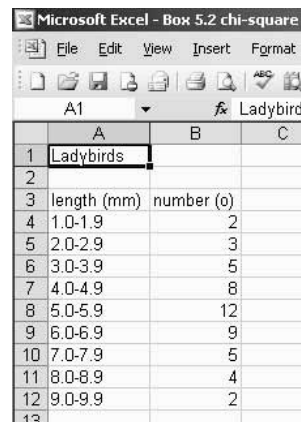


### 5.1.3. How to check if your data have a normal distribution using a goodness of fit chi-squared test

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

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

**Step 1.** Enter the data into the Excel spreadsheet using suitable column headings. The '(o)' means 'observed' – to distinguish it from the **expected** values we are going to calculate later.



The screenshot shows a Microsoft Excel spreadsheet titled "Microsoft Excel - Box 5.2 chi-square". The spreadsheet has three columns: A, B, and C. The data is as follows:

	A	B	C
1	Ladybirds		
2			
3	length (mm)	number (o)	
4	1.0-1.9		2
5	2.0-2.9		3
6	3.0-3.9		5
7	4.0-4.9		8
8	5.0-5.9		12
9	6.0-6.9		9
10	7.0-7.9		5
11	8.0-8.9		4
12	9.0-9.9		2
13			

**Step 2.** Calculate the expected values. To do this, we first need to find the total number of ladybirds, the **mean** length and the **standard deviation** of the length. The total number can be found by using the formula '=sum(b4:b12)' in a suitable cell (make sure it is labelled so that you know what the number means). Type in the label, click in the cell where you want the number to appear, and type the formula into the formula bar.

	A	B	C
1	Ladybirds		
2			
3	length (mm)	number (o)	
4	1.0-1.9	2	
5	2.0-2.9	3	
6	3.0-3.9	5	
7	4.0-4.9	8	
8	5.0-5.9	12	
9	6.0-6.9	9	
10	7.0-7.9	5	
11	8.0-8.9	4	
12	9.0-9.9	2	
13			
14	total	=sum(b4:b12)	
15			

Click on the green tick, or press 'return'.

	A	B	C
1	Ladybirds		
2			
3	length (mm)	number (o)	
4	1.0-1.9	2	
5	2.0-2.9	3	
6	3.0-3.9	5	
7	4.0-4.9	8	
8	5.0-5.9	12	
9	6.0-6.9	9	
10	7.0-7.9	5	
11	8.0-8.9	4	
12	9.0-9.9	2	
13			
14	total	50	
15			
16			

We now need the centre of each range (class). This is found by taking the top and bottom of the range, adding them together, and dividing by 2. Thus, for the first range, the top is 1.9 and the bottom is 1.0, so the mid-point is 1.45. The others can be found in a similar fashion.

	A	B	C
1	Ladybirds		
2			
3	length (mm)	number (o)	midpoint (mm)
4	1.0-1.9	2	1.45
5	2.0-2.9	3	2.45
6	3.0-3.9	5	3.45
7	4.0-4.9	8	4.45
8	5.0-5.9	12	5.45
9	6.0-6.9	9	6.45
10	7.0-7.9	5	7.45
11	8.0-8.9	4	8.45
12	9.0-9.9	2	9.45
13			
14	total	50	
15			

Calculating the **mean** and **standard deviation** is messy if we don't have the actual values. Fortunately, they have already been calculated in Chapter 3: enter them in cells b15 and b16.

	A	B	C
1	Ladybirds		
2			
3	length (mm)	number (o)	midpoint (mm)
4	1.0-1.9	2	1.45
5	2.0-2.9	3	2.45
6	3.0-3.9	5	3.45
7	4.0-4.9	8	4.45
8	5.0-5.9	12	5.45
9	6.0-6.9	9	6.45
10	7.0-7.9	5	7.45
11	8.0-8.9	4	8.45
12	9.0-9.9	2	9.45
13			
14	total	50	
15	mean	5.08	
16	st. dev.	1.94	
17			

Next, we use Excel to calculate the frequency for a normal distribution at each of these mid-points. In cell d4, type the formula '= normdist (c4,\$b\$15,\$b\$16,false)'. In this formula:

- 'c4' is a relative reference to the cell immediately to the left, which contains the value at which we want the frequency: this will change as we drag the formula down the column;
- '\$b\$15' is an absolute reference to the cell containing the mean of the distribution: this will NOT change as we drag the formula down the column;

'\$b\$16' is another absolute reference, this time to the cell containing the standard deviation of the distribution;  
 'false' is a logical argument telling the computer that we want the probability mass function, not the cumulative distribution function.

	A	B	C	D
1	Ladybirds			
2				
3	length (mm)	number (o)	midpoint (mm)	probability
4	1.0-1.9	2	1.45	=normdist(c4,\$b\$15,\$b\$16,false)
5	2.0-2.9	3	2.45	
6	3.0-3.9	5	3.45	
7	4.0-4.9	8	4.45	
8	5.0-5.9	12	5.45	
9	6.0-6.9	9	6.45	
10	7.0-7.9	5	7.45	
11	8.0-8.9	4	8.45	
12	9.0-9.9	2	9.45	
13				
14	total	50		
15	mean	5.08		
16	st. dev.	1.94		

Click on the green tick, or press 'return'.

	A	B	C	D	E
1	Ladybirds				
2					
3	length (mm)	number (o)	midpoint (mm)	probability	
4	1.0-1.9	2	1.45	=normdist(c4,\$b\$15,\$b\$16,false)	
5	2.0-2.9	3	2.45		
6	3.0-3.9	5	3.45		
7	4.0-4.9	8	4.45		
8	5.0-5.9	12	5.45		
9	6.0-6.9	9	6.45		
10	7.0-7.9	5	7.45		
11	8.0-8.9	4	8.45		
12	9.0-9.9	2	9.45		
13					
14	total	50			
15	mean	5.08			
16	st. dev.	1.94			
17					

Drag the formula down into cells d5 to d12 to find the other probabilities.

	A	B	C	D	E
1	Ladybirds				
2					
3	length (mm)	number (o)	midpoint (mm)	probability	
4	1.0-1.9	2	1.45	0.0357145	
5	2.0-2.9	3	2.45	0.0820401	
6	3.0-3.9	5	3.45	0.1444821	
7	4.0-4.9	8	4.45	0.1950781	
8	5.0-5.9	12	5.45	0.2019341	
9	6.0-6.9	9	6.45	0.1602572	
10	7.0-7.9	5	7.45	0.0975061	
11	8.0-8.9	4	8.45	0.0454834	
12	9.0-9.9	2	9.45	0.016266	
13					
14	total	50			
15	mean	5.08			
16	st. dev.	1.94			
17					

The final step is to calculate the expected numbers. To do this, we multiply the probabilities by the total number of ladybirds (as stored in cell b14). Put a suitable title (e.g. ‘number (e)’ – where the ‘(e)’ means ‘expected’) in cell e3, then click in cell e4. Type in the formula ‘=d4\*\$b\$14’, where ‘d4’ is a relative reference to the cell to the left and ‘\$b\$14’ is an absolute reference to the cell containing the total number of ladybirds.

	A	B	C	D	E	F
1	Ladybirds					
2						
3	length (mm)	number (o)	midpoint (mm)	probability	number (e)	
4	1.0-1.9	2	1.45	0.0357145	=d4*\$b\$14	
5	2.0-2.9	3	2.45	0.0820401		
6	3.0-3.9	5	3.45	0.1444821		
7	4.0-4.9	8	4.45	0.1950781		
8	5.0-5.9	12	5.45	0.2019341		
9	6.0-6.9	9	6.45	0.1602572		
10	7.0-7.9	5	7.45	0.0975061		
11	8.0-8.9	4	8.45	0.0454834		
12	9.0-9.9	2	9.45	0.016266		
13						
14	total	50				
15	mean	5.08				
16	st. dev.	1.94				
17						

Click on the green tick, or press ‘return’.

	A	B	C	D	E	F
1	Ladybirds					
2						
3	length (mm)	number (o)	midpoint (mm)	probability	number (e)	
4	1.0-1.9	2	1.45	0.0357145	1.7857263	
5	2.0-2.9	3	2.45	0.0620401		
6	3.0-3.9	5	3.45	0.1444821		
7	4.0-4.9	8	4.45	0.1950781		
8	5.0-5.9	12	5.45	0.2019341		
9	6.0-6.9	9	6.45	0.1602572		
10	7.0-7.9	5	7.45	0.0975061		
11	8.0-8.9	4	8.45	0.0454834		
12	9.0-9.9	2	9.45	0.016266		
13						
14	total	50				
15	mean	5.08				
16	st. dev.	1.94				
17						

Now drag the formula down to cells e5 –e12.

	A	B	C	D	E	F
1	Ladybirds					
2						
3	length (mm)	number (o)	midpoint (mm)	probability	number (e)	
4	1.0-1.9	2	1.45	0.0357145	1.7857263	
5	2.0-2.9	3	2.45	0.0620401	4.1020046	
6	3.0-3.9	5	3.45	0.1444821	7.2241045	
7	4.0-4.9	8	4.45	0.1950781	9.7539051	
8	5.0-5.9	12	5.45	0.2019341	10.096705	
9	6.0-6.9	9	6.45	0.1602572	8.0128578	
10	7.0-7.9	5	7.45	0.0975061	4.8753055	
11	8.0-8.9	4	8.45	0.0454834	2.2741697	
12	9.0-9.9	2	9.45	0.016266	0.8132995	
13						
14	total	50				
15	mean	5.08				
16	st. dev.	1.94				
17						

**Step 3.** Work out goodness of fit chi-squared. Unless you are a maths whiz, it is best to do this in small steps.

First, note that several of the expected values are less than 5, which violates one of the criteria for a chi-squared test. We need to group together some of the classes to reduce this problem, and we choose to combine the first two and the last two classes.

Further down the spreadsheet, we can copy most of the important parts of our table to make a new table. Set up the new lengths as shown:

13		
14	total	50
15	mean	5.08
16	st. dev.	1.94
17		
18		
19		
20	length (mm)	
21	1.0-2.9	
22	3.0-3.9	
23	4.0-4.9	
24	5.0-5.9	
25	6.0-6.9	
26	7.0-7.9	
27	8.0-9.9	
28		

(Note that the first and last class are now twice the size of the others. This is only acceptable as a step within a calculation such the goodness of fit chi-squared. You should use the usual equal-sized classes at all other times.)

The data can be copied down for the middle classes by copying the numbers from their original locations. Either type them in by hand, or use formulae of the form ‘=b6’ in cell b22, and drag it down into cells b23 to b26.

17			
18			
19			
20	length (mm)	number (o)	number (e)
21	1.0-2.9		
22	3.0-3.9	5	250
23	4.0-4.9	8	400
24	5.0-5.9	12	600
25	6.0-6.9	9	450
26	7.0-7.9	5	250
27	8.0-9.9		
28			
29			
30			

The combined classes will need formulae: in cell b21, type ‘=b4 + b5’ to combine the observed values for the first two classes. The other combination can be performed similarly.

18			
19			
20	length (mm)	number (o)	number (e)
21	1.0-2.9	5	5.887730942
22	3.0-3.9	5	7.224104502
23	4.0-4.9	8	9.753905109
24	5.0-5.9	12	10.09670492
25	6.0-6.9	9	8.012857761
26	7.0-7.9	5	4.875305492
27	8.0-9.9	6	3.087469234
28			
29			
30			

Unless you are a maths whiz, it is best to do the next bit in small steps.

First, find the difference between the observed and expected values. Enter '= b21-c21' into cell d21, say, then drag the result down into cells c22 to d27.

18					
19					
20	length (mm)	number (o)	number (e)	obs-exp	
21	1.0-2.9	5	5.887730942	-0.8877309	
22	3.0-3.9	5	7.224104502	-2.2241045	
23	4.0-4.9	8	9.753905109	-1.7539051	
24	5.0-5.9	12	10.09670492	1.9032951	
25	6.0-6.9	9	8.012857761	0.9871422	
26	7.0-7.9	5	4.875305492	0.1246945	
27	8.0-9.9	6	3.087469234	2.9125308	
28					
29					

Next, square these values by entering '= d21^2' into cell e21, say, and drag the result down into cells e22 to e27.

18					
19					
20	length (mm)	number (o)	number (e)	obs-exp	(obs-exp) <sup>2</sup>
21	1.0-2.9	5	5.887730942	-0.8877309	0.7880662
22	3.0-3.9	5	7.224104502	-2.2241045	4.9466408
23	4.0-4.9	8	9.753905109	-1.7539051	3.0761831
24	5.0-5.9	12	10.09670492	1.9032951	3.6225322
25	6.0-6.9	9	8.012857761	0.9871422	0.9744498
26	7.0-7.9	5	4.875305492	0.1246945	0.0155487
27	8.0-9.9	6	3.087469234	2.9125308	8.4828355
28					
29					

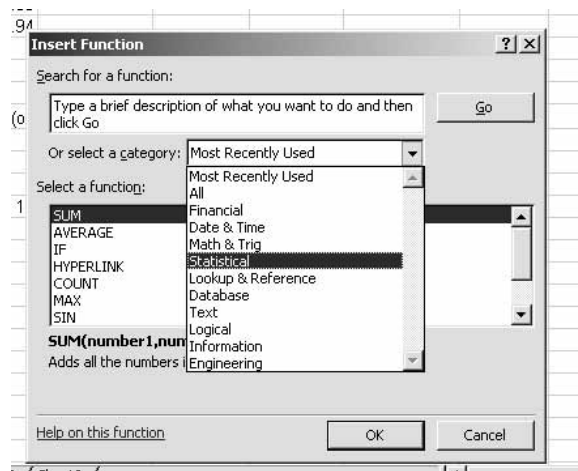
Then divide this by the expected number of ladybirds by entering '= e21/c21' into cell f21, say, and dragging down into cells f22 to f27.

18						
19						
20	length (mm)	number (o)	number (e)	obs-exp	(obs-exp) <sup>2</sup>	(obs-exp) <sup>2</sup> /exp
21	1.0-2.9	5	5.887730942	-0.8877309	0.7880662	0.133848885
22	3.0-3.9	5	7.224104502	-2.2241045	4.9466408	0.684741041
23	4.0-4.9	8	9.753905109	-1.7539051	3.0761831	0.315379645
24	5.0-5.9	12	10.09670492	1.9032951	3.6225322	0.358783601
25	6.0-6.9	9	8.012857761	0.9871422	0.9744498	0.12161077
26	7.0-7.9	5	4.875305492	0.1246945	0.0155487	0.003189281
27	8.0-9.9	6	3.087469234	2.9125308	8.4828355	2.74750445
28						
29						

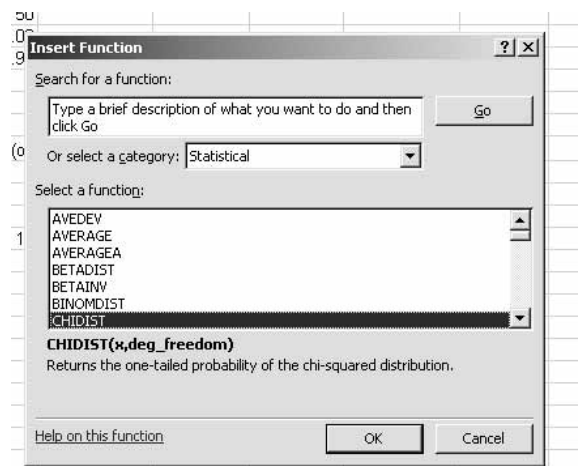
Finally, find chi-squared by adding together all the values of  $(\text{Obs} - \text{Exp})^2 / \text{Exp}$  by entering '= SUM(f21:f27)' into cell f29, say.

	length (mm)	number (o)	number (e)	obs-exp	(obs-exp) <sup>2</sup>	(obs-exp) <sup>2</sup> /exp
20	1.0-2.9	5	5.887730942	-0.8877309	0.7880662	0.133848885
22	3.0-3.9	5	7.224104502	-2.2241045	4.9466408	0.684741041
23	4.0-4.9	8	9.753905109	-1.7539051	3.0761831	0.315379645
24	5.0-5.9	12	10.09670492	1.9032951	3.6225322	0.358783601
25	6.0-6.9	9	8.012857761	0.9871422	0.9744498	0.12161077
26	7.0-7.9	5	4.875305492	0.1246945	0.0155487	0.003189281
27	8.0-9.9	6	3.087469234	2.9125308	8.4828355	2.74750445
28						
29					chi-sq.	4.365057674
30						
31						

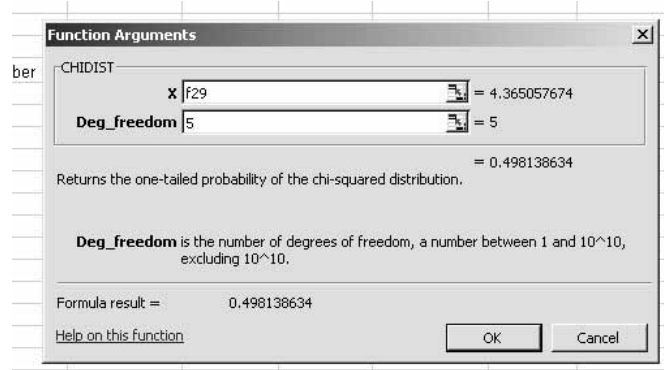
**Step 4.** Perform the test. We are going to work out the probability. To do this, select another cell, f31, say, and start the function wizard ('Insert', 'Function'). From the function category, select 'Statistical'.



From the function name list select 'CHIDIST'.



Click on 'OK'. In the following dialogue box, in the first box enter the cell location where the value of chi-squared is held (you can enter the value of chi-squared manually, but entering the cell number eliminates the possibility of typing mistakes). In the second box, enter the number of **degrees of freedom** (5 in this case).



Now hit 'OK'. The number that appears in the cell is the probability that the null hypothesis will not be rejected.

18						
19						
20	length (mm)	number (o)	number (e)	obs-exp	(obs-exp) <sup>2</sup>	(obs-exp) <sup>2</sup> /exp
21	1.0-2.9	5	5.887730942	-0.8877309	0.7880662	0.133848885
22	3.0-3.9	5	7.224104502	-2.2241045	4.9466408	0.684741041
23	4.0-4.9	8	9.753905109	-1.7539051	3.0761831	0.315379645
24	5.0-5.9	12	10.09670492	1.9032951	3.6225322	0.358783601
25	6.0-6.9	9	8.012857761	0.9871422	0.9744498	0.12161077
26	7.0-7.9	5	4.875305492	0.1246945	0.0155487	0.003189281
27	8.0-9.9	6	3.087469234	2.9125308	8.4828355	2.74750445
28						
29					chi-sq.	4.365057674
30						
31					probability	0.498138634
32						

In this case,  $p$  is large (0.49) and above the threshold of  $p = 0.05$ , so we do not reject the null hypothesis. There is no significant difference ( $\chi^2_{\text{calculated}} = 4.37$ ,  $p = 0.498$ ) 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.