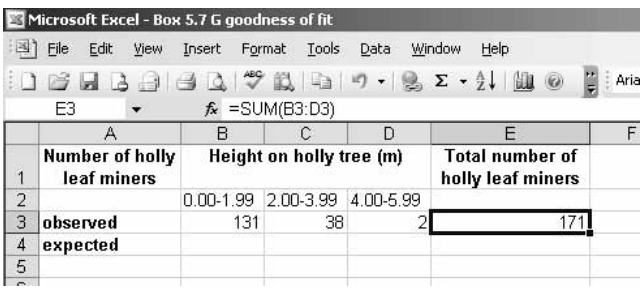
5.5.1. G goodness of fit test

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

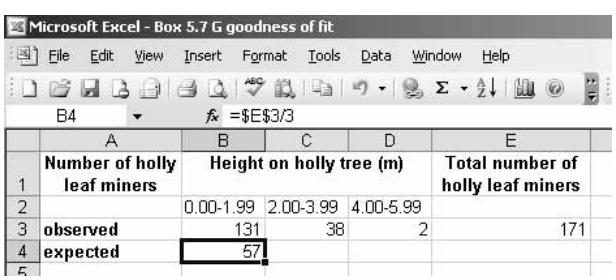
BOX 5.7. How to calculate a goodness of fit *G* test

Step 1. Enter the data into the spreadsheet using suitable headings. The format of a results table is usually OK, and we can get Excel to calculate the total number of observations for us.



	A	B	C	D	E	F
1	Number of holly leaf miners	Height on holly tree (m)			Total number of holly leaf miners	
2		0.00-1.99	2.00-3.99	4.00-5.99		
3	observed	131	38	2	171	
4	expected					
5						

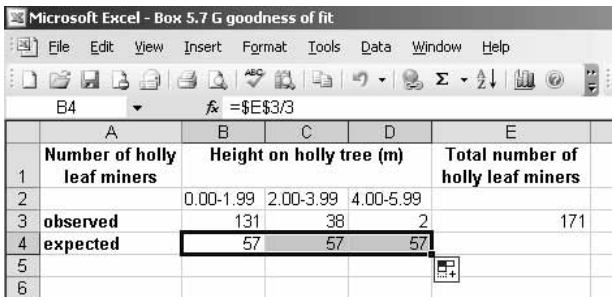
Step 2. Calculate the expected values. In this case, we are testing against a random distribution, so we expect the same number of holly leaf miners in each height range. We have three ranges, and a total of 171 holly leaf miners (recorded in cell e3 in the example above), so we expect 171/3 holly leaf miners in each range. Let Excel do the calculation for you by entering '=e3/3' into cell b4. (The $ signs mean an absolute (rather than relative) reference to the contents of cell e3. We use this because we are going to drag the formula across into the other three cells.)



	A	B	C	D	E
1	Number of holly leaf miners	Height on holly tree (m)			Total number of holly leaf miners
2		0.00-1.99	2.00-3.99	4.00-5.99	
3	observed	131	38	2	171
4	expected	57			
5					

Place the cursor (usually a thick vertical–horizontal cross) on the bottom right-hand corner of cell b4. It should change to a thin horizontal–vertical

cross. Hold down the left mouse button and drag the cursor across until cells c4 and d4 are highlighted, then release the mouse button.

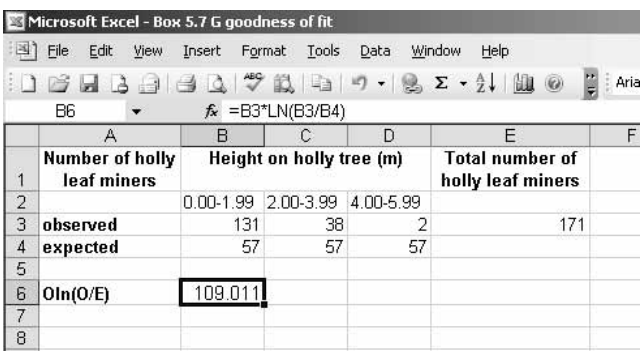


	A	B	C	D	E
1	**Number of holly leaf miners**	**Height on holly tree (m)**			**Total number of holly leaf miners**
2		0.00-1.99	2.00-3.99	4.00-5.99	
3	**observed**	131	38	2	171
4	**expected**	57	57	57	
5					
6					

We now have our 'expected' values in the table.

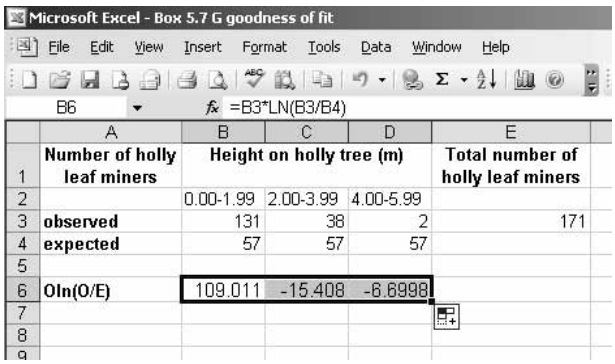
Step 3. Perform the *G* goodness of fit test and decide what the result means.

First, we need to calculate $O \times \ln(O/E)$, where O and E are the observed and expected values for each category. Create a new row for this (we shall use row 6), and in cell b6 enter the formula '=b3*ln(b3/b4)'. In this formula, the function 'ln' gives the logarithm to base e, known as a 'natural logarithm'. Click on the green tick, or press 'return'.



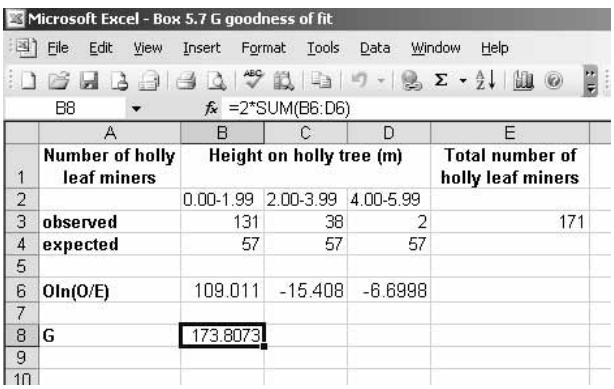
	A	B	C	D	E	F
1	**Number of holly leaf miners**	**Height on holly tree (m)**			**Total number of holly leaf miners**	
2		0.00-1.99	2.00-3.99	4.00-5.99		
3	**observed**	131	38	2	171	
4	**expected**	57	57	57		
5						
6	**Oln(O/E)**	109.011				
7						
8						

Drag this across into cells c6 and d6. Make sure that cell b6 is highlighted, and hover the cursor over the bottom right-hand corner. It should change from an open horizontal–vertical cross into an addition sign. Hold down the left mouse button, and drag the cursor across. When you get to cell d6, release the mouse button.



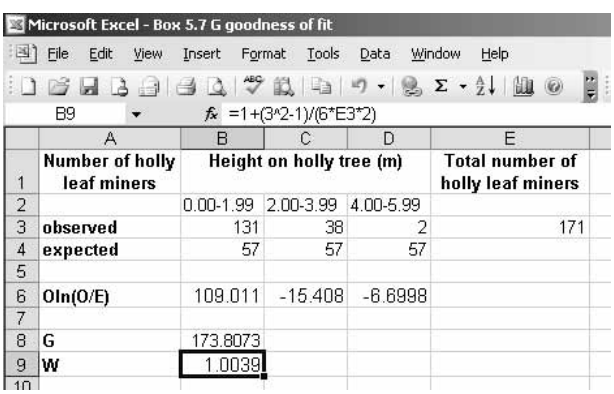
	A	B	C	D	E
1	Number of holly leaf miners	Height on holly tree (m)			Total number of holly leaf miners
2		0.00-1.99	2.00-3.99	4.00-5.99	
3	observed	131	38	2	171
4	expected	57	57	57	
5					
6	OIn(O/E)	109.011	-15.408	-6.6998	
7					
8					
9					

Add these values together and multiply by 2 to give the value of G. Use the formula '$=2$*sum(b6:d6)' to do this.



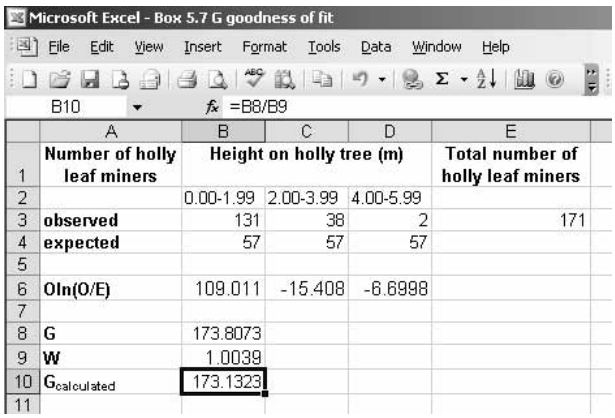
	A	B	C	D	E
1	Number of holly leaf miners	Height on holly tree (m)			Total number of holly leaf miners
2		0.00-1.99	2.00-3.99	4.00-5.99	
3	observed	131	38	2	171
4	expected	57	57	57	
5					
6	OIn(O/E)	109.011	-15.408	-6.6998	
7					
8	G	173.8073			
9					
10					

Next, we apply the Williams' correction. First, calculate $W = 1 + (a^2 - 1)/6nv$, where a is the number of categories (3 in this case), n is the total number of observations (stored in cell e3), and v is the number of degrees of freedom (2 in this case). Use the formula '$=1+(3\text{^}2-1)/(6$*e3*$2)$'.



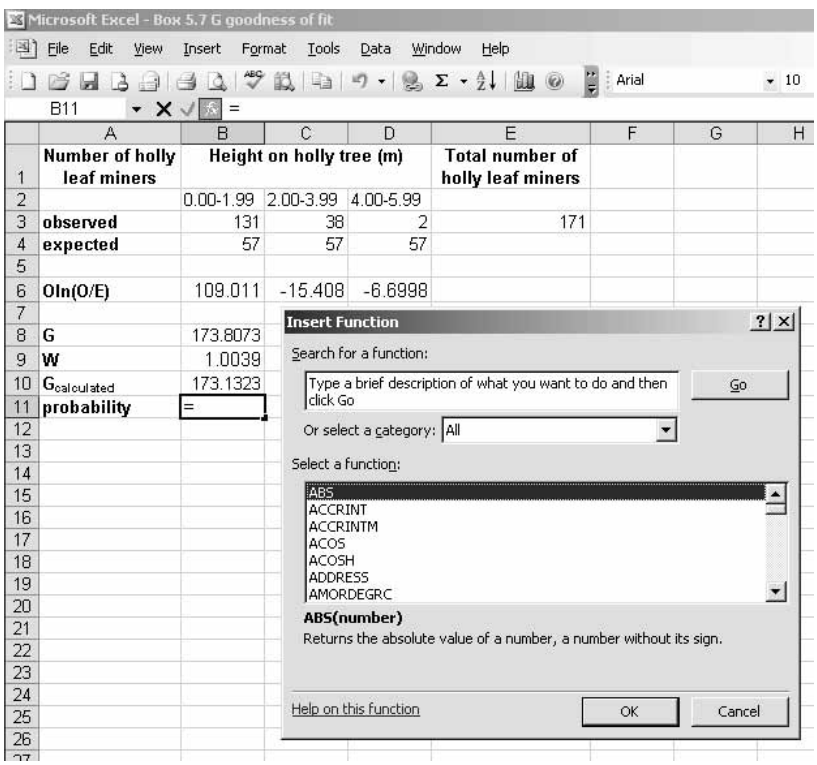
	A	B	C	D	E
1	Number of holly leaf miners	Height on holly tree (m)			Total number of holly leaf miners
2		0.00-1.99	2.00-3.99	4.00-5.99	
3	observed	131	38	2	171
4	expected	57	57	57	
5					
6	OIn(O/E)	109.011	-15.408	-6.6998	
7					
8	G	173.8073			
9	W	1.0039			
10					

$G_{calculated}$ is given by G/W, so we use the formula '=b8/b9'. Enter this in a suitable cell, the click on the green tick or press 'return'.



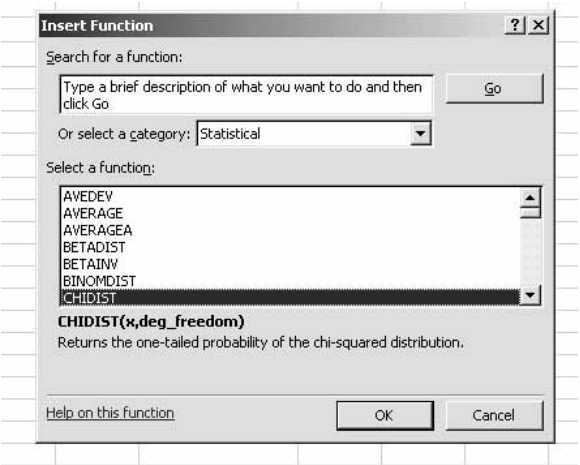
	A	B	C	D	E
1	**Number of holly leaf miners**	**Height on holly tree (m)**			**Total number of holly leaf miners**
2		0.00-1.99	2.00-3.99	4.00-5.99	
3	**observed**	131	38	2	171
4	**expected**	57	57	57	
5					
6	**OIn(O/E)**	109.011	-15.408	-6.6998	
7					
8	**G**	173.8073			
9	**W**	1.0039			
10	**$G_{calculated}$**	173.1323			
11					

We now use the chi-squared distribution to find the probability of accepting the null hypothesis. Using a suitable free cell, go to 'Insert', 'Function'.

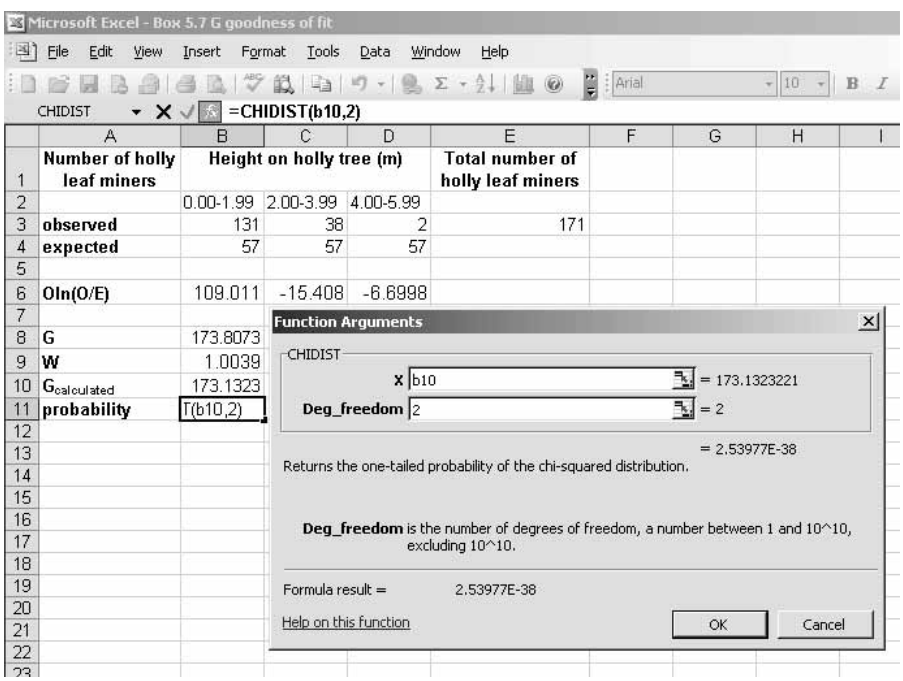


	A	B	C	D	E	F	G	H
1	**Number of holly leaf miners**	**Height on holly tree (m)**			**Total number of holly leaf miners**			
2		0.00-1.99	2.00-3.99	4.00-5.99				
3	**observed**	131	38	2	171			
4	**expected**	57	57	57				
5								
6	**OIn(O/E)**	109.011	-15.408	-6.6998				
7								
8	**G**	173.8073						
9	**W**	1.0039						
10	**$G_{calculated}$**	173.1323						
11	**probability**	=						

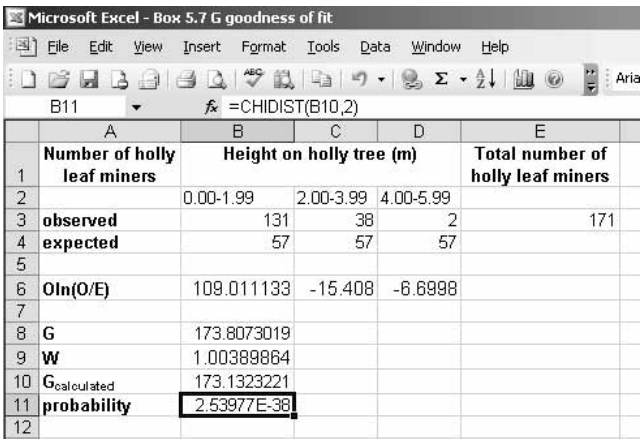
Select the category ‘Statistical’, and click on ‘CHIDIST’.



Click on ‘OK’. Enter the cell where *G* is stored and the number of degrees of freedom (b10 and 2 in this case).



Click on 'OK'.



	A	B	C	D	E
1	**Number of holly leaf miners**	**Height on holly tree (m)**			**Total number of holly leaf miners**
2		0.00-1.99	2.00-3.99	4.00-5.99	
3	**observed**	131	38	2	171
4	**expected**	57	57	57	
5					
6	**OIn(O/E)**	109.011133	-15.408	-6.6998	
7					
8	**G**	173.8073019			
9	**W**	1.00389864			
10	**$G_{calculated}$**	173.1323221			
11	**probability**	2.53977E-38			
12					

The probability of not rejecting the null hypothesis is very small ('2.53977E-38' means '2.53977×10^{-38}', which is a number that starts with 37 zeros after the decimal point), so we reject the null hypothesis, and conclude that there is a significant difference ($G = 173.13$, $p < 0.001$) 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.