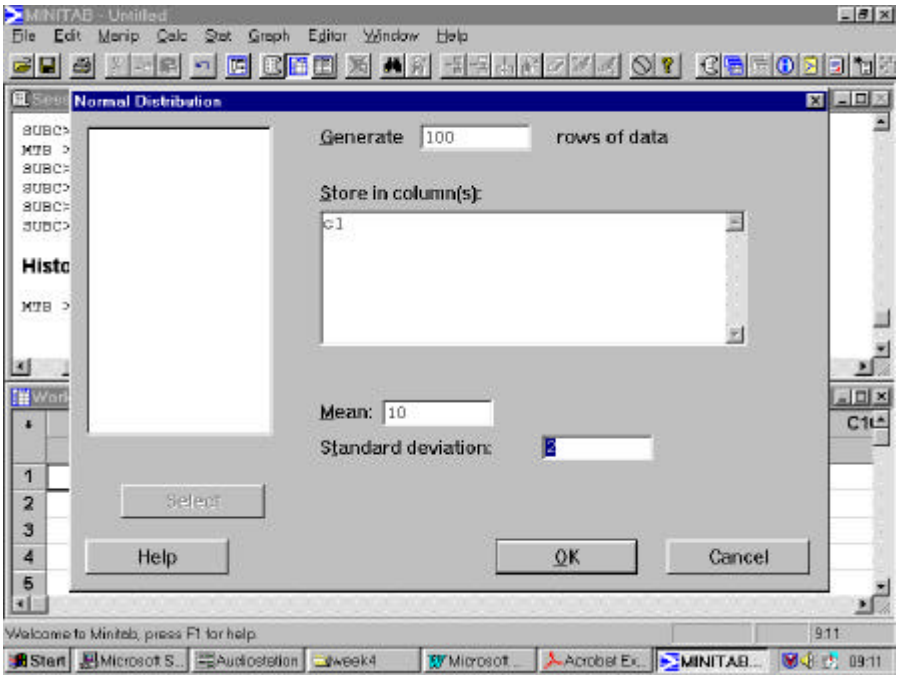
Research Methods 2

Week 4: Exercise Sheet 1

Solution sheet

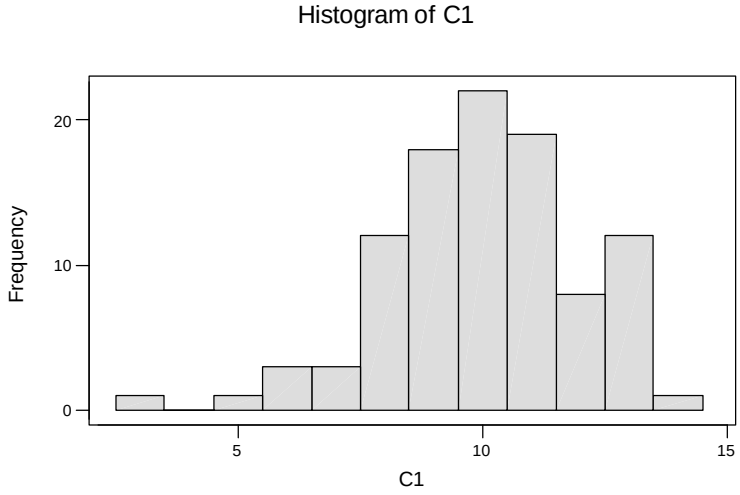
Question 1.

If you follow the sequence of commands specified in the question you will obtain the screen shown below.



Clicking on OK will produce a sample of 100 numbers in column C1, that comprise a sample from the Normal population which has mean 10 and SD 2.

Drawing a histogram of the data in column C1, using the option for drawing a histogram found from the **Graphs...** box in the dialogue box which appears after clicking the sequence **Stat** -> **Basic Statistics** -> **Display Descriptive Statistics**....



(n.b. this was the 'easier' way of producing a histogram described in question 2 of Exercise sheet 2 in Week 3).

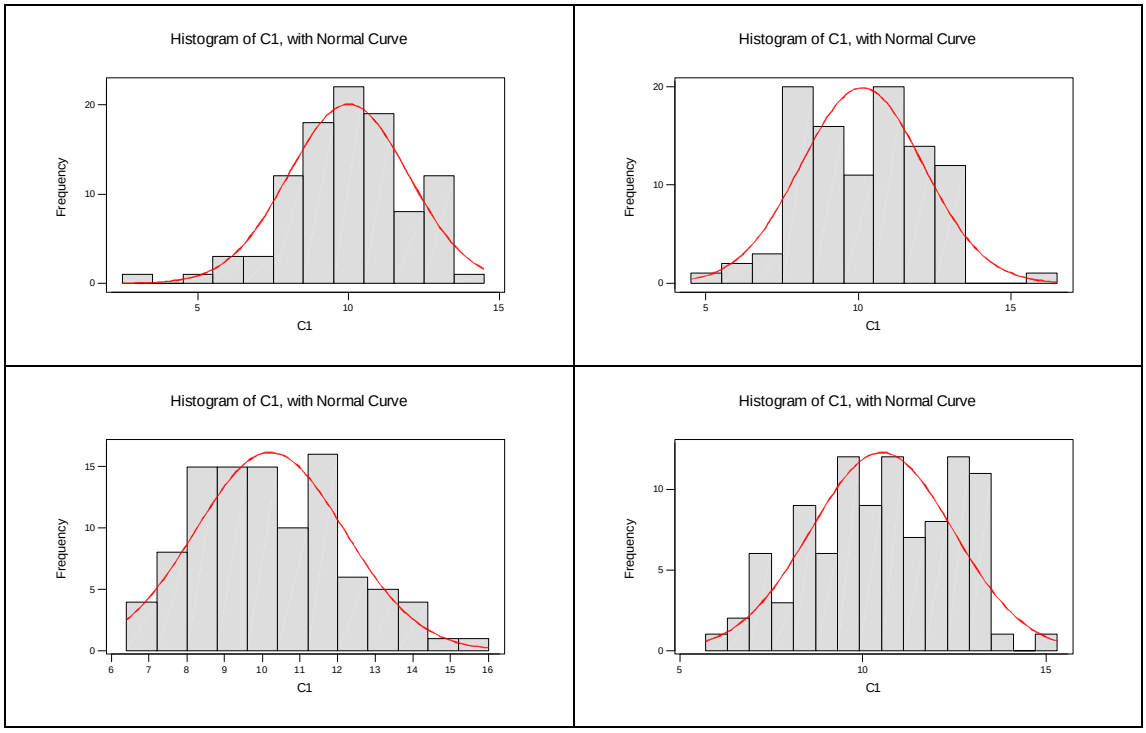
Note that the peak of the histogram occurs around the value 10 and the data are spread between about 5 and 15, with one point below 5.

> Important note: when generating artificial data in this way, it is important to realise that the data you generate will not be exactly the same as the data generated by other students or that used to illustrate the answers in this solution sheet. This is an almost inescapable consequence of this way of working. However, while the details of the answers given here will not coincide exactly with yours, the important features of the solutions will be the same.

You should also notice that although the data *are* Normal (because this is how they have been generated) this particular sample looks a little skew, with a longer tail to the left than to the right. This illustrates the point that you should not expect a sample from a Normal distribution always to look like the text-book picture of the bell-shaped curve that *defines the population*. It also illustrates that in practice it can be quite awkward to assess whether or not a sample plausibly comes from a Normal population. However, we will not now give any further consideration to this important practical problem.

When you produce a histogram from the **Graphs...** option box of the **Display Descriptive Statistics**... command, there is an option to plot the histogram with a Normal curve superimposed.

The above histogram has been redrawn using this facility and is shown in the top left panel below. The other panels show the result of repeating (three times) the command used initially to generate the data. This gives a total of four samples from the Normal population.



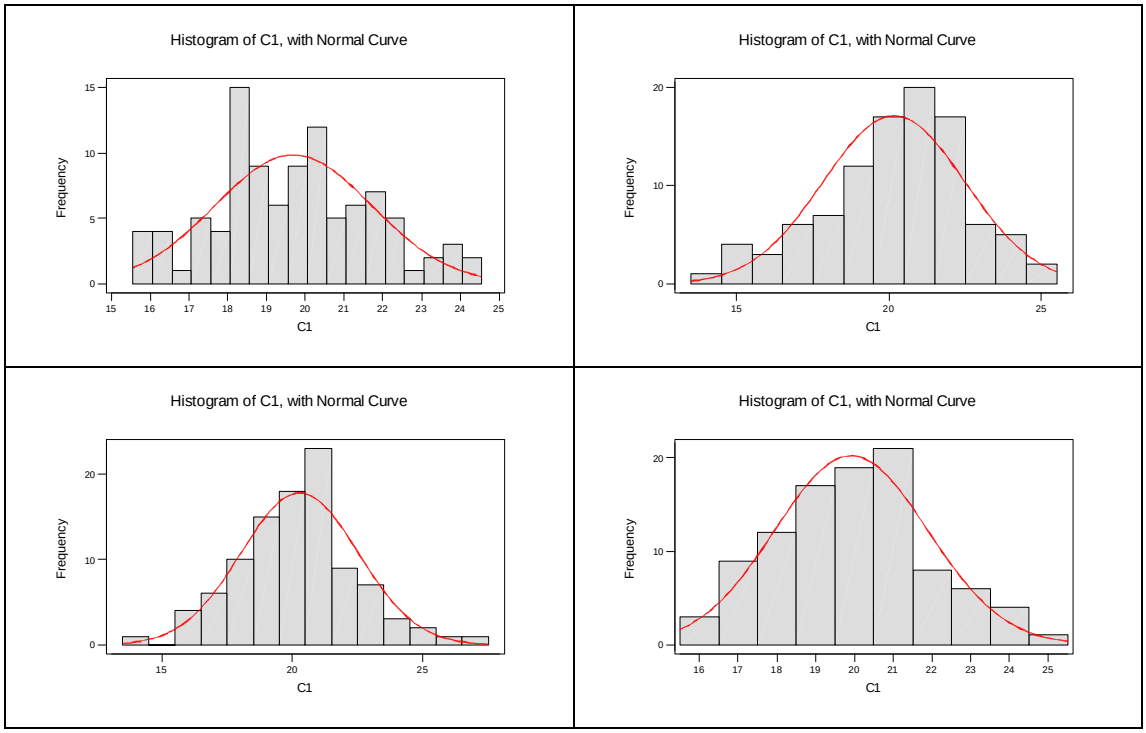
You should note several features of these graphs.

1. The histograms all look rather different. This illustrates that samples, even those as large as 100, can look quite different, even when they are known to have been drawn from the same population. {Note: only by using our technique of artificially generating data from a know population can we observe this. If four histogram such as those above had been produced from real data then the temptation would have been to ascribe the differences, at least in part, to differences in the underlying populations}

2. There are, nonetheless, important similarities. All the samples are centred near 10 and the data are almost always between 5 and 15, with only very occasional points lying outside this range. Moreover the bulk of the data falls between about 8 and 12.

Question 2

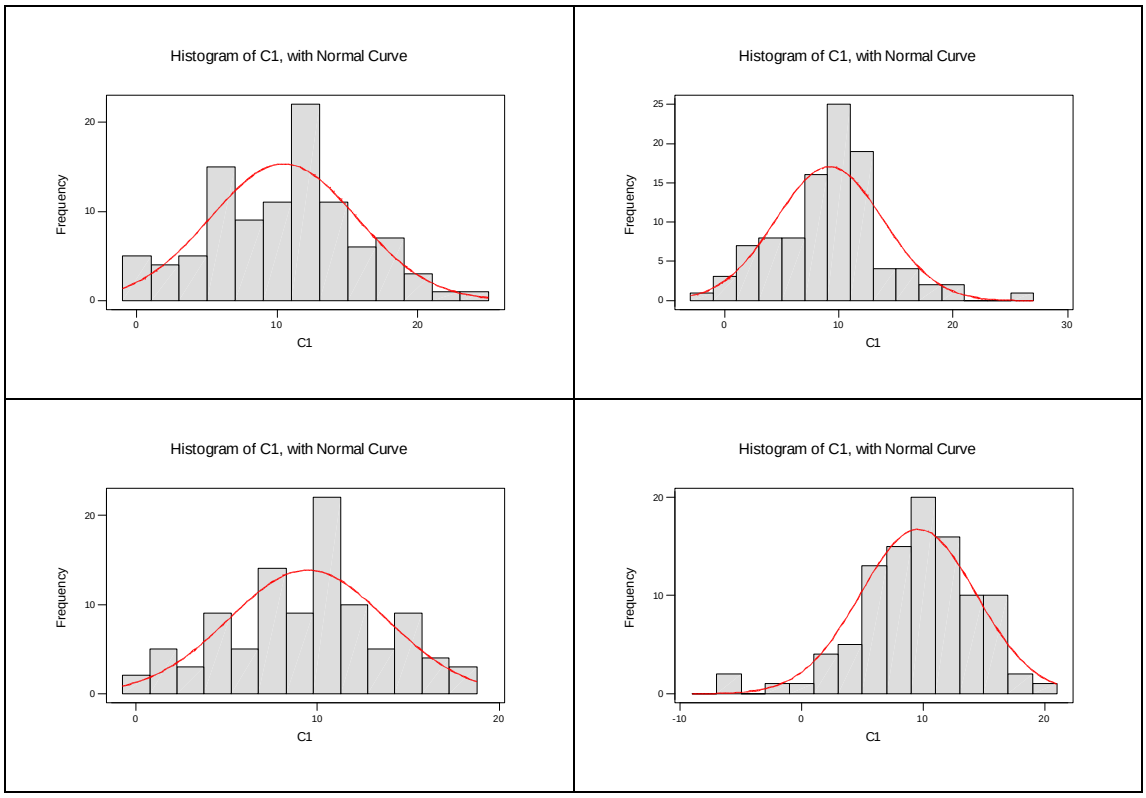
The above exercise can be repeated using a population mean of 20. Only the panel of four histograms is shown. Again, bear in mind that the details of your samples will differ from those shown here.



The broad picture is very similar to that obtained in question 1. In points of detail the histograms look rather different, but they share important features. The main difference here from the histograms in question 1 is that they are all centred on 20, as opposed to 10. The spread about this point is, however, similar to that in question 1. The data are almost always with ±5 of 20 and the bulk of the data are within ±2.

Question 3.

Here the population mean has been reset to 10 and the SD changed to 5. The panel of four histograms is shown below.



Some of the comments made previously apply here too. There are broad similarities between the histogram but also noticeable difference in detail. As in question 1, the histograms are once again centred on 10. However, the data are now much more dispersed around this value than was the case in question 1. Values above 20 and below 0 are encountered. Also, it would no longer be true to say that the bulk of the data are with ±2 of the central value. An interval of width at least 5 on either side of the mean would now be more appropriately quoted.

End of solution sheet