sign), then drag the cursor down to cell f5.

	A	B	C	D	E	F	G
1		Shell pattern and colour in <i>Cepea nemoralis</i>					
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total	
3	Coastal observed	10	19	5	16	50	
4	Hedgerow observed	17	8	19	11	55	
5	Total					0	
6							
7							

Note that cell f5 has '0' in it, because there is nothing in the cells whose contents are to be added up.

Now we total the columns. In cell b5, type ' $=b3 + b4$ '. Click on the green tick, or press 'return'.

	A	B	C	D	E	F	G
1		Shell pattern and colour in <i>Cepea nemoralis</i>					
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total	
3	Coastal observed	10	19	5	16	50	
4	Hedgerow observed	17	8	19	11	55	
5	Total	27				27	
6							
7							

(Note that cell f5 has changed to reflect what we have just done.)

Drag this across into cells c5 to e5.

	A	B	C	D	E	F	G
1		Shell pattern and colour in <i>Cepea nemoralis</i>					
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total	
3	Coastal observed	10	19	5	16	50	
4	Hedgerow observed	17	8	19	11	55	
5	Total	27	27	24	27	105	
6							
7							

(We could have done this the other way round by calculating the column totals, then the row totals.)

Step 3. Calculate $\ln(o)$ for all these values, where 'o' is an observed value, and 'ln' means 'the natural logarithm of'. This is done by the Excel function 'ln'. The easiest way to do this is to create another table of identical dimensions to the one we already have, and populate it with the numbers we need.

Highlight the cells a1 to f5 (place the cursor in cell a1, hold down the left mouse button, drag the cursor to cell f5, and release the button.) Got to 'Edit', 'Copy' (or hold down 'ctrl' while typing 'c'). This copies the selection to the clipboard. Select an appropriate location for the new table (we shall use the block with top left cell at a8), click in the cell, and go to 'Edit', 'Paste' (or hold down the 'ctrl' key while typing 'v').

Microsoft Excel - G text rxc association						
File Edit View Insert Format Tools Data Window Help						
A8						
A	B	C	D	E	F	G
1	Shell pattern and colour in <i>Cepea nemoralis</i>					
	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
2	Coastal observed	10	19	5	16	50
3	Hedgerow observed	17	8	19	11	55
4	Total	27	27	24	27	105
5						
6						
7						
8	Shell pattern and colour in <i>Cepea nemoralis</i>					
	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
9	Coastal observed	10	19	5	16	50
10	Hedgerow observed	17	8	19	11	55
11	Total	27	27	24	27	105
12						
13						
14						
15						

Make sure we remember what this table is about to contain by typing 'o ln(o)' in cell a8. (If you want to, you can make it more conspicuous by making it bold and changing its colour.)

The screenshot shows a Microsoft Excel spreadsheet with two identical contingency tables. The first table is in rows 1-5, and the second is in rows 8-12. The second table's cell A8 contains the text 'o ln(o)'. The data in both tables is as follows:

Shell pattern and colour in <i>Cepea 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

Remove all the numbers from this new table, because we are going to replace them with values of $o \ln(o)$. Highlight all the data cells, and press 'delete' (or go to 'Edit', 'Clear', 'Contents').

Microsoft Excel - G test rxc association

File Edit View Insert Format Tools Data Window Help

B10

	A	B	C	D	E	F
1		Shell pattern and colour in <i>Cepea nemoralis</i>				
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
3	Coastal observed	10	19	5	16	50
4	Hedgerow observed	17	8	19	11	55
5	Total	27	27	24	27	105
6						
7						
8	$o \ln(o)$	Shell pattern and colour in <i>Cepea nemoralis</i>				
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
10	Coastal observed					
11	Hedgerow observed					
12	Total					
13						

Now calculate the values of $o \ln(o)$. We can do this by typing a formula into one cell and dragging it into all the others, because our grids are the same shape and size. In cell b10, type ' $=b3*\ln(b3)$ ', then click on the green tick or press 'return'.

Microsoft Excel - G test $r \times c$ association						
File Edit View Insert Format Tools Data Window Help						
B10 $=B3*LN(B3)$						
	A	B	C	D	E	F
1		Shell pattern and colour in <i>Cepea nemoralis</i>				
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
3	Coastal observed	10	19	5	16	50
4	Hedgerow observed	17	8	19	11	55
5	Total	27	27	24	27	105
6						
7						
8	$o \ln(o)$	Shell pattern and colour in <i>Cepea</i>				
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
10	Coastal observed	23.0258509				
11	Hedgerow observed					
12	Total					
13						

Now drag this cell across the rest of the table. (You may have to drag across a row, and then drag the row down.)

Microsoft Excel - G test rxc association

File Edit View Insert Format Tools Data Window Help

B10 $=B3*LN(B3)$

	A	B	C	D	E	F	G
1		Shell pattern and colour in <i>Cepea nemoralis</i>					
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total	
3	Coastal observed	10	19	5	16	50	
4	Hedgerow observed	17	8	19	11	55	
5	Total	27	27	24	27	105	
6							
7							
8	$\ln(o)$	Shell pattern and colour in <i>Cepea nemoralis</i>					
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total	
10	Coastal observed	23.0258509	55.9443	8.04719	44.3614	195.601	
11	Hedgerow observed	48.1646268	16.6355	55.9443	26.3768	220.403	
12	Total	88.9875954	88.9876	76.2733	88.9876	488.666	
13							
14							
15							

Following the outline in the book, we next add together all the values of $\ln(o)$ for the individual measurements, and place them in a convenient (labelled) cell, say b14. Type in the formula ' $=\text{sum}(b10:e11)$ ', then click on the green tick or press 'return'.

7							
8	$\ln(o)$	Shell pattern and colour in <i>Cepea nemoralis</i>					
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total	
10	Coastal observed	23.0258509	55.9443	8.04719	44.3614	195.601	
11	Hedgerow observed	48.1646268	16.6355	55.9443	26.3768	220.403	
12	Total	88.9875954	88.9876	76.2733	88.9876	488.666	
13							
14	measurements	278.5001484					
15							
16							

$\ln(o)$ for the grand total is stored in cell f12, so we simply note that it is there: we will need it soon. The next thing to do is to add the values of $\ln(o)$ for the rows and columns together, and put them somewhere

convenient (b15). Type the formula ‘=sum(b12:e15) + f10 + f11’ into cell b15, then click on the green tick or press ‘return’.

7						
8	o ln(o)	Shell pattern and colour in <i>Cepea</i>				
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
10	Coastal observed	23.0258509	55.9443	8.04719	44.3614	195.601
11	Hedgerow observed	48.1646268	16.6355	55.9443	26.3768	220.403
12	Total	88.9875954	88.9876	76.2733	88.9876	488.666
13						
14	measurements	278.5001484				
15	rows & columns	759.2405535				
16						
17						

Taking values of $o \ln(o)$, we now need to find $G = 2 \times (\text{measurements} + \text{grand total} - \text{rows \& columns})$. Using cell b16, type ‘=2*(b14 + f12 – b15)’, then click on the green tick or press ‘return’.

7						
8	o ln(o)	Shell pattern and colour in <i>Cepea</i>				
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
10	Coastal observed	23.0258509	55.9443	8.04719	44.3614	195.601
11	Hedgerow observed	48.1646268	16.6355	55.9443	26.3768	220.403
12	Total	88.9875954	88.9876	76.2733	88.9876	488.666
13						
14	measurements	278.5001484				
15	rows & columns	759.2405535				
16	G	15.85086334				
17						

To find the Williams' correction, first work out 1/each row total, and add these values together. Multiply by the grand total. Subtract 1. To do this, use a formula in a convenient cell, say, e15. The formula is ' $= (1/f3 + 1/f4) * f5 - 1$ '.

Microsoft Excel - G test rxc association						
File Edit View Insert Format Tools Data Window Help						
Σ						
E15 $= (1/F3+1/F4)*F5-1$						
	A	B	C	D	E	F
1		Shell pattern and colour in <i>Cepea nemoralis</i>				
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
3	Coastal observed	10	19	5	16	50
4	Hedgerow observed	17	8	19	11	55
5	Total	27	27	24	27	105
6						
7						
8	$\ln(o)$	Shell pattern and colour in <i>Cepea</i>				
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
10	Coastal observed	23.0258509	55.9443	8.04719	44.3614	195.601
11	Hedgerow observed	48.1646268	18.6355	55.9443	26.3768	220.403
12	Total	88.9875954	88.9876	76.2733	88.9876	488.668
13						
14	measurements	278.5001484		Williams		
15	rows & columns	759.2405535		rows	3.009091	
16	G	15.85086334				
17						

Do the same thing for the columns. The formula is $' = (1/b5 + 1/c5 + 1/d5 + 1/e5) * f5 - 1'$.

Microsoft Excel - G test $r \times c$ association

File Edit View Insert Format Tools Data Window Help

Σ ↓

E16 $= (1/B5+1/C5+1/D5+1/E5)*F5-1$

	A	B	C	D	E	F	G
1		Shell pattern and colour in <i>Cepea nemoralis</i>					
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total	
3	Coastal observed	10	19	5	16	50	
4	Hedgerow observed	17	8	19	11	55	
5	Total	27	27	24	27	105	
6							
7							
8	$\ln(\phi)$	Shell pattern and colour in <i>Cepea</i>					
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total	
10	Coastal observed	23.0258509	55.9443	8.04719	44.3614	195.601	
11	Hedgerow observed	48.1646268	16.6355	55.9443	26.3768	220.403	
12	Total	88.9875954	88.9876	76.2733	88.9876	488.668	
13							
14	measurements	278.5001484		Williams			
15	rows & columns	759.2405535		rows	3.009091		
16	G	15.85086334		columns	15.04167		
17							

Multiply these two together: the formula is ' $=e15*e16$ '.

Microsoft Excel - G test rxc association						
File Edit View Insert Format Tools Data Window Help						
E17 =E15*E16						
	A	B	C	D	E	F
1		Shell pattern and colour in <i>Cepea nemoralis</i>				
2	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
3	Coastal observed	10	19	5	16	50
4	Hedgerow observed	17	8	19	11	55
5	Total	27	27	24	27	105
6						
7						
8		Shell pattern and colour in <i>Cepea nemoralis</i>				
9	Habitat	Banded Yellow	No bands yellow	Banded pink	No bands pink	Total
10	Coastal observed	23.0258509	55.9443	8.04719	44.3614	195.601
11	Hedgerow observed	48.1646268	16.6355	55.94434	26.3768	220.403
12	Total	88.9875954	88.9876	76.27329	88.9876	488.666
13						
14	measurements	278.5001484		Williams		
15	rows & columns	759.2405535		rows	3.009091	
16	G	15.85086334		columns	15.04167	
17				rows x cols	45.26174	
18				$6n(r-1)(c-1)$	1890	
19						

Now we calculate $6n(\text{rows} - 1)(\text{columns} - 1)$, where n is the total number of observations, and 'rows' and 'columns' are the numbers of rows and columns. Use the formula ' $=6*f5*(2-1)*(4-1)$ '.

13					
14	measurements	278.5001484		Williams	
15	rows & columns	759.2405535		rows	3.009091
16	G	15.85086334		columns	15.04167
17				rows x cols	45.26174
18				$6n(r-1)(c-1)$	1890
19					

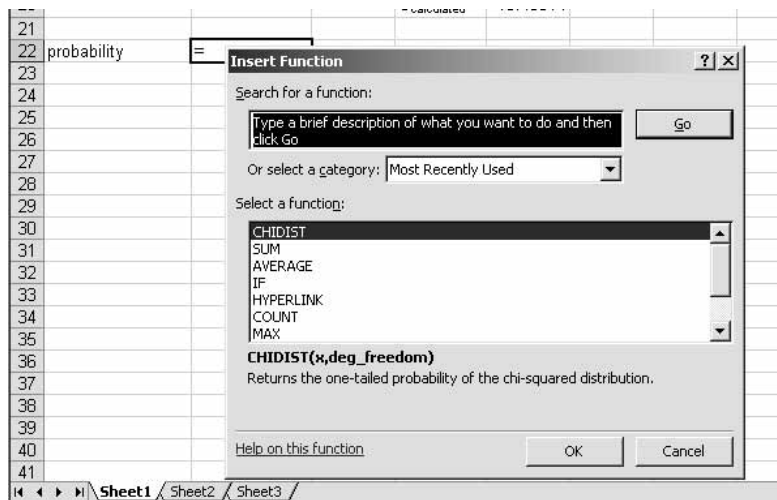
The Williams' correction factor is $W = 1 + (\text{rows} \times \text{cols}) / (6n(r - 1)(c - 1))$. Use the formula ' $=1 + e17/e18$ '.

13					
14	measurements	278.5001484		Williams	
15	rows & columns	759.2405535		rows	3.009091
16	G	15.85086334		columns	15.04167
17				rows x cols	45.26174
18				$6n(r-1)(c-1)$	1890
19				W	1.023948
20					

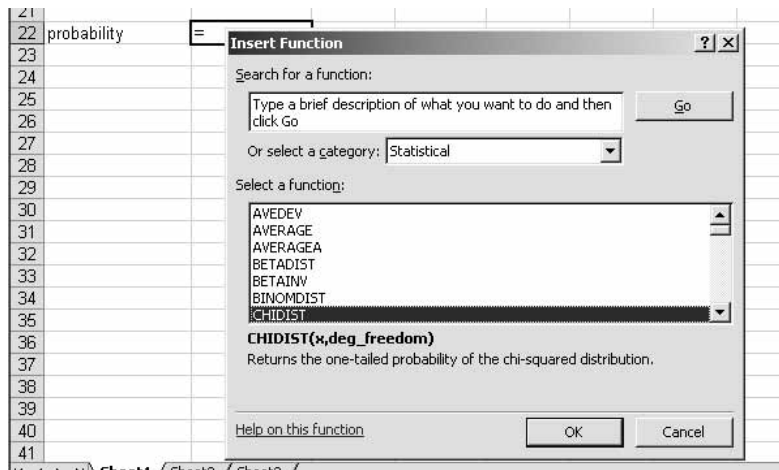
$G_{\text{calculated}} = G/W$. The formula is ' $= b16/e19$ '.

14	measurements	278.5001484	Williams	
15	rows & columns	759.2405535	rows	3.009091
16	G	15.85086334	columns	15.04167
17			rows x cols	45.26174
18			$6n(r-1)(c-1)$	1890
19			W	1.023948
20			$G_{\text{calculated}}$	15.48014
21				
22				

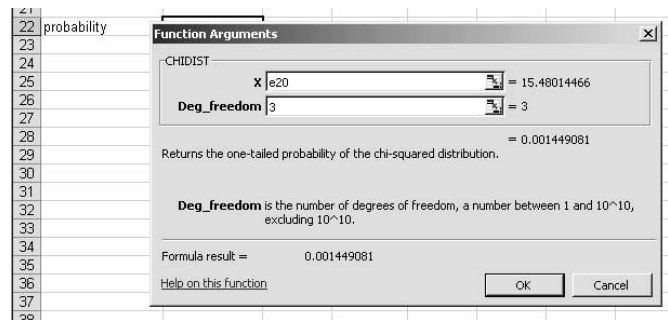
Step 4. We now use the chi-squared distribution to find the probability of not rejecting the null hypothesis. Using a suitable free cell, go to 'Insert', 'Function'.



Select the category 'Statistical', and click on 'CHIDIST'.



Click on 'OK'. Enter the cell where G is stored and the number of **degrees of freedom** (e20 and 3 in this case).



Click on 'OK'.

14	measurements	278.5001484	Williams	
15	rows & columns	759.2405535	rows	3.009091
16	G	15.85086334	columns	15.04167
17			rows x cols	45.26174
18			$6n(r-1)(c-1)$	1890
19			W	1.023948
20			$G_{\text{calculated}}$	15.48014
21				
22	probability	0.001449081		
23				
24				

This gives us the probability of not rejecting the null hypothesis, and this is very small ($p=0.0015$). There is a highly significant association ($G = 15.48$, $p = 0.0015$) between the distribution of shell patterns and habitat (coastal and hedgerow) of *Cepea nemoralis*.