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

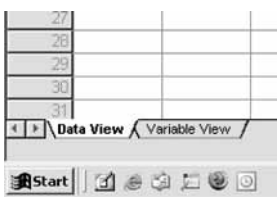
EXAMPLE 3.7. Length (mm) of two-spot ladybirds (*Adalia bipunctata*)

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

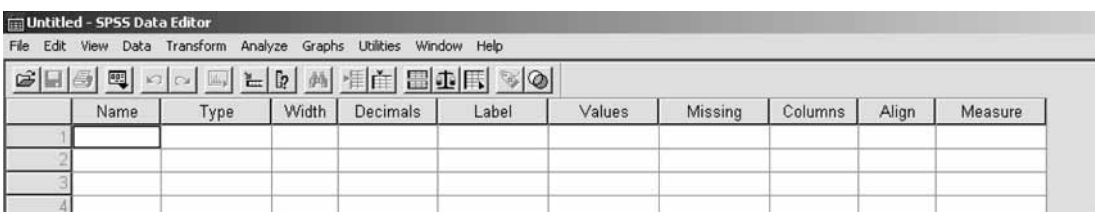
1. Set up the variables.

(i) When SPSS starts, select the 'Type in data' option.

(ii) Then choose 'Variable View' from the tabs at the bottom left.

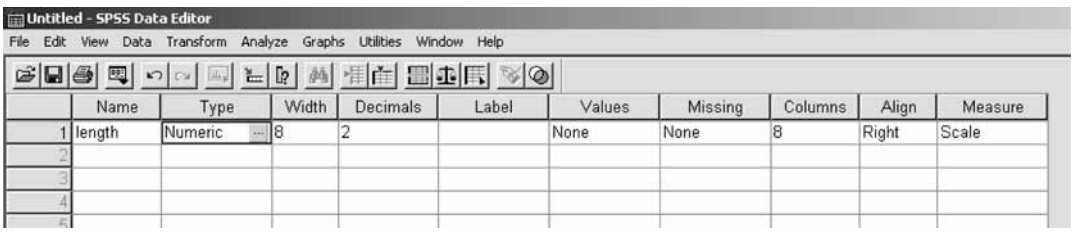


You will see a screen something like this:

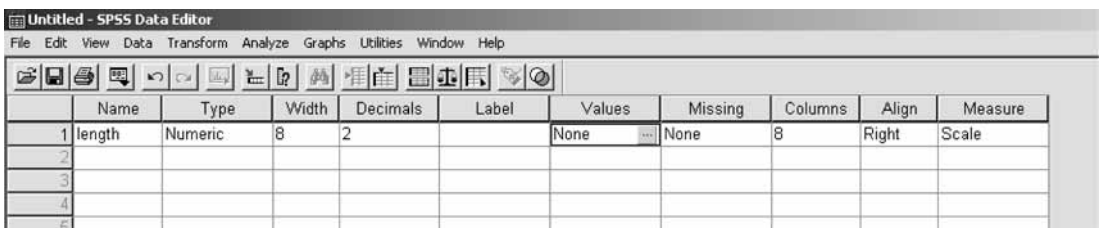


Each row represents a variable for the analysis.

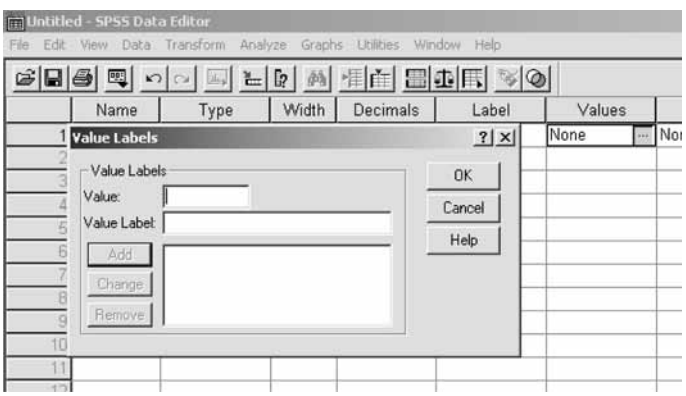
(iii) In the name for variable 1, type 'length' (SPSS won't accept capital letters as parts of a Name, and Names can be no more than eight characters long). Most of the other characteristics of the variable will be give default values as below:



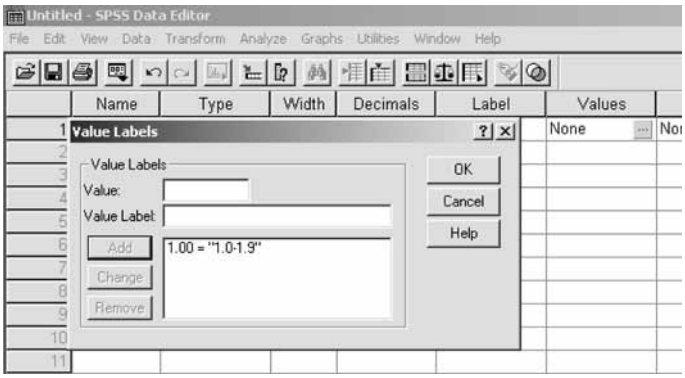
(iv) To input the values for the ranges, click in the 'Values' cell for 'length'. This will produce a grey area at the right of the cell.



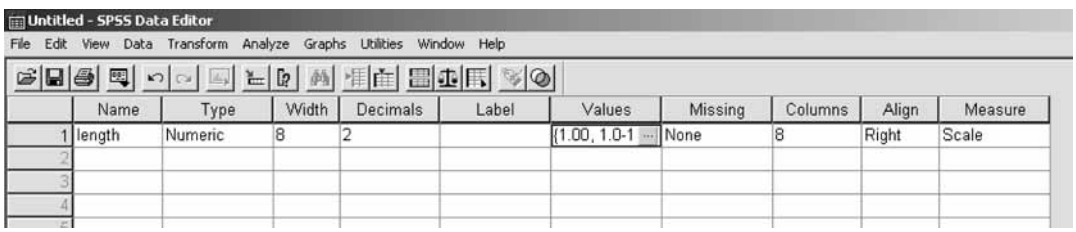
and clicking on this will give a dialogue box for inputting the values.



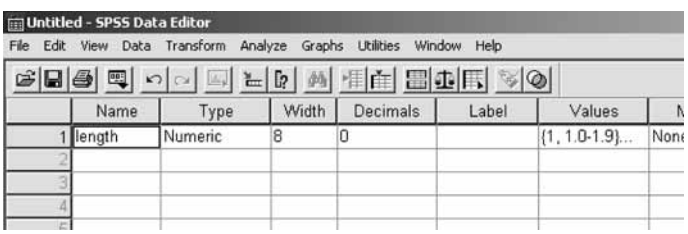
Type in a value of 1 and a value label of '1.0–1.9'. Then click on 'Add'. This will add your new value–label pair to the window at the bottom of the dialogue box.



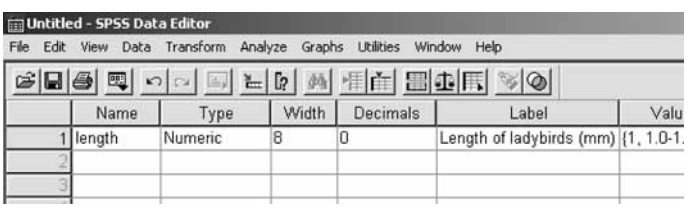
Repeat for the other eight value-label pairs, then click on 'OK'. You should finish up with something like this:



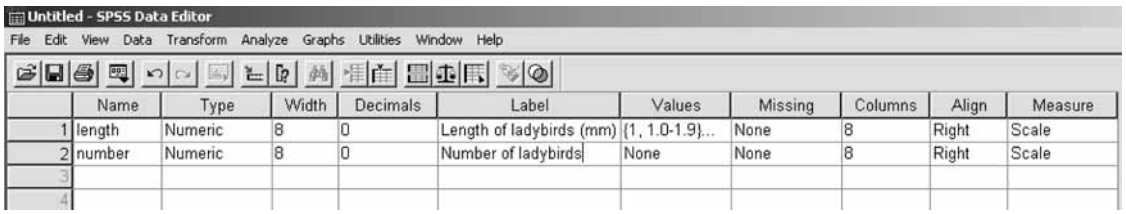
(If you click on the 'decimals' cell, and change the number of decimal places to zero using the arrows, the '1.00' at the start of the Values becomes simply '1'.)



(v) In the 'label' column, enter a description of the ranges, for example 'Length of ladybird (mm)'. Note that the cell width expands to fit the text.



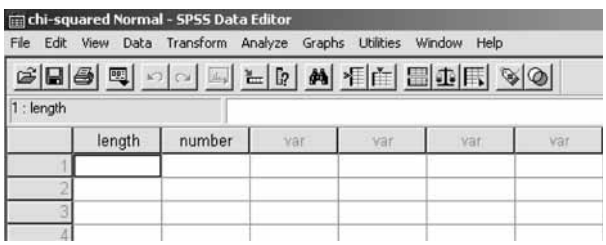
(vi) Next set up the variable to contain the actual counts of ladybirds. Give it the name 'number', and set decimals to zero (we can't have fractions of a ladybird). Put in the label 'Number of ladybirds'.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	length	Numeric	8	0	Length of ladybirds (mm)	{1, 1.0-1.9}...	None	8	Right	Scale
2	number	Numeric	8	0	Number of ladybirds	None	None	8	Right	Scale
3										
4										

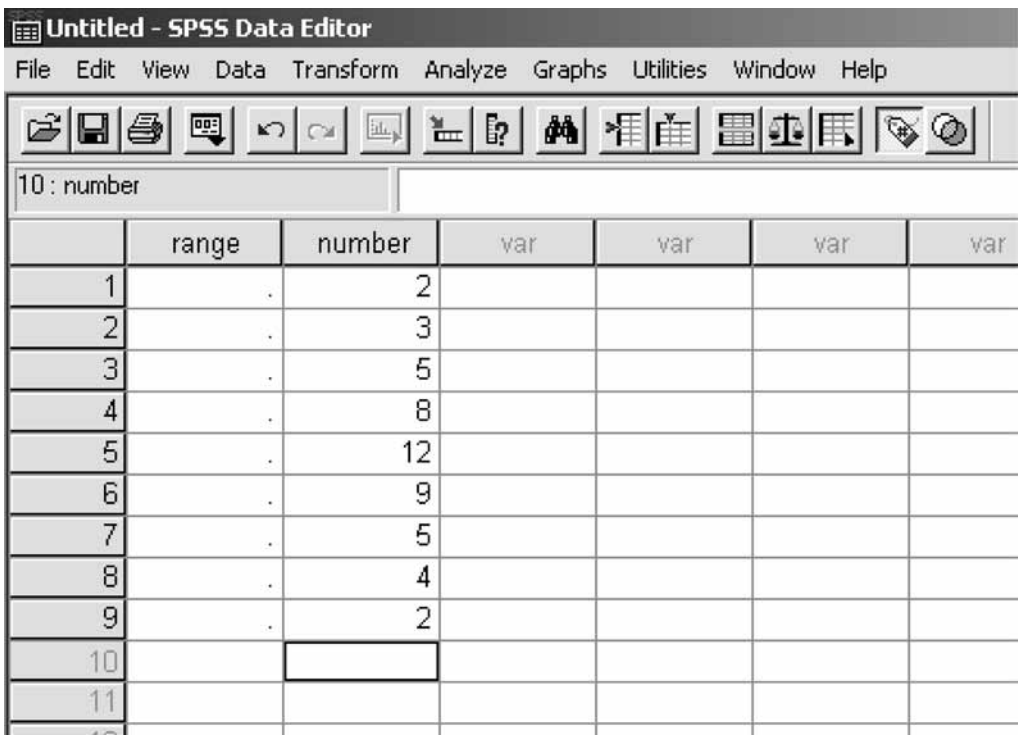
2. Enter the Data

(i) Transfer to 'Data View' using the tab at the bottom left of the screen. You should get something like this:

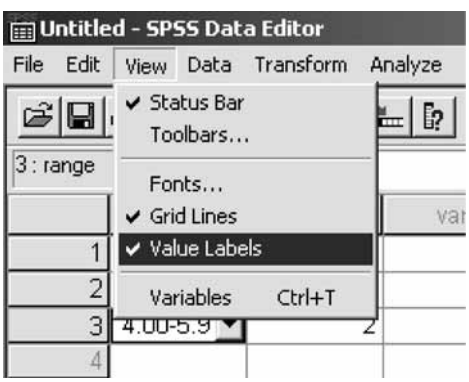


	length	number	var	var	var	var
1						
2						
3						
4						

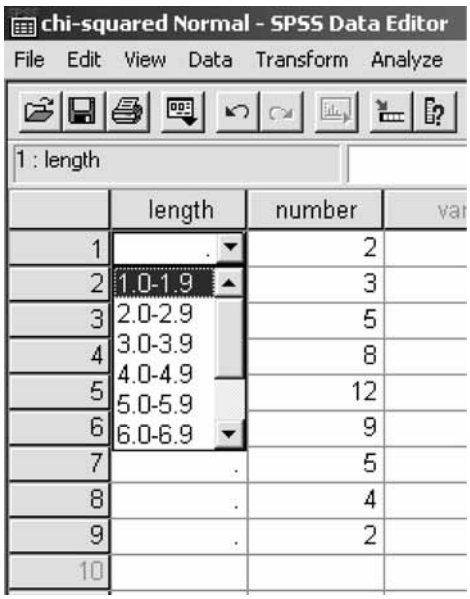
(ii) Another quirk of SPSS is that you have to input the numbers before it lets you put in the labels. Put the numbers of ladybirds into the second column:



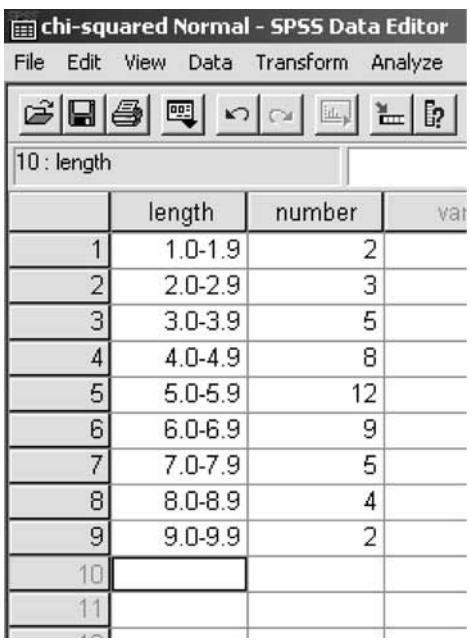
(iii) Check that value labels are enabled by going to 'View' and ensuring that 'Value Labels' is selected.



Now click in the first 'range' cell. You will get a drop-down menu of the values you put in while in variable view.

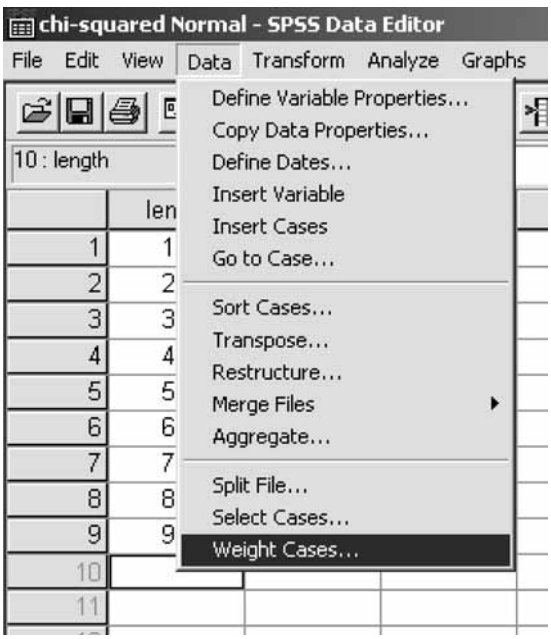


Select the first one (1.0–1.9). Repeat for the other ranges.

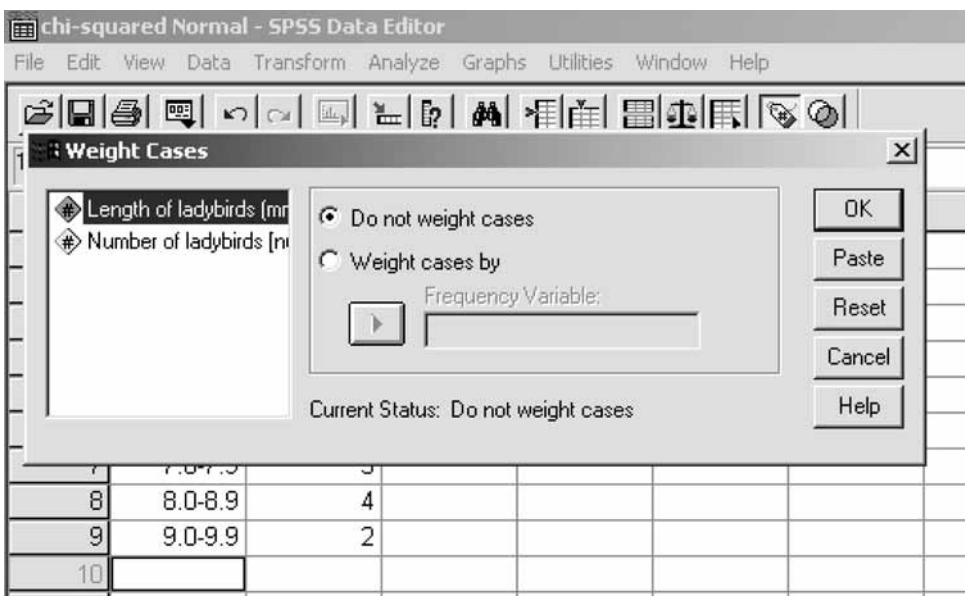


3. Perform the test.

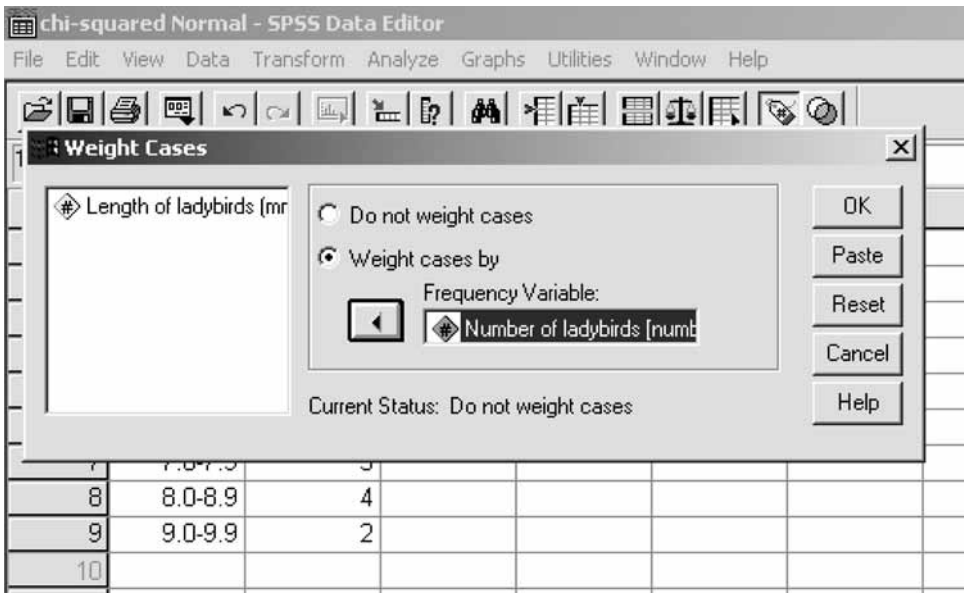
(i) The ranges are our test variables, and the weightings for each variable are the number of ladybirds. The first thing to do is assign the weightings. Go to 'Data' and select 'Weight Cases'.



This brings up a dialogue box.



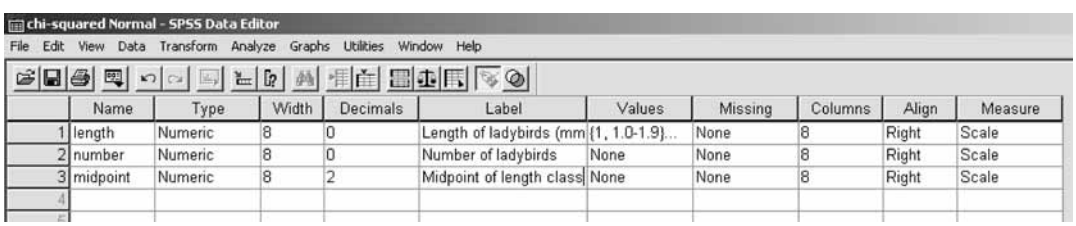
Make sure that the 'Weight cases by' radio button is clicked. In the left-hand window, there will be a list of variables. Select 'Number of ladybirds from the list, and click on the arrow to transfer it to the box labelled 'Frequency Variable'.



Click on 'OK'.

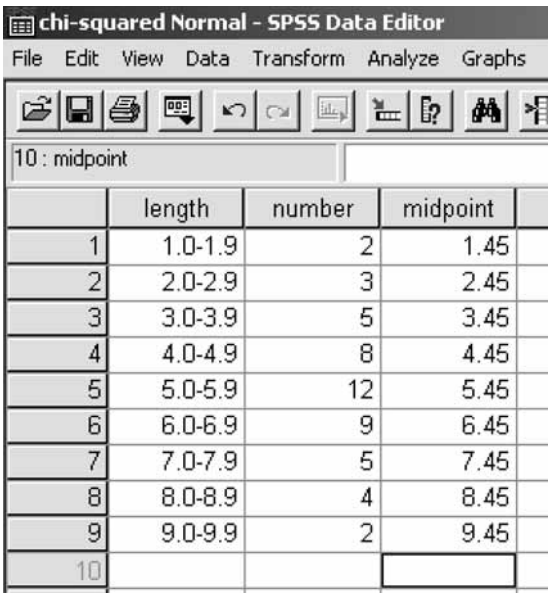
(ii) To calculate our expected values, we need some descriptive statistics of our data, in particular the mean (5.08 mm) and standard deviation (1.936 mm)—these were calculated from the original data in BOXES 3.2 and 3.1 respectively.

(iii) Next, we use these to calculate our expected frequencies. These will be based on the midpoints of our classes, so we need to calculate these and make them available to the software. Go back to 'Variable view' using the tabs at bottom left, and set up another variable called 'midpoint'. Give it the label 'Midpoint of length class'.



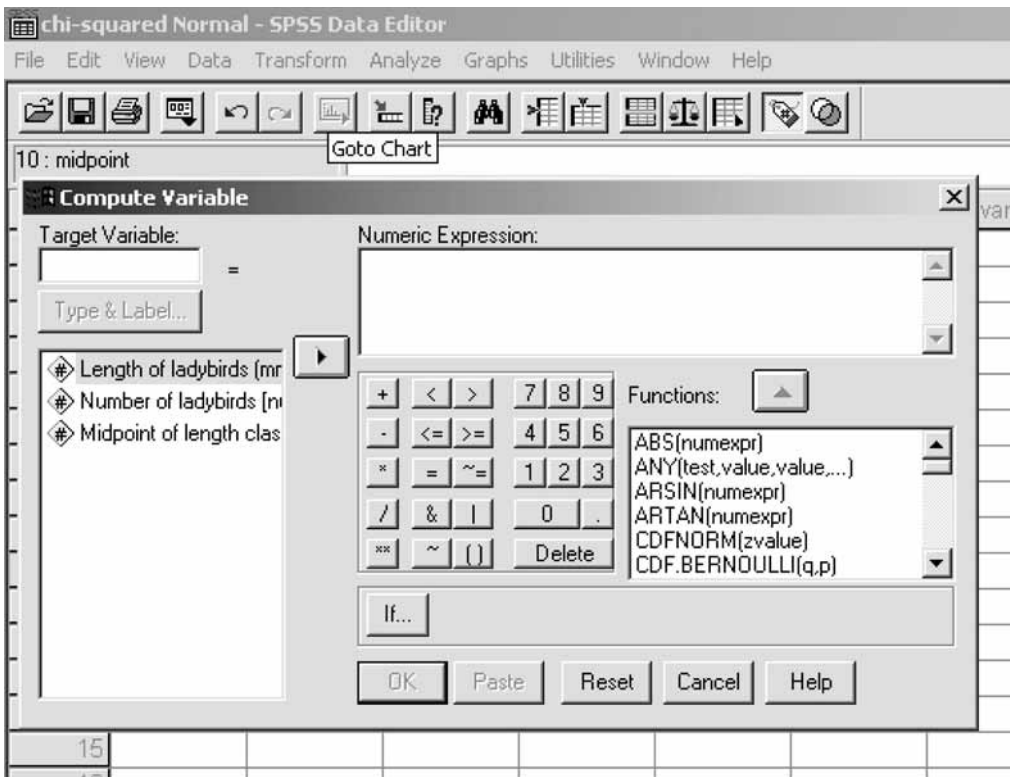
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	length	Numeric	8	0	Length of ladybirds (mm	{1, 1.0-1.9}...	None	8	Right	Scale
2	number	Numeric	8	0	Number of ladybirds	None	None	8	Right	Scale
3	midpoint	Numeric	8	2	Midpoint of length class	None	None	8	Right	Scale
4										

Return to 'Data View' using the tabs at bottom-left, and type in the midpoints of the ranges.

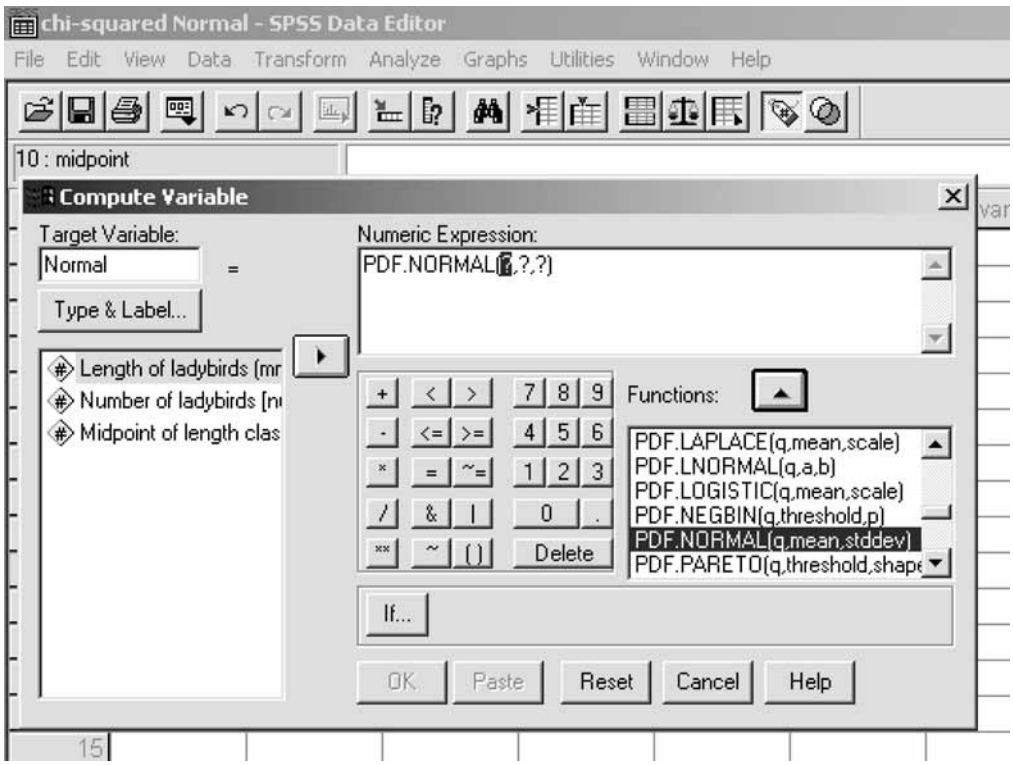


	length	number	midpoint
1	1.0-1.9	2	1.45
2	2.0-2.9	3	2.45
3	3.0-3.9	5	3.45
4	4.0-4.9	8	4.45
5	5.0-5.9	12	5.45
6	6.0-6.9	9	6.45
7	7.0-7.9	5	7.45
8	8.0-8.9	4	8.45
9	9.0-9.9	2	9.45
10			

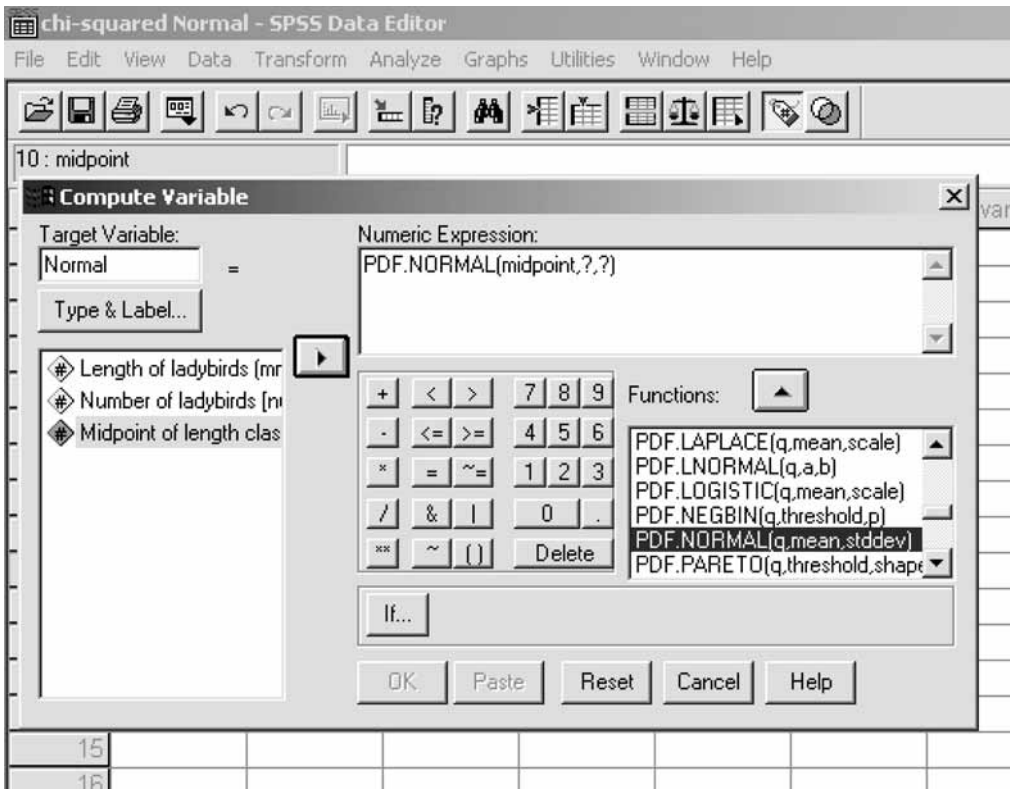
Go to ‘Transform’ and ‘Compute’. You will get a window like this:



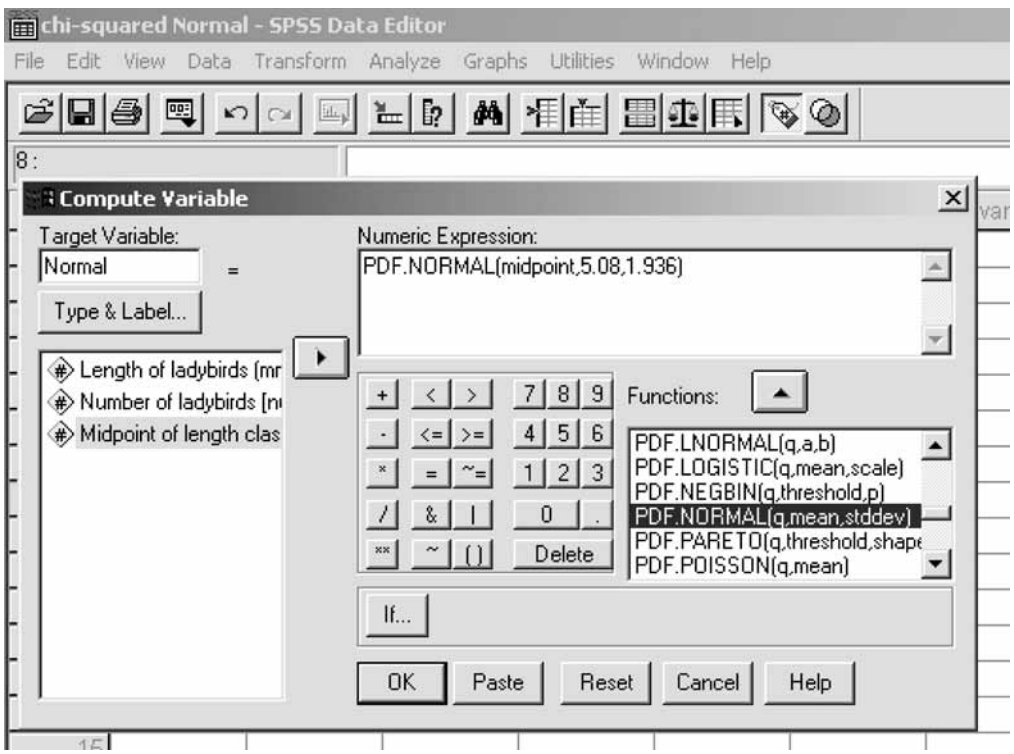
In the 'Target Variable' window, type the name of the quantity to be computed: in this case, 'Normal' would be a good name. Then scroll down the list of available functions, and select PDF.NORMAL(q,mean,stddev). PDF stands for 'probability density function', and q is the value at which the function is to be calculated. Click on the up arrow to move this into the top right window.



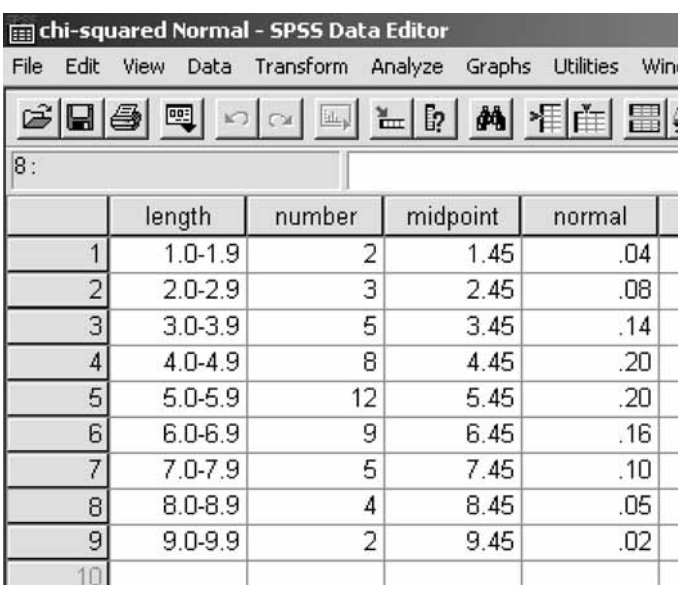
When the first question mark is highlighted, select 'Midpoint of length class' from the bottom left-hand window, and click on the right-arrow to transfer it into the expression we are building.



The second question mark represents the mean (5.08), and the third represents the **standard deviation** (1.936), so we will have to type them in by hand.

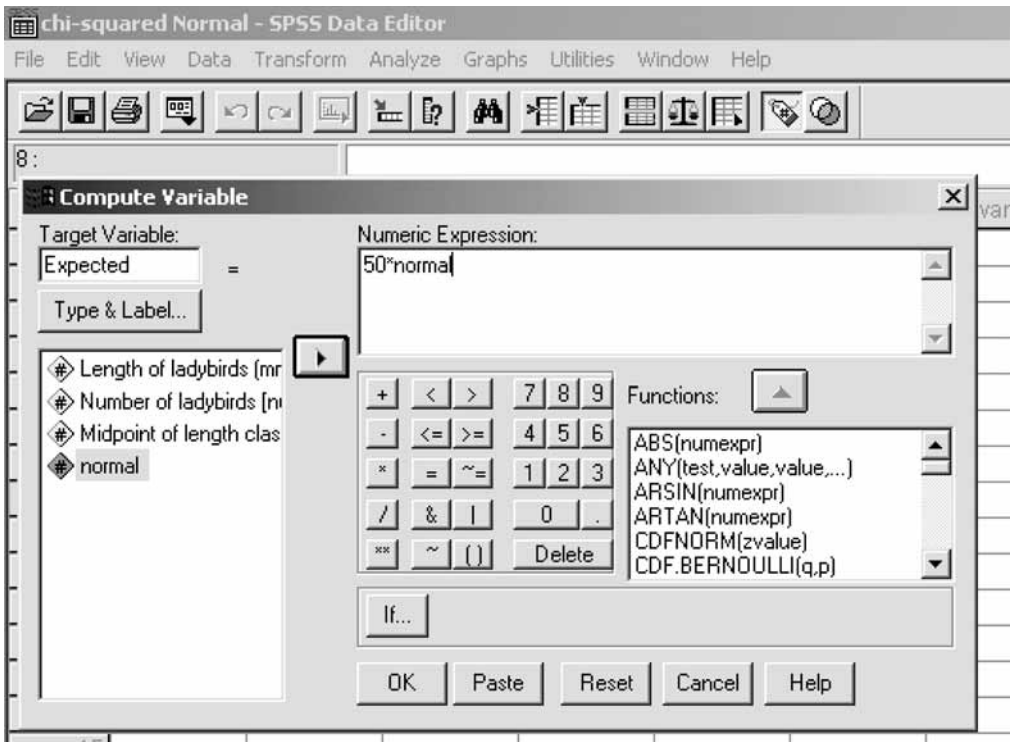


Finally, click on 'OK'.

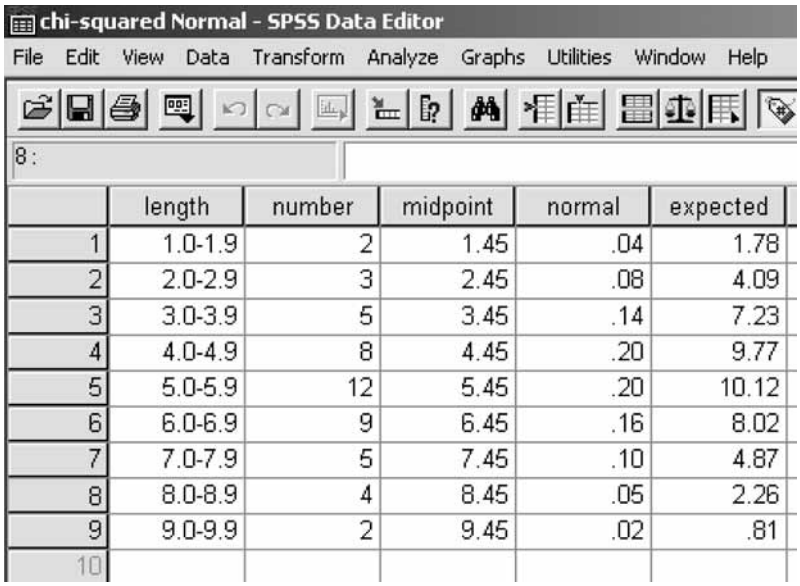


	length	number	midpoint	normal
1	1.0-1.9	2	1.45	.04
2	2.0-2.9	3	2.45	.08
3	3.0-3.9	5	3.45	.14
4	4.0-4.9	8	4.45	.20
5	5.0-5.9	12	5.45	.20
6	6.0-6.9	9	6.45	.16
7	7.0-7.9	5	7.45	.10
8	8.0-8.9	4	8.45	.05
9	9.0-9.9	2	9.45	.02
10				

We have generated a list of the expected frequencies if the distribution were Normal. To calculate the expected values, we simply multiply these by the total number of ladybirds (50). Go to 'Transform', 'Compute'. In the 'Target variable' window, type 'expected'. Into the expression window, type '50*', and then highlight 'Normal' in the left-hand window by clicking on it, and click on the arrow to transfer it into the expression.



Click on 'OK'.

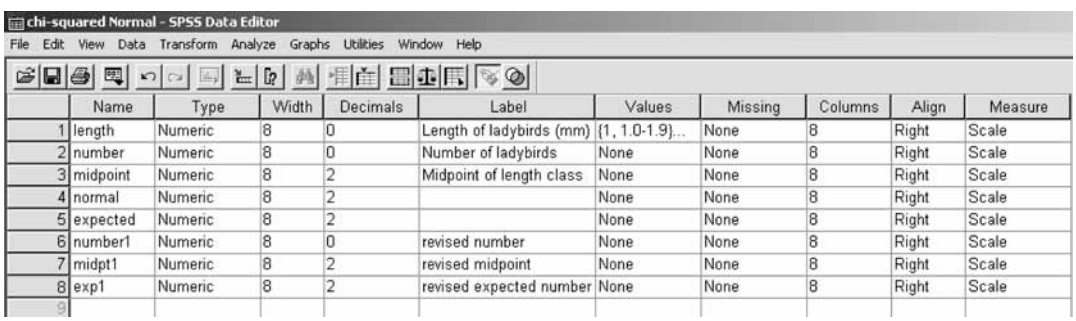

chi-squared Normal - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

8 :

	length	number	midpoint	normal	expected
1	1.0-1.9	2	1.45	.04	1.78
2	2.0-2.9	3	2.45	.08	4.09
3	3.0-3.9	5	3.45	.14	7.23
4	4.0-4.9	8	4.45	.20	9.77
5	5.0-5.9	12	5.45	.20	10.12
6	6.0-6.9	9	6.45	.16	8.02
7	7.0-7.9	5	7.45	.10	4.87
8	8.0-8.9	4	8.45	.05	2.26
9	9.0-9.9	2	9.45	.02	.81
10					

Note that five of these values are less than 5, which is not good for a chi-squared test. To overcome this, we need to combine some of our classes. We will need three new variables, number1, midpt1 and exp1. The first will be an integer (set the 'decimals' property to zero), but the other two will need two decimal places. Set them up in variable view, and then return to data view.


chi-squared Normal - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	length	Numeric	8	0	Length of ladybirds (mm)	{1, 1.0-1.9}...	None	8	Right	Scale
2	number	Numeric	8	0	Number of ladybirds	None	None	8	Right	Scale
3	midpoint	Numeric	8	2	Midpoint of length class	None	None	8	Right	Scale
4	normal	Numeric	8	2		None	None	8	Right	Scale
5	expected	Numeric	8	2		None	None	8	Right	Scale
6	number1	Numeric	8	0	revised number	None	None	8	Right	Scale
7	midpt1	Numeric	8	2	revised midpoint	None	None	8	Right	Scale
8	exp1	Numeric	8	2	revised expected number	None	None	8	Right	Scale
9										

We combine the lowest two classes, and the highest two classes. The numbers and expected numbers can be found by adding the numbers in the two combined classes; and the midpoint of the combined class can be found by averaging the midpoints of the two combined classes. Other data values can be transferred to the new columns by copying and pasting.

chi-squared Normal - SPSS Data Editor

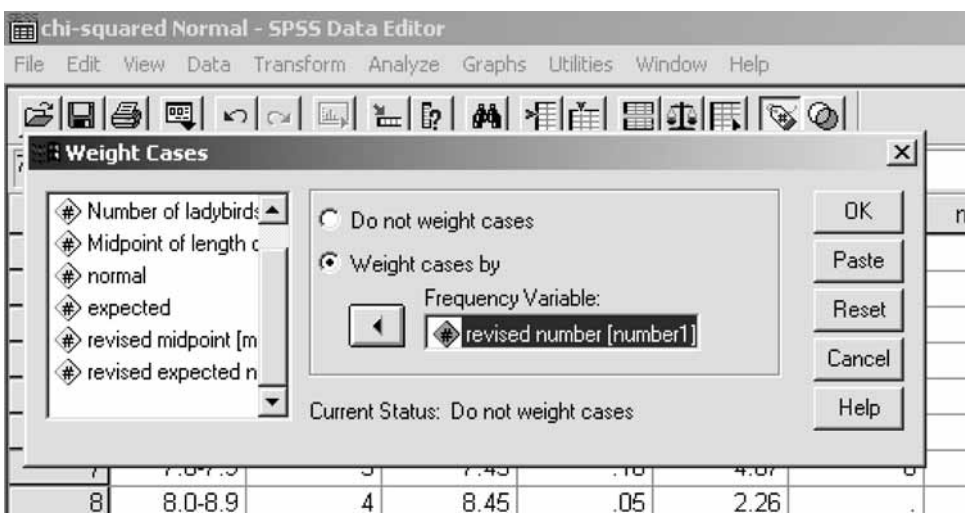
File Edit View Data Transform Analyze Graphs Utilities Window Help

7 : exp1 3.07

	length	number	midpoint	normal	expected	number1	midpt1	exp1
1	1.0-1.9	2	1.45	.04	1.78	5	1.95	5.87
2	2.0-2.9	3	2.45	.08	4.09	5	3.45	7.23
3	3.0-3.9	5	3.45	.14	7.23	8	4.45	9.77
4	4.0-4.9	8	4.45	.20	9.77	12	5.45	10.12
5	5.0-5.9	12	5.45	.20	10.12	9	6.45	8.02
6	6.0-6.9	9	6.45	.16	8.02	5	7.45	4.87
7	7.0-7.9	5	7.45	.10	4.87	6	8.95	3.07
8	8.0-8.9	4	8.45	.05	2.26	.	.	.
9	9.0-9.9	2	9.45	.02	.81	.	.	.
10								

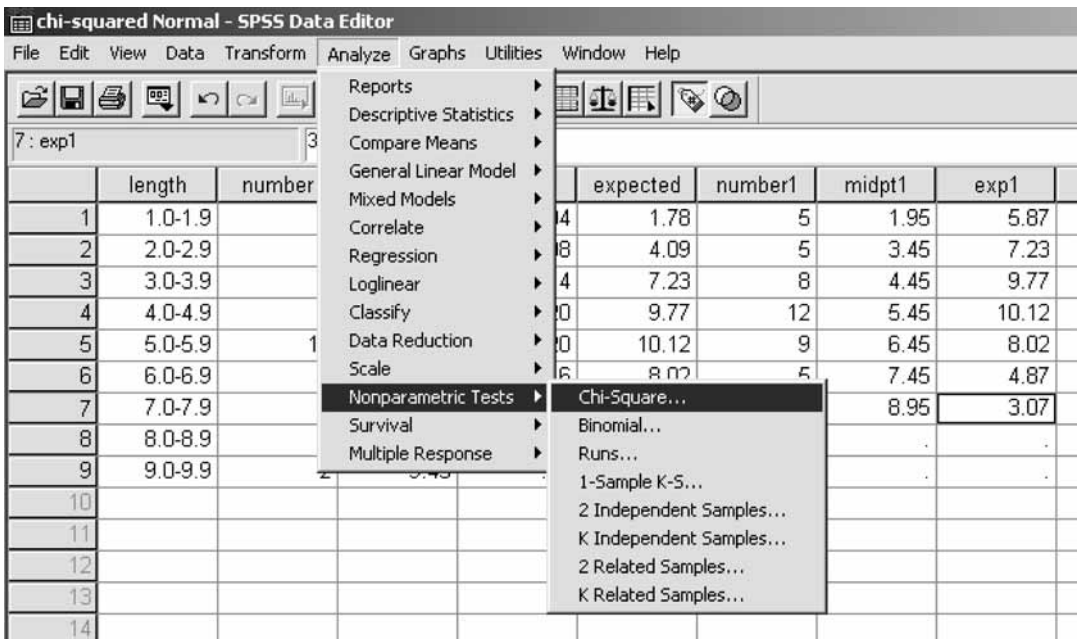
Two of the expected values are still less than 5, but this is a substantial improvement.

(iv) Now we go back to the test. The first step is to apply the appropriate weighting to our variables. Go to 'Data', 'Weight cases', and transfer 'Revised number' to the 'Frequency Variable' window. (You may have to evict 'number' first: highlight it, and click on the arrow.)

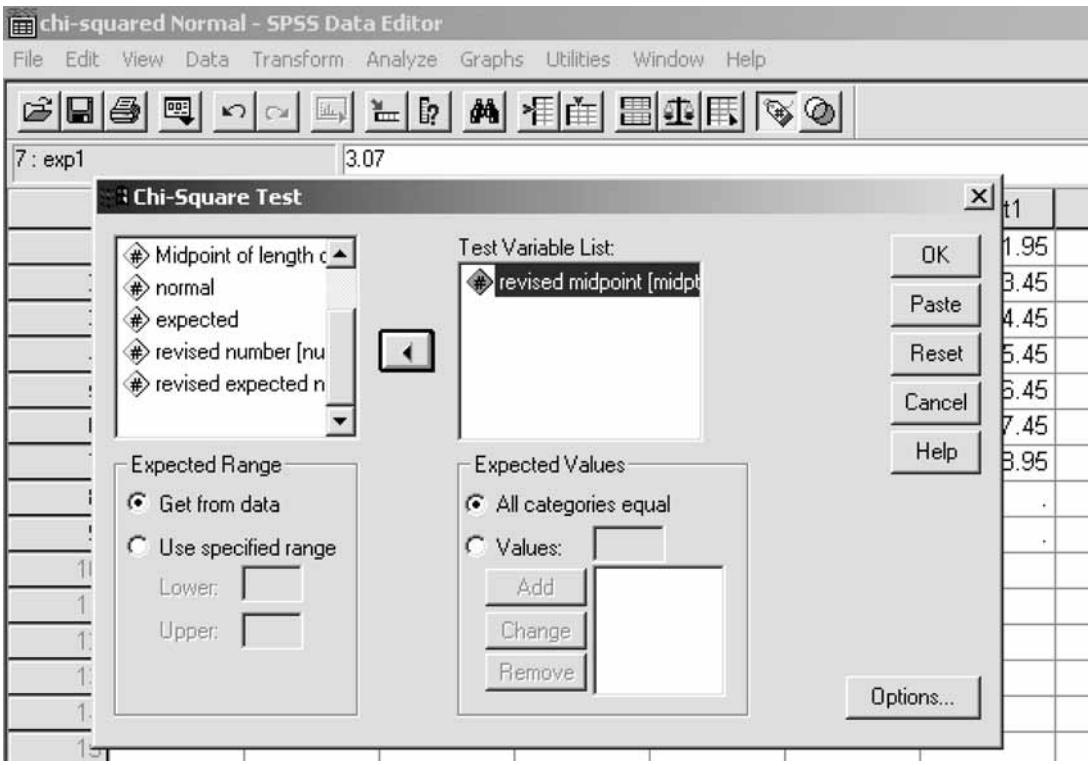


Click on 'OK'.

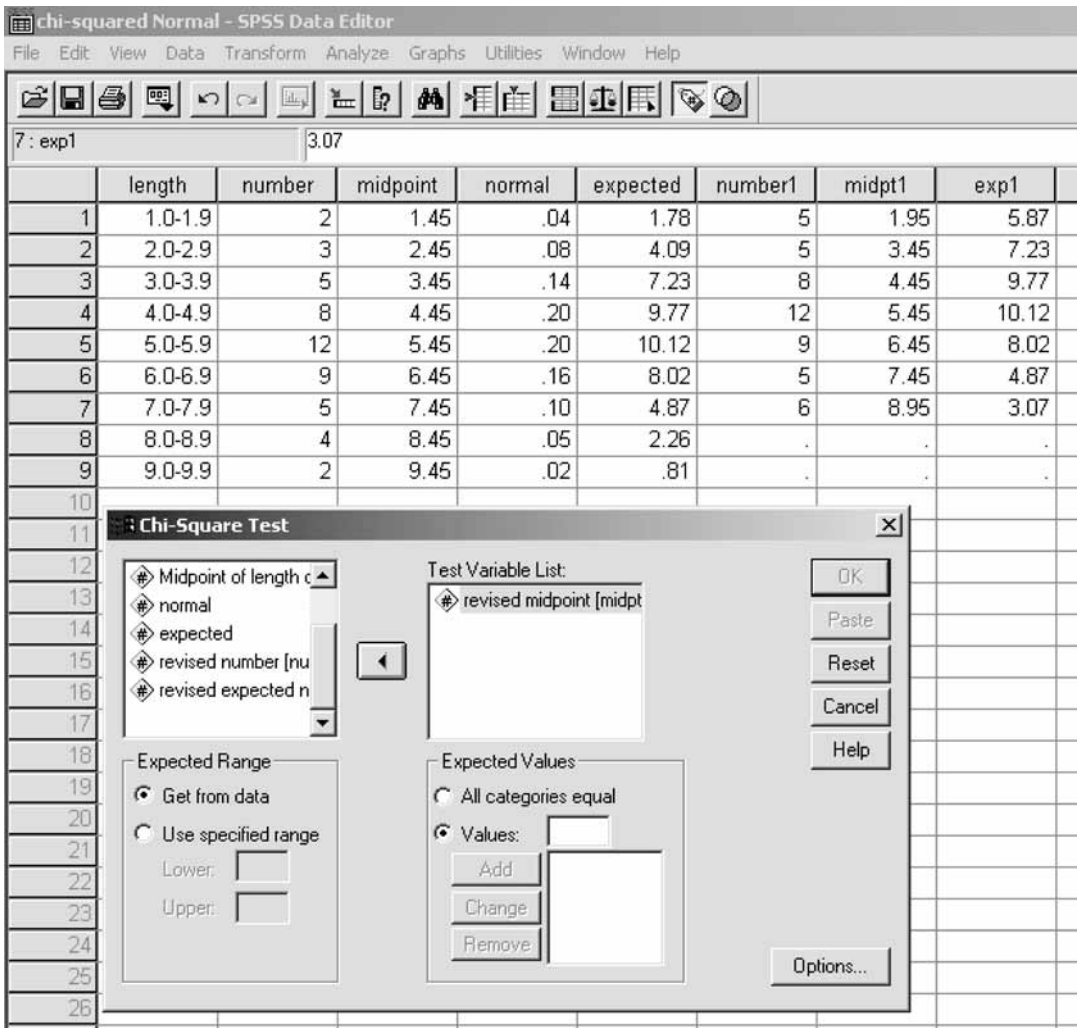
Go to 'Analyze' (sic), '**Nonparametric** Tests' and 'Chi-Square'.



Select 'Revised Midpoint' and click on the right-pointing arrow to transfer it to the 'Test Variable List' window.

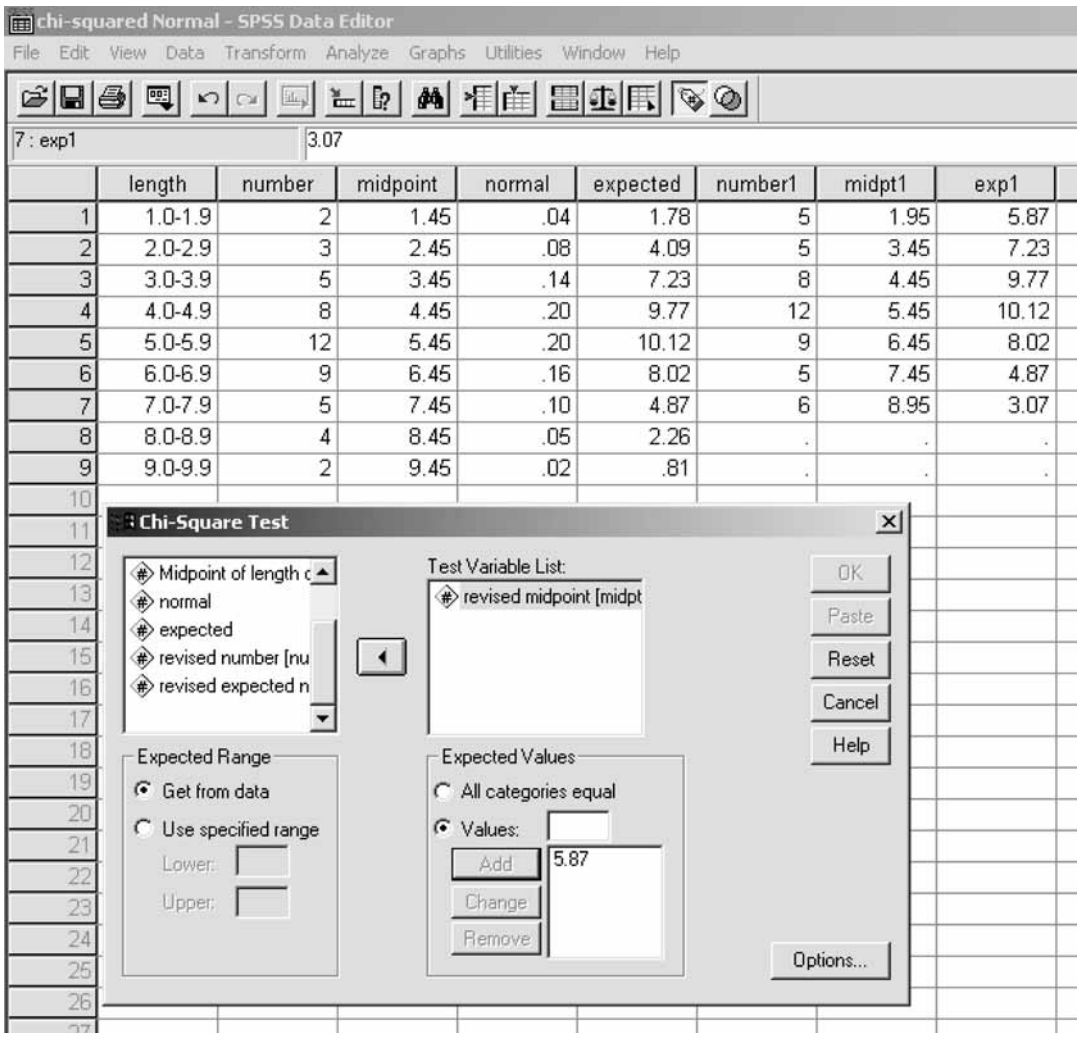


Next, we need to enter our expected values. Select the 'Values' radio button in the 'Expected Values' box. At this stage, it would help to pull the 'Chi-Square Test' box down (click and drag on its title bar) so that the data behind can be seen.

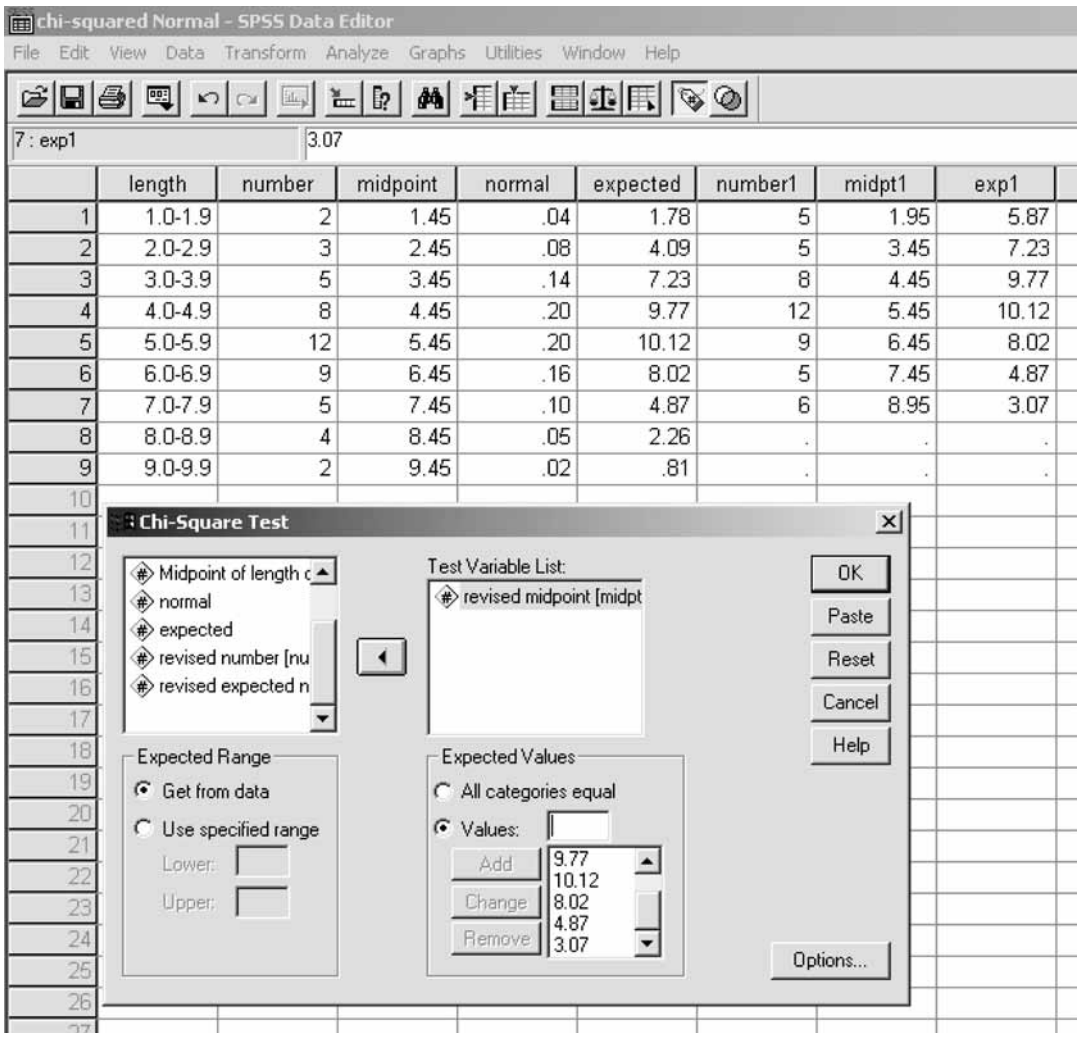


	length	number	midpoint	normal	expected	number1	midpt1	exp1
1	1.0-1.9	2	1.45	.04	1.78	5	1.95	5.87
2	2.0-2.9	3	2.45	.08	4.09	5	3.45	7.23
3	3.0-3.9	5	3.45	.14	7.23	8	4.45	9.77
4	4.0-4.9	8	4.45	.20	9.77	12	5.45	10.12
5	5.0-5.9	12	5.45	.20	10.12	9	6.45	8.02
6	6.0-6.9	9	6.45	.16	8.02	5	7.45	4.87
7	7.0-7.9	5	7.45	.10	4.87	6	8.95	3.07
8	8.0-8.9	4	8.45	.05	2.26	.	.	.
9	9.0-9.9	2	9.45	.02	.81	.	.	.
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								

It is important that the values are entered in the correct order. In the text box next to the label 'Values', enter the first value from the 'exp1' column (5.87). Click on 'Add', and it will appear in the window below.



Repeat the process for all the other values in the 'exp1' column.



	length	number	midpoint	normal	expected	number1	midpt1	exp1
1	1.0-1.9	2	1.45	.04	1.78	5	1.95	5.87
2	2.0-2.9	3	2.45	.08	4.09	5	3.45	7.23
3	3.0-3.9	5	3.45	.14	7.23	8	4.45	9.77
4	4.0-4.9	8	4.45	.20	9.77	12	5.45	10.12
5	5.0-5.9	12	5.45	.20	10.12	9	6.45	8.02
6	6.0-6.9	9	6.45	.16	8.02	5	7.45	4.87
7	7.0-7.9	5	7.45	.10	4.87	6	8.95	3.07
8	8.0-8.9	4	8.45	.05	2.26	.	.	.
9	9.0-9.9	2	9.45	.02	.81	.	.	.

Click on 'OK', and the results will appear in a separate window.

NPar Tests

Chi-Square Test

Frequencies

revised midpoint

	Observed N	Expected N	Residual
1.95	5	6.0	−1.0
3.45	5	7.4	−2.4
4.45	8	10.0	−2.0
5.45	12	10.3	1.7
6.45	9	8.2	.8
7.45	5	5.0	.0
8.95	6	3.1	2.9
Total	50		

Test Statistics

	revised midpoint
Chi-Square(a)	4.292
df	6
Asymp. Sig.	.637

a 1 cells (14.3%) have expected frequencies less than 5. The minimum expected cell frequency is 3.1.

>**Warning** # 3211
>On at least one case, the value of the weight variable was zero, negative, >or missing. Such cases are invisible to statistical procedures and graphs >which need positively weighted cases, but remain on the file and are >processed by non-statistical facilities such as LIST and SAVE.

The value of chi-squared is 4.292, the number of degrees of freedom is 6, and the probability that the data came from a normally distributed population is 0.637.

Notes:

1. In the book, the number of degrees of freedom is reduced to 5, because of the method of estimating the population mean. This can't be included in the test in SPSS.
2. The expected values entered only add up to 48.95 instead of 50 because the tails of the distribution (lengths less than 1.0 mm and greater than 9.9 mm) have not been included. SPSS recalculates the expected values assuming a total of 50, and keeping the same proportions. This explains the different expected values in the results, and the different value of chi-squared (4.292) instead of the calculated value of 4.404.
3. The warning at the end refers to rows 8 and 9 of the table, where SPSS is expecting values (because some columns do have values in these rows), but not getting them.
4. Slightly better results could have been obtained by increasing the number of decimal places in the columns for the Normal probabilities and the expected values, but the expected values can only be entered to four significant figures.
5. The asymptotic significance (Asymp. Sig.) value of 0.637 indicates that there is a high probability that the data comes from a Normally distributed population.