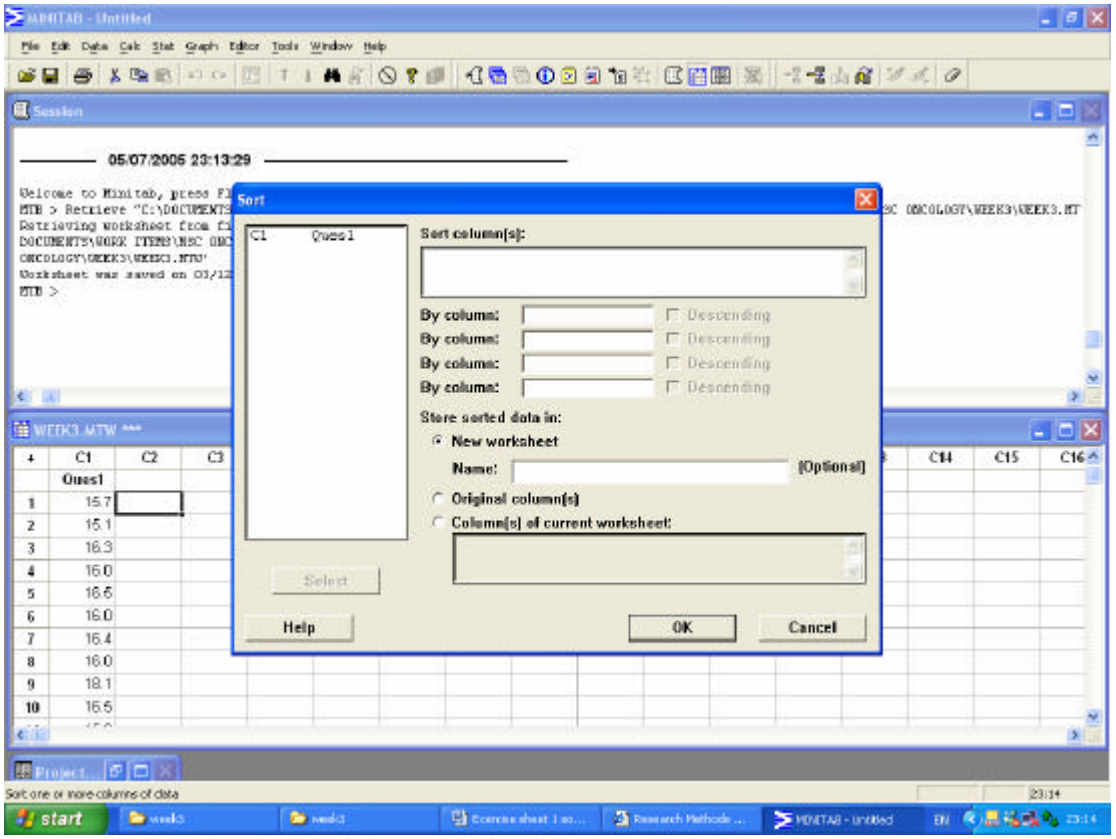
Research Methods 2

Week 3: Exercise Sheet 1, numerical summaries

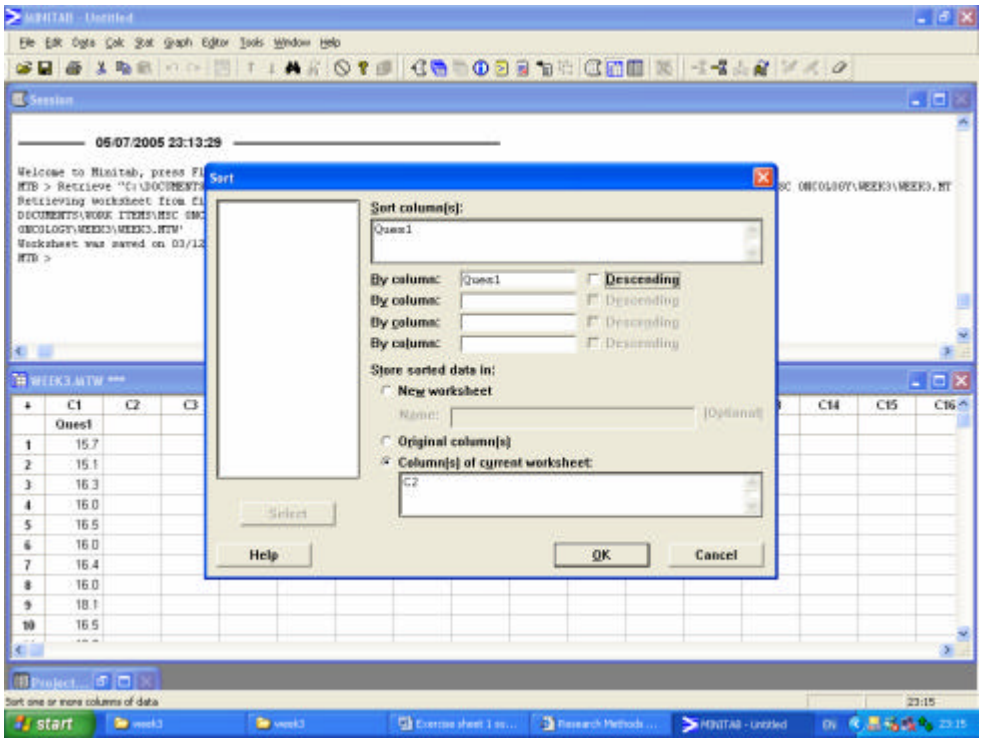
Solution sheet

Question 1.

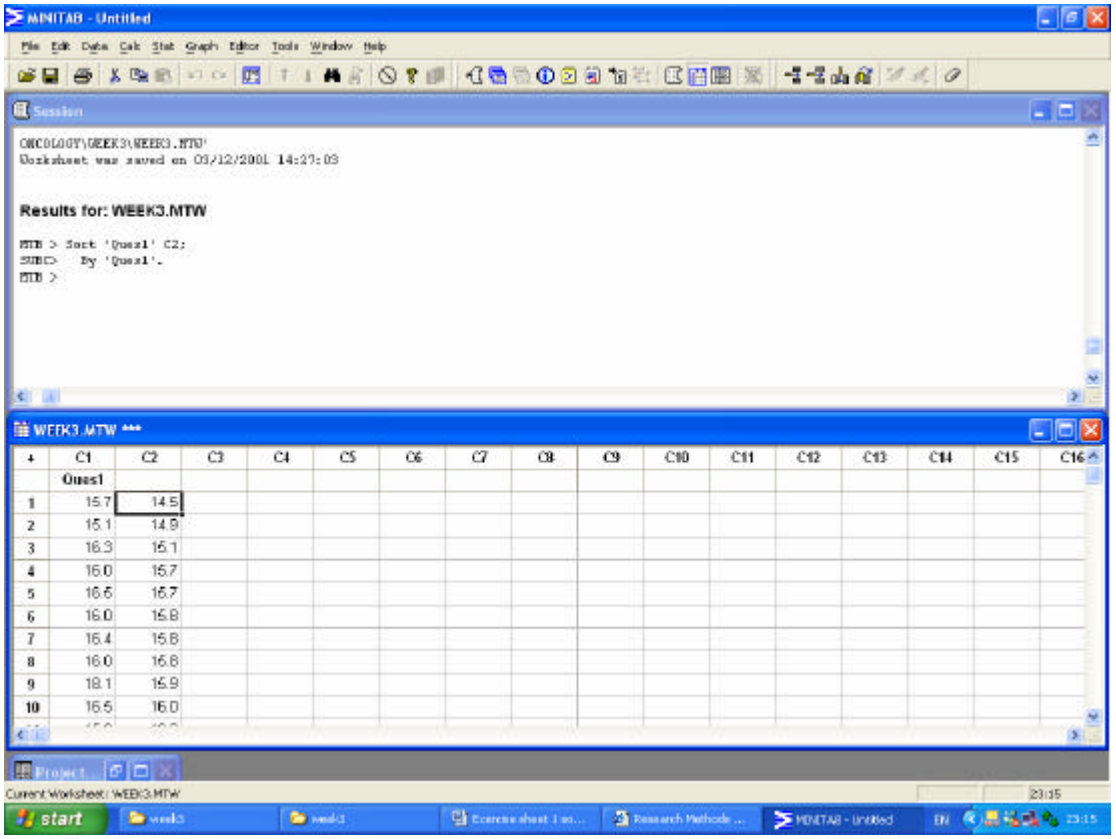
When you click on Data and then Sort, you are confronted with the following dialogue box.



If the boxes are filled in according to the hint, then the following screen will be obtained,



and clicking on OK gives



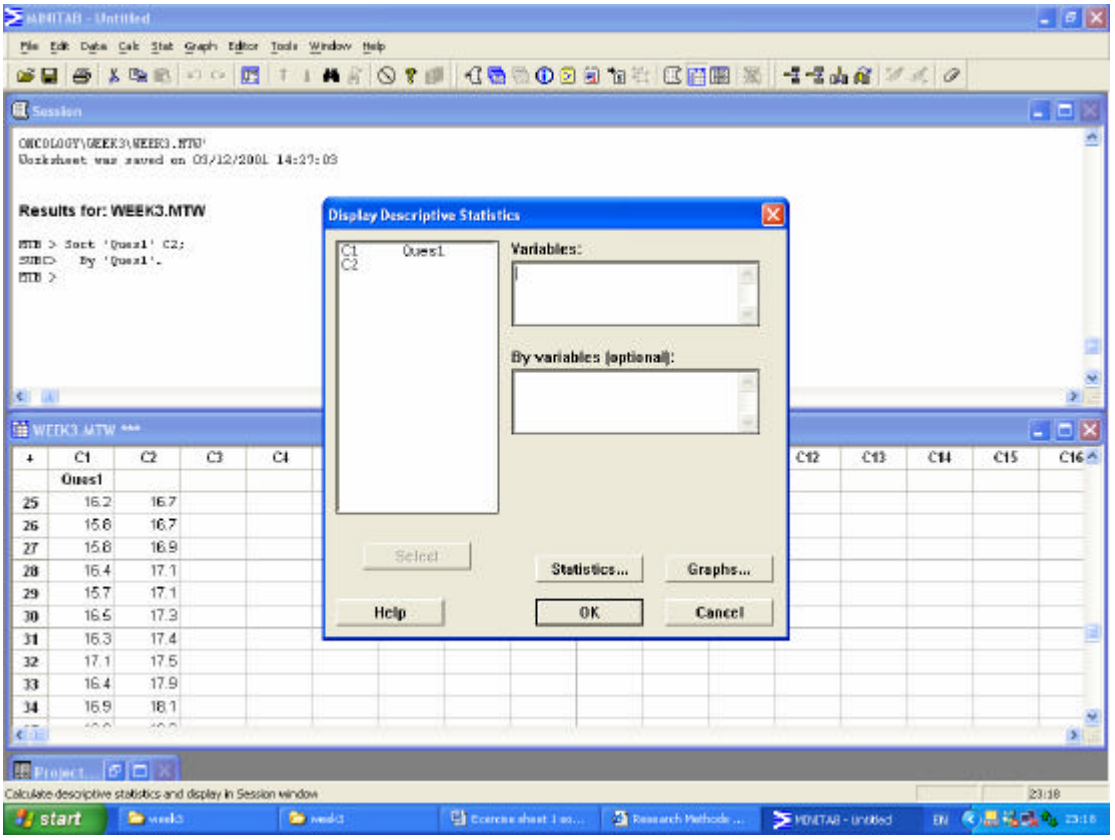
Scrolling down the data window until you reach row 9, you find the value in C2 is 15.9 g/dl. Carrying on until row 18 gives you 16.4 g/dl and in row 27 the value in C2 is 16.9 g/dl. There are 35 values in each of these columns. So there are 17 values less than the value in row 18 of C2 and 17 values bigger than this

value, so the value in row 18 is the middle value and therefore is the median. In terms of the formula given in Appendix 1, the sample size is n = 35, and the median is the ½(n +1) largest value, i.e. the ½ (35+ 1) =18th largest value, which is 16.4 g/dl..

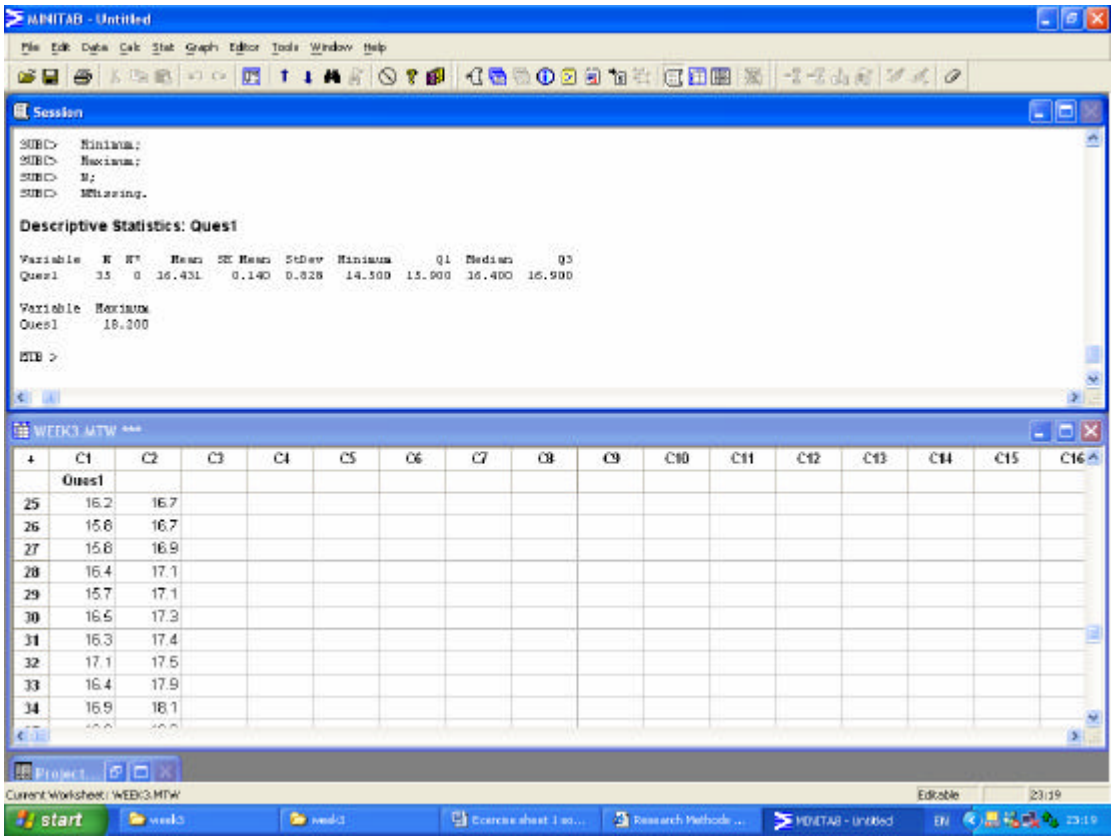
Similarly, the lower quartile is the ¼(n +1)th largest value, i.e. the ¼(35 +1) = 9th value in the sample put in ascending order, which is 15.9 g/dl. Also the upper quartile is the ¾(35 + 1) = 27th value in ascending order in the sample, which is 16.9 g/dl.

Question 2.

If you choose Stat from the main menu bar in Minitab and then select Basic Statistics and then Display Descriptive Statistics... you will be presented with the screen show below.



Double clicking on Ques1 in the left hand box will place Ques1 in the Variables: box. Clicking on OK gives the following screen.



The results of the calculation are shown in the Session window. Several quantities have been calculated. The ones of primary interest are: the Median, which is 16.4 g/dl, as obtained previously, and the quartiles, which Minitab labels Q1, for the lower quartile and Q3 for the upper quartile. These are, respectively, 15.9 g/dl and 16.9 g/d, again agreeing with the results of question l.

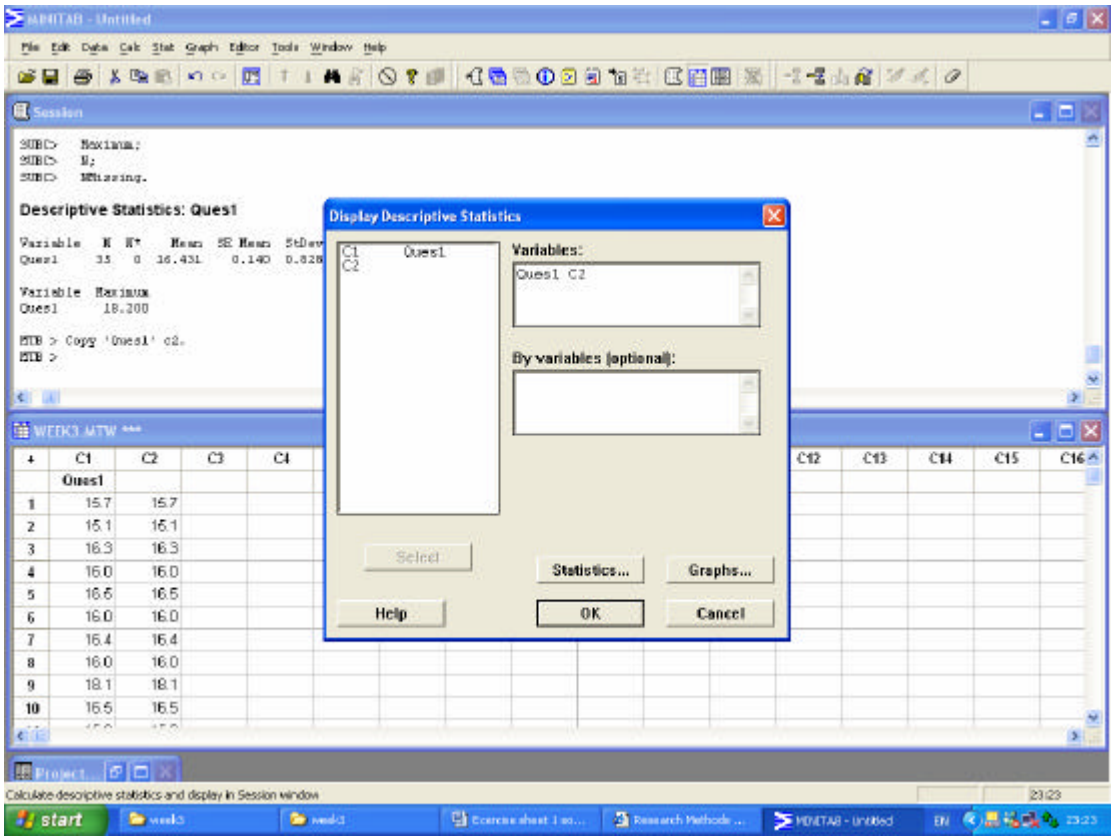
Some of the other results calculated by this command may already be familiar, and in any case soon will be. The mean and standard deviation (labelled StDev) will be described next week and the Standard Error of the mean (SE Mean) will be considered the week after.

The only other values given are the maximum and minimum values in the sample, which together with the median and quartiles allows us to quote the *five number summary* for this sample as:

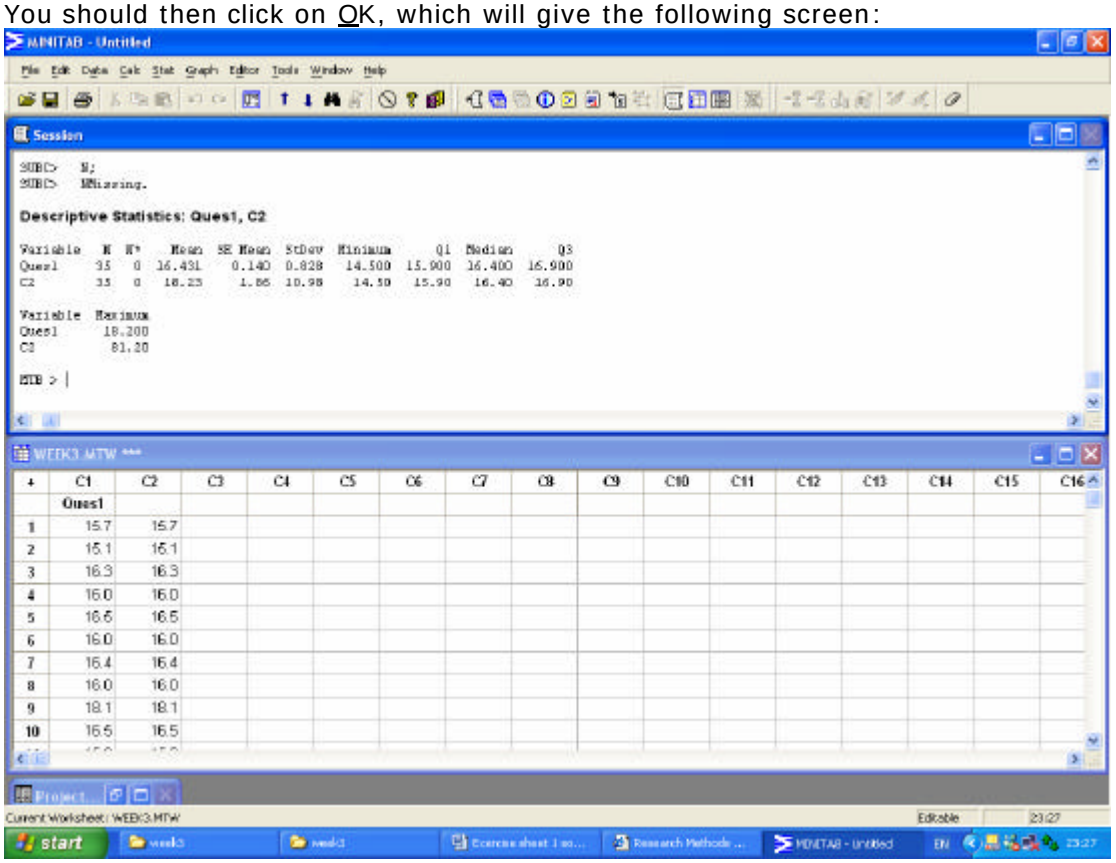
14.5 15.9 16.4 16.9 18.2 (all g/dl).

Question 3.

By following the instructions in the hint, a copy of 'Ques1' is placed in C2 and the final element of the latter changed to 81.2. The method of question 2 requires you to click on Stat from the main menu bar and then select Basic Statistics and then Display Descriptive Statistics.... You then select 'Ques1' as described in the answer to question 2 and then repeat the process to select C2 as well. This will leave the dialogue box looking as follows:



You should then click on OK, which will give the following screen:



It can be seen that the median and quartiles (Q1 and Q3) are unchanged by the perturbation to the largest value in the data. However the mean and standard deviation (StDev) have changed markedly.

This is a feature of the median and quartiles. Once you have placed the sample in ascending order, the median and quartiles are (crudely speaking) located a quarter, a half and three quarters of the way up the sample. If the sample size, n, is such that the point $\frac{3}{4}(n+1)$ of the way up the sample is not the largest value, then changes to the largest value will have no effect on the median or quartiles of the sample. Indeed, all the values above the point $\frac{3}{4}$ of the way up the sample could be changed without any effect on the quartiles or median. The lack of sensitivity of these summaries to changes in the data can be both an advantage and a disadvantage, depending on the circumstances. Note, by way of a contrast, how the mean and standard deviation are very sensitive to changes in the data.

End of solution sheet