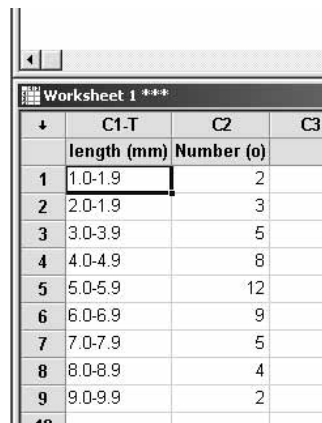


5.1.3. How to check if your data have a normal distribution using a goodness of fit chi-squared test

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. Using the data for the ladybirds, enter the raw numbers into the spreadsheet part of the Minitab window in the form of a summary results table. The two columns will be 'Length (mm)' and 'Number (o)', where the '(o)' means 'observed'.



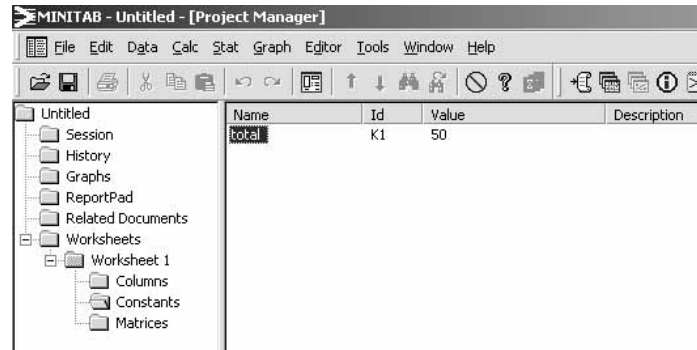
The screenshot shows a Minitab spreadsheet window titled 'Worksheet 1 ***'. The spreadsheet contains a table with the following data:

	C1-T	C2	C3
	length (mm)	Number (o)	
1	1.0-1.9	2	
2	2.0-1.9	3	
3	3.0-3.9	5	
4	4.0-4.9	8	
5	5.0-5.9	12	
6	6.0-6.9	9	
7	7.0-7.9	5	
8	8.0-8.9	4	
9	9.0-9.9	2	

Step 2. Next, we need to calculate the expected values. To do this, we will need to calculate the normal distribution values based on the **mean** and **standard deviation** of our sample.

For the mean, the first thing we need is the total number of observations. Go to 'Calc', 'Calculator', and enter 'K1' in the 'Store result in variable' window. (This will store the total as a constant, K1, that can be used in the future.) Now enter 'sum(c2)' in the 'Expression' window. (Alternatively, scroll down the function list, click on 'sum', and click on 'select'; then choose 'C2 number (o)' from the left-hand window and click on 'select'.) Click on 'OK'. The screen won't have changed.

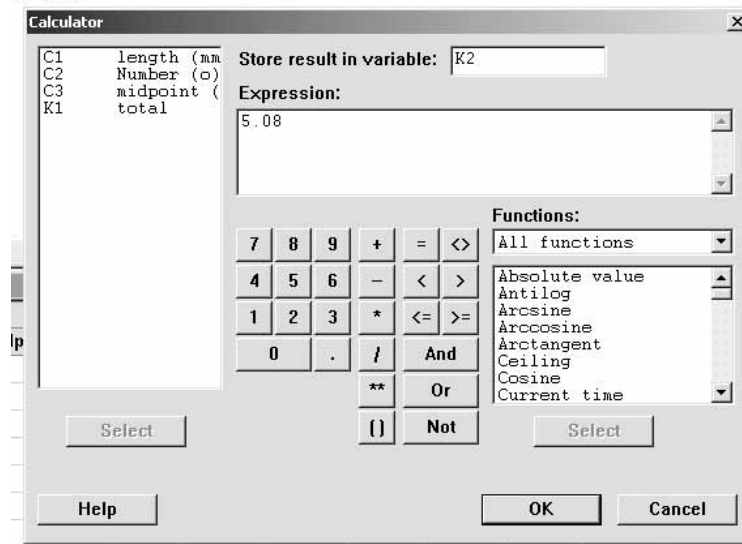
Maximize the project manager (tab at bottom left), select the 'constants' folder, right-click on the name for K1, select 'rename', and call it 'total'.



Minimize the project manager.

The **mean** and **variance** of the ladybird lengths have already been calculated; they are 5.08 mm and 3.74857 mm² respectively. These need to be stored as constants. Go to 'Calc', 'Calculator', enter 'K2' in the 'Store result as variable' window, and type '5.08' into the expression window.

for help.



Click on 'OK'. Go to the project manager, and rename K2 as 'mean'. Repeat for the variance (storing it in K3).

The **standard deviation** is the square root of the **variance**. Go to 'Calc', 'Calculator', type 'sqrt(k3)', and store the result in K4. Rename K4 as 'standard deviation'.

The screenshot shows the Minitab Project Manager interface. On the left, a tree view shows the project structure: Untitled, Session, History, Graphs, ReportPad, Related Documents, Worksheets, and Worksheet 1 (containing Columns, Constants, and Matrices). The main window displays a table of statistical results:

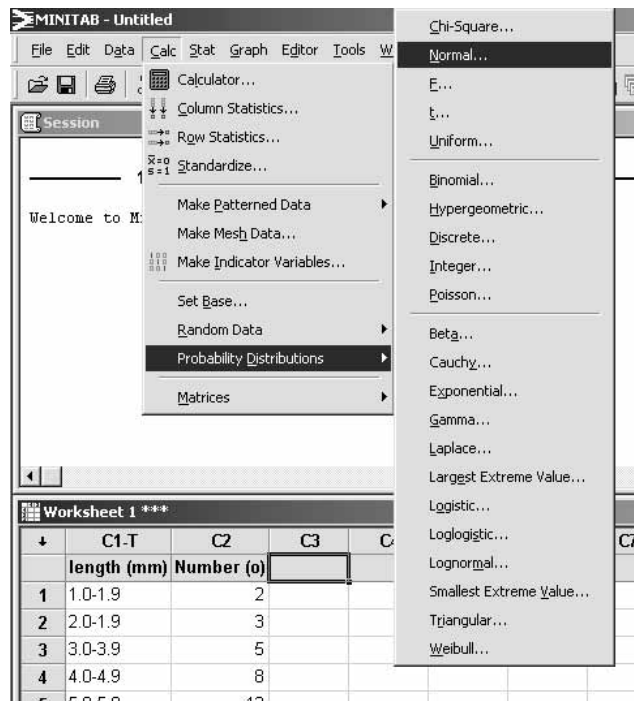
Name	Id	Value
total	K1	50
mean	K2	5.08
variance	K3	3.74857
standard deviation	K4	1.93612

Find the mid-points of the ranges. These are given by half the sum of the top and bottom limits, so they can be calculated manually and entered by hand.

The screenshot shows a Minitab worksheet titled 'Worksheet 1 ***'. The data is organized into columns: C1-T (length (mm)), C2 (Number (o)), C3 (midpoint (mm)), and C4. The data is as follows:

	C1-T	C2	C3	C4
	length (mm)	Number (o)	midpoint (mm)	
1	1.0-1.9	2	1.45	
2	2.0-1.9	3	2.45	
3	3.0-3.9	5	3.45	
4	4.0-4.9	8	4.45	
5	5.0-5.9	12	5.45	
6	6.0-6.9	9	6.45	
7	7.0-7.9	5	7.45	
8	8.0-8.9	4	8.45	
9	9.0-9.9	2	9.45	
10				
11				

The expected values can be found as follows: go to 'Calc', 'Probability Distributions', 'Normal',

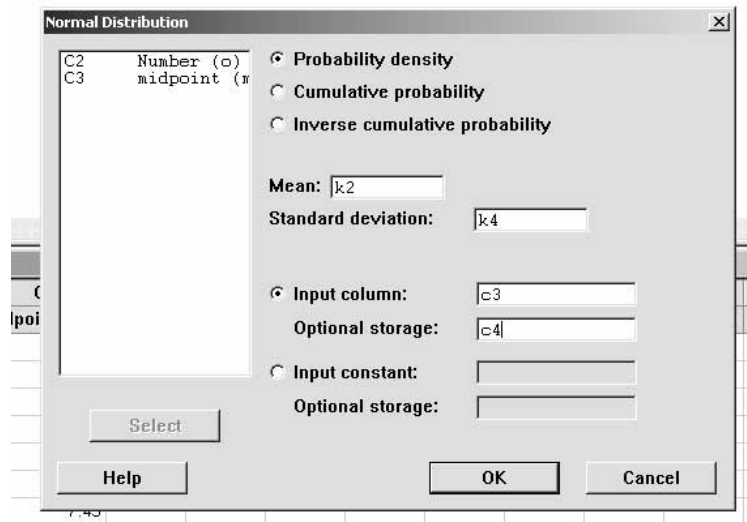


Select 'Probability density'.

Type in 'k2' for the mean and 'k4' for the standard deviation.

Type in c3 for the input column, and c4 for the output column.

(Alternatively, these can be entered by clicking in the windows, selecting from the left-hand window and clicking on 'Select'.)

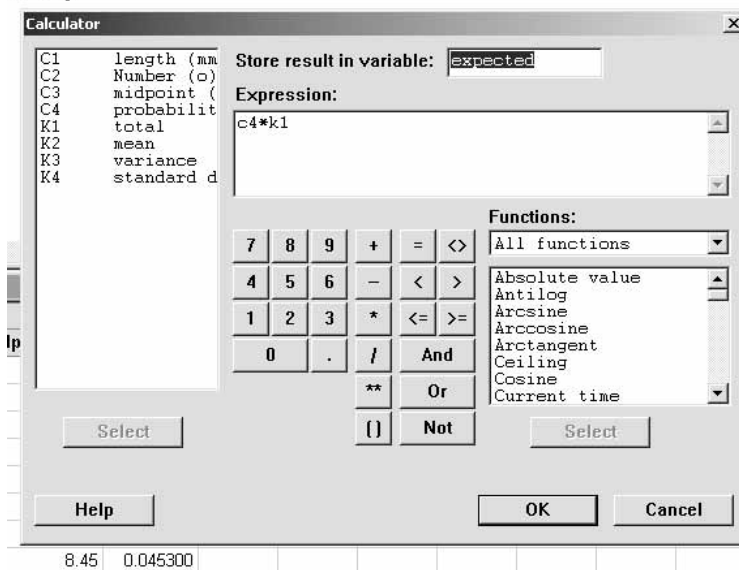


Click on 'OK', and add the name 'probability' to column 4.

	C1-T	C2	C3	C4	C5
	length (mm)	Number (o)	midpoint (mm)	probability	
1	1.0-1.9	2	1.45	0.035536	
2	2.0-1.9	3	2.45	0.081902	
3	3.0-3.9	5	3.45	0.144567	
4	4.0-4.9	8	4.45	0.195427	
5	5.0-5.9	12	5.45	0.202324	
6	6.0-6.9	9	6.45	0.160418	
7	7.0-7.9	5	7.45	0.097410	
8	8.0-8.9	4	8.45	0.045300	
9	9.0-9.9	2	9.45	0.016134	
10					

To find the actual expected values, we need to multiply the probabilities by the total number of observations, N . Go to 'Calc', 'Calculator', enter 'expected' in the 'Store results in variable' window, and type 'c4*k1' in the 'Expression' window.

or neip.

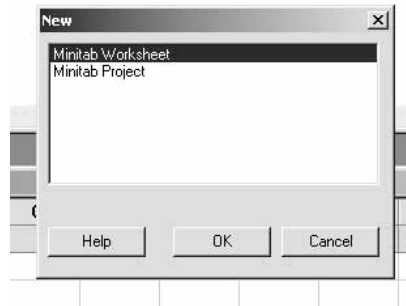


Click on 'OK'.

	C1-T	C2	C3	C4	C5	C6
	length (mm)	Number (o)	midpoint (mm)	probability	expected	
1	1.0-1.9	2	1.45	0.035536	1.7768	
2	2.0-1.9	3	2.45	0.081902	4.0951	
3	3.0-3.9	5	3.45	0.144567	7.2283	
4	4.0-4.9	8	4.45	0.195427	9.7714	
5	5.0-5.9	12	5.45	0.202324	10.1162	
6	6.0-6.9	9	6.45	0.160418	8.0209	
7	7.0-7.9	5	7.45	0.097410	4.8705	
8	8.0-8.9	4	8.45	0.045300	2.2650	
9	9.0-9.9	2	9.45	0.016134	0.806680	

Step 3. We can now do the test to compare the actual values (in column 2) with the **expected** ones from a normal distribution (in column 5).

First, we note that several of the expected values are less than 5, which violates one of the requirements of a chi-squared test. To mitigate this problem, we combine the lowest and highest two length ranges. This is probably easiest to do by starting a new worksheet. Go to 'File', 'New', and select 'Minitab Worksheet'.



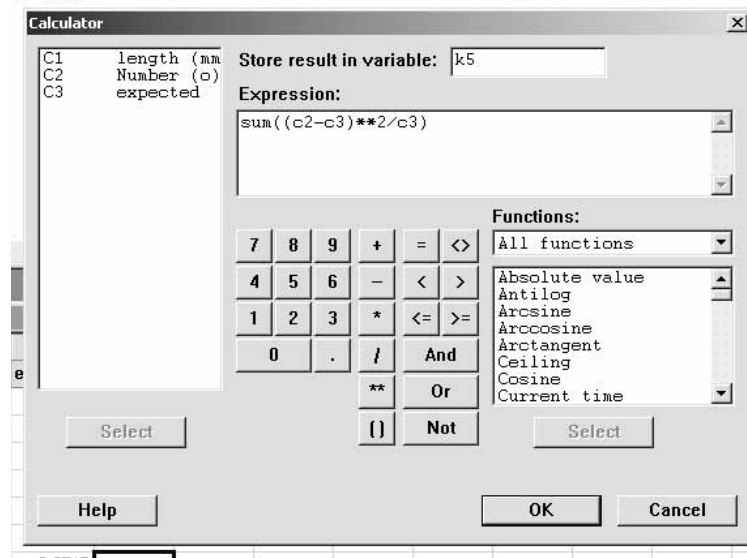
Click on 'OK', and a new worksheet will appear.

Transfer over the data, making sure that the numbers are added together as appropriate.

	C1-T	C2	C3	C4
	length (mm)	Number (o)	expected	
1	1.0-2.9	5	5.8719	
2	3.0-3.9	5	7.2283	
3	4.0-4.9	8	9.7714	
4	5.0-5.9	12	10.1162	
5	6.0-6.9	9	8.0209	
6	7.0-7.9	5	4.8705	
7	8.0-9.9	6	3.0717	
8				

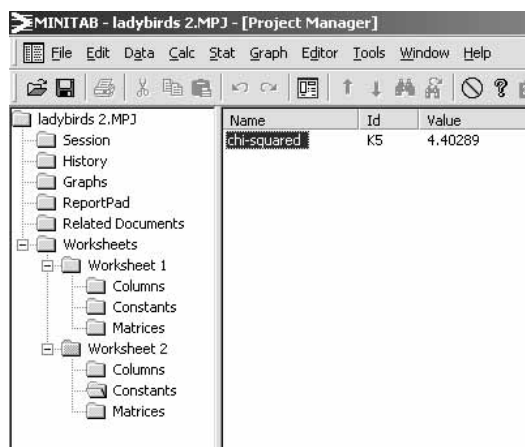
Next, calculate chi-squared. This is done by going to 'Calc', 'Calculator', typing 'k5' in the 'Store results in variable' window, and 'sum((c2 - c5)**2/c5)' in the 'Expression' window.

for help.



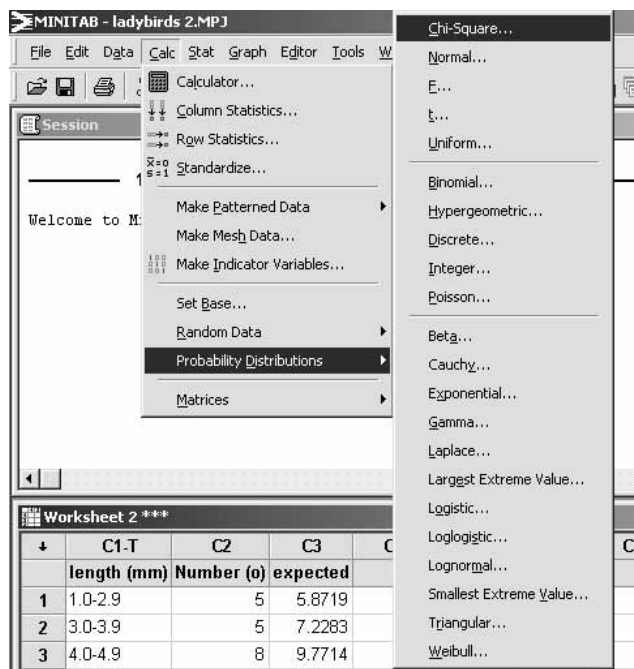
Then hit 'OK'.

Go to the project manager, open the constants folder for worksheet 2, and rename k5 as 'chi-squared'.

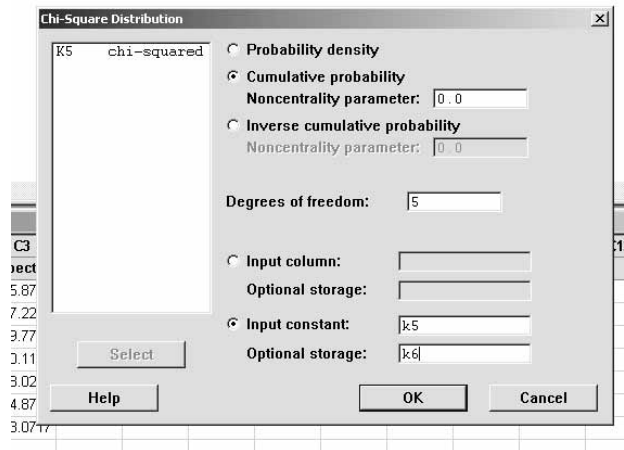


Minimize the project manager.

Go to 'Calc', 'Probability Distributions', 'Chi-squared'.

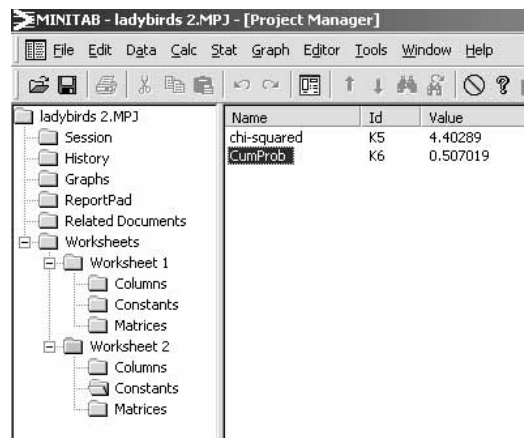


Select 'Cumulative Probability', and enter '5' (for this example) in the 'degrees of freedom' window. Select 'Input Constant' and enter 'k5', and for 'Optional Storage' enter 'k6'.



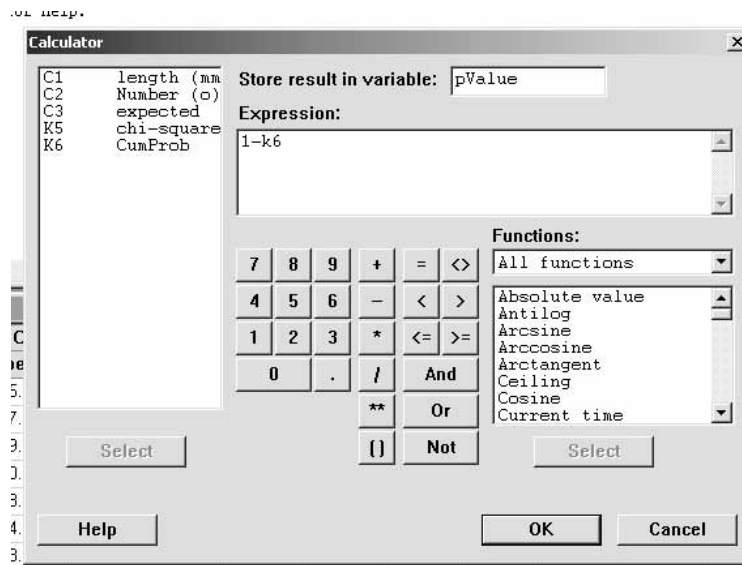
Then click on 'OK'.

Got to the project manager, and rename k6 as 'CumProb'.



Minimize the project manager.

The final step is to find the probability that the null hypothesis (the observed value) can be rejected. Go to 'Calc', 'Calculator'. Enter 'pValue' in the 'Store results in variable' window, and type '1 - k6' in the 'Expression' window.



Click on 'OK'.

	C1-T	C2	C3	C4
	length (mm)	Number (o)	expected	pValue
1	1.0-2.9	5	5.8719	0.492981
2	3.0-3.9	5	7.2283	
3	4.0-4.9	8	9.7714	
4	5.0-5.9	12	10.1162	
5	6.0-6.9	9	8.0209	
6	7.0-7.9	5	4.8705	
7	8.0-9.9	6	3.0717	
8				

The p value is the probability that we do not reject the null hypothesis. In this example $p = 0.49$, which is greater than the threshold of $p = 0.05$, so we do not reject the null hypothesis. There is no significant difference ($\chi^2_{\text{calculated}} = 4.40$, $p = 0.49$) 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.