Research Methods 2

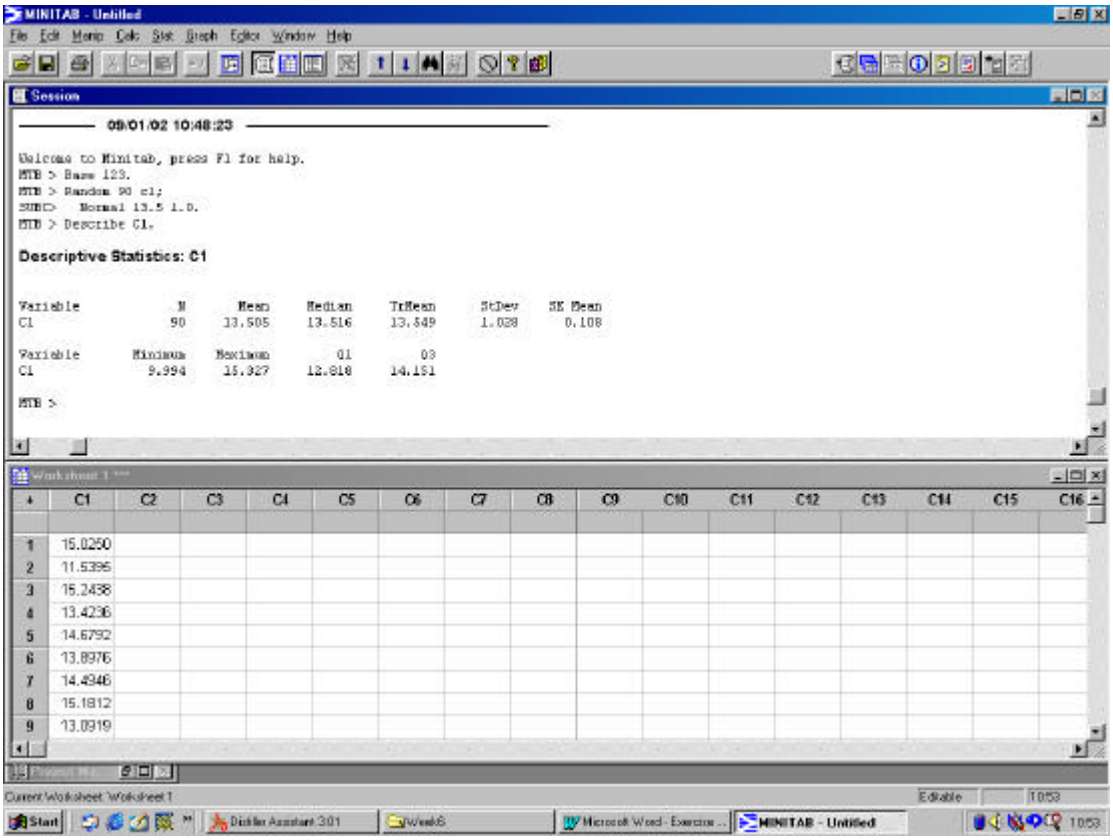
Week 6: Exercise Sheet 1

Solution sheet

[Remember: for questions 1 to 4, the exercises are based on randomly generated data, so while the numerical answers you obtain ought to be quite similar to those given here, they will not agree precisely. However, the important points of principle behind the numerical examples will be the same]

Question 1

Generating data as suggested in the question, and then using **Display Descriptive Statistics...** item gives the following screen.



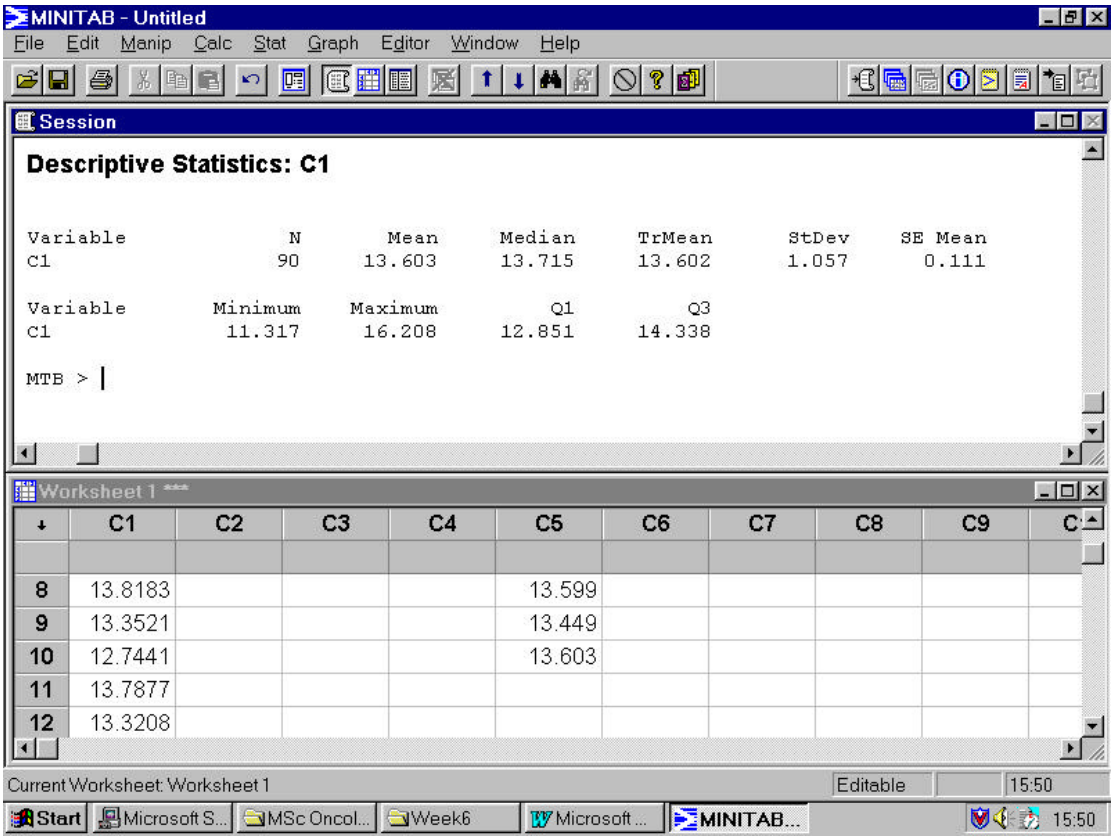
The standard deviation of this sample is 1.028 g/dl. You can estimate the standard error of the mean of this sample by dividing 1.028 by √90, which gives

Standard Error = 1.028/√90 = 1.028 / 9.487 = 0.108.

Notice that this is the same value that Minitab prints under SE Mean. In other words Minitab automatically calculates the Standard error for you.

Note that your sample will have a slightly different SD and SE, but in all cases the result of dividing the SD by √90 will be exactly the SE.

Question 2



The above screen shows the result of generating 10 samples, each of size 90 from the population used in question 1. Each sample is stored in C1, the mean of C1 is found using the well-trodden route through **Stat** → **Basic Statistics** → **Display Descriptive Statistics**. The mean is found and copied into the next available row of column C5. The screen shows the situation where C1 holds the tenth sample to be generated, the Session window gives the statistics for C1, including the mean of 13.603, which has been copied into the last row of C5.

At this stage C5 contains the means of ten samples, each of size 90, from the Normal population with mean 13.5 and SD 1. The next stage is to find the sample SD of these ten means. This is done by applying **Display Descriptive Statistics** to C5. This gives output

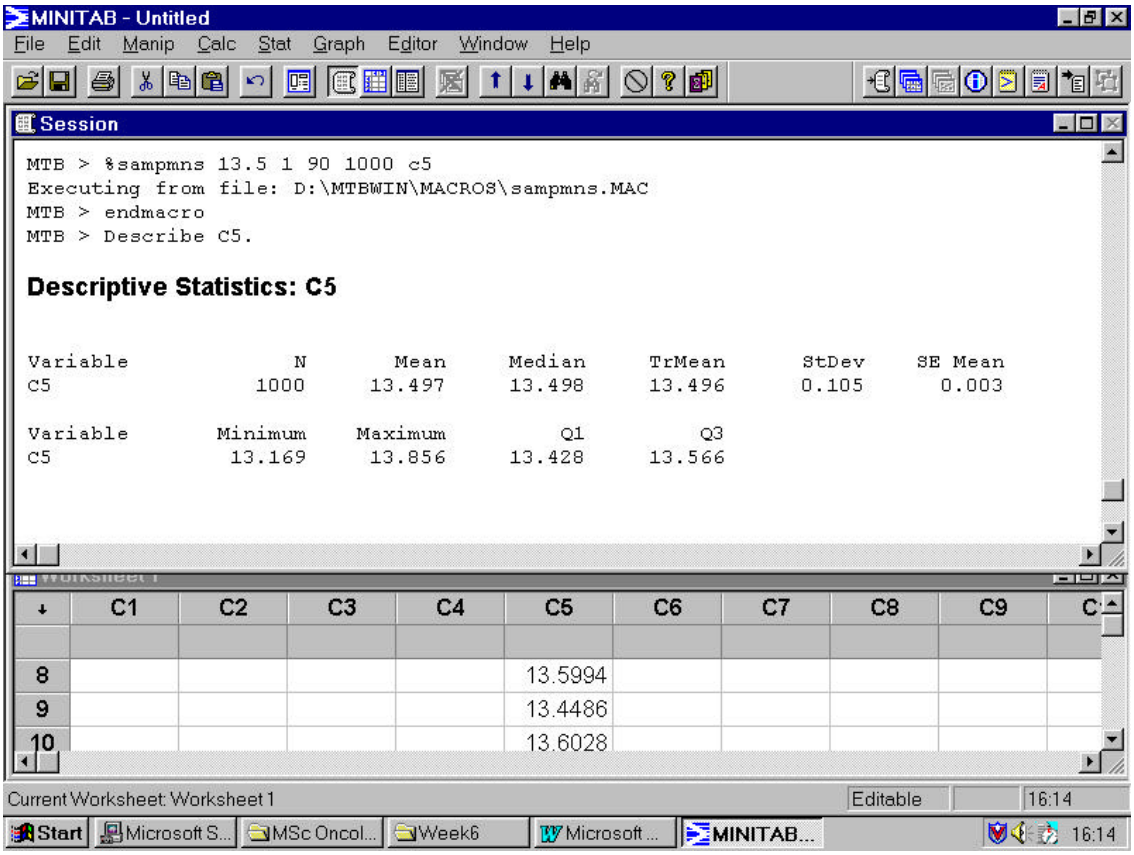
Descriptive Statistics: C5

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C5	10	13.560	13.590	13.568	0.098	0.031

Variable	Minimum	Maximum	Q1	Q3
C5	13.364	13.694	13.491	13.635

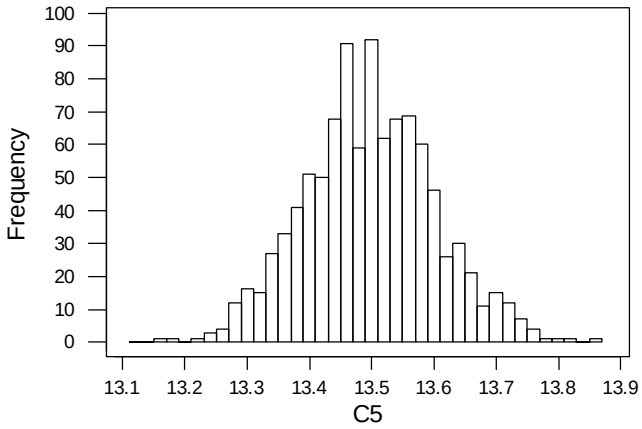
Thus the SD of this collection of sample mean is 0.098 g/dl. This is an estimate of the SE of the Mean of a single sample and is similar to that obtained in question 1, namely 0.108 g/dl. {note that the SE Mean entry in the above (i.e. 0.031) has no meaning as C5 is not raw data but a collection of sample means}.

Question 3



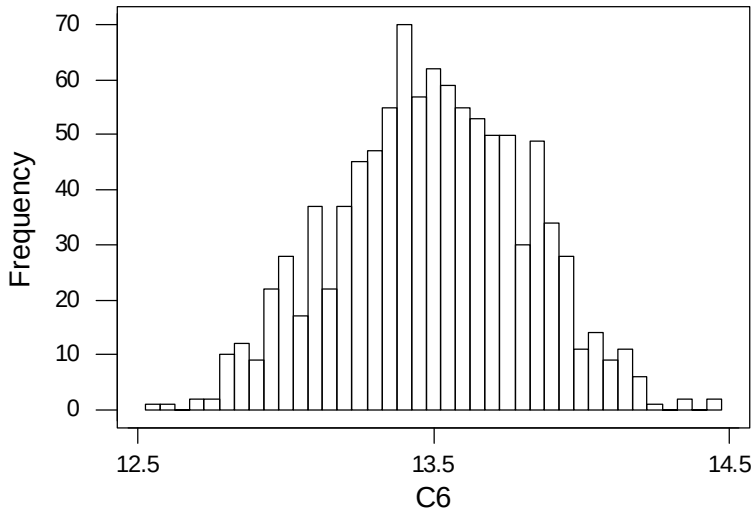
By using the macro sampmns macro the above screen is obtained and the SD of the 1000 sample means is 0.105 g/dl. As this SD is based on 1000 means, rather than the 10 used in question 2, it should provide a more precise estimate of the SE of the mean. Indeed, this is the case as the estimate based on 1000 means, namely 0.105 g/dl is closer to that found from the formula used in question 1 (0.108) than the value based on just ten means in question 2.

The histogram of the 1000 sample means is below



The histogram shows that if you take the mean of a sample of 90 haemoglobin concentrations, then this will very likely be within 0.4 g/dl of the population mean. This is because the 1000 means you have generated all lie within a range of 0.8 g/dl, between 13.1 and 13.9 g/dl. Moreover, most of them lie in a narrower central band and the sample means themselves seem to follow a Normal distribution.

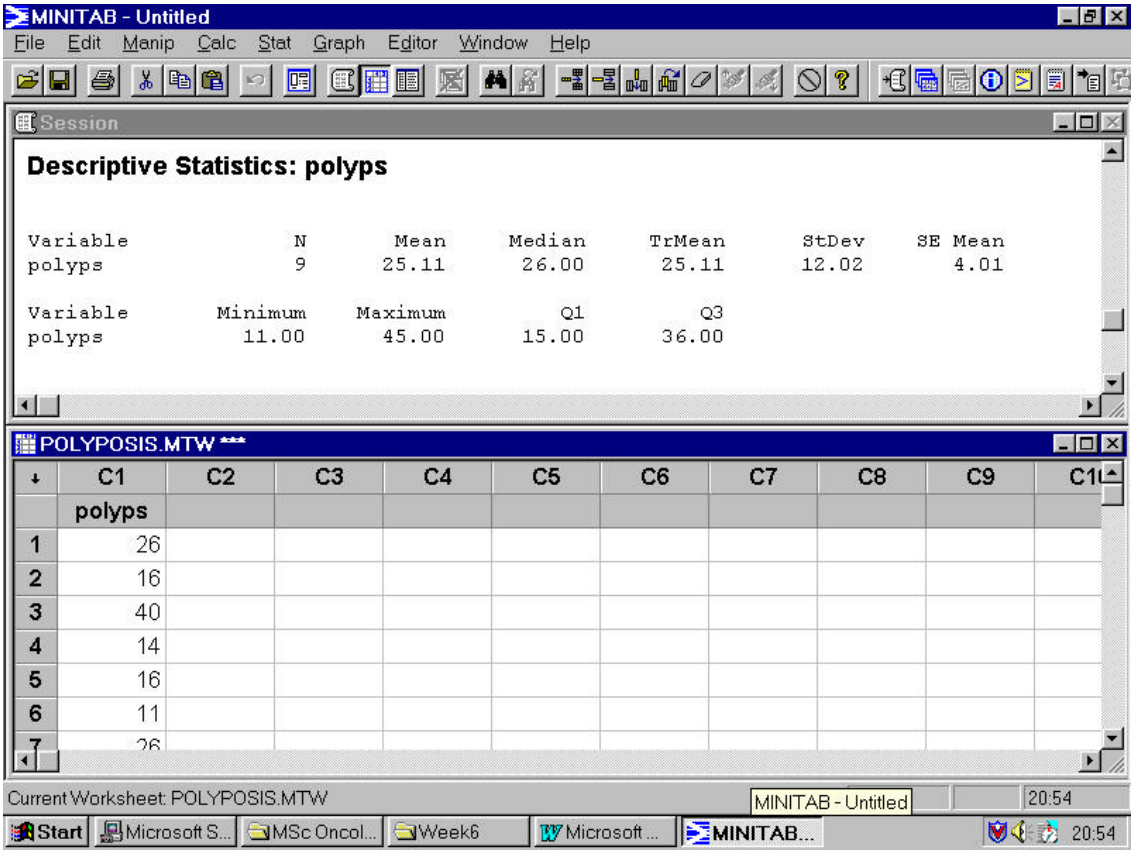
The following histogram is analogous to the one above but shows means of samples of size 9



The broad picture is the same, namely the sample means seem to have a Normal distribution. It is also centred on what we know is the population mean (13.5 g/dl). However, it is much more dispersed than the previous histogram. Therefore, if you take a mean of a sample of size 9, rather than 90, then it will very likely lie within 1 g/dl of the population mean, as the above shows that the 1000 means you have generated lie between 12.5 and 14.5 g/dl.

Question 4

If you type the data given in the question into column C1 and then apply **Stat** → **Basic Statistics** → **Display Descriptive Statistics** to C1, you get the following screen

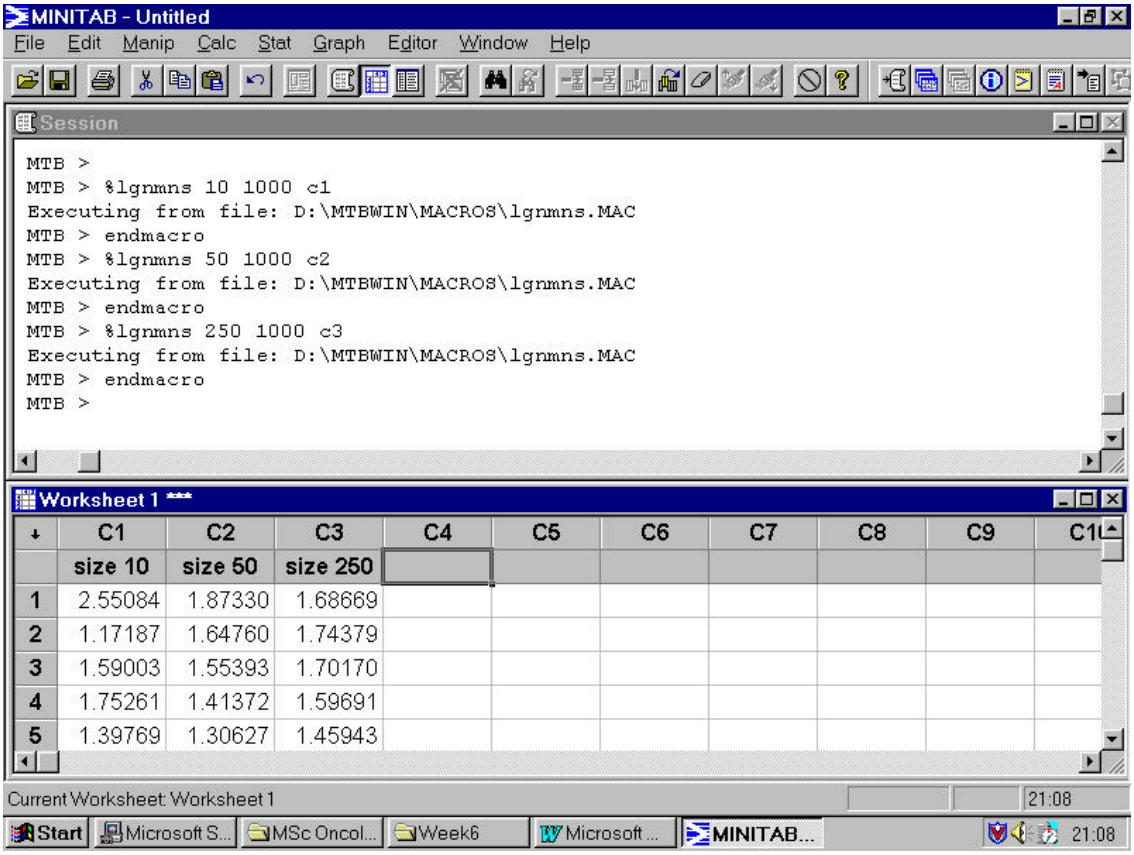


The mean number of polyps per patient in this sample is 25.1. The SE of this mean is 4.01. Therefore the mean of the population from which the sample has been drawn is estimated to be 25.1. The standard error, which measures how good is this estimate of the mean, is 4.01.

Just quoting a SE in this way is not very meaningful. More helpful ways of using the SE to quantify the uncertainty in a sample mean will be discussed next week.

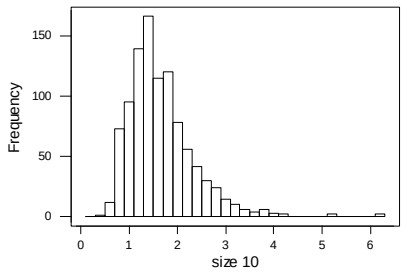
Question 5

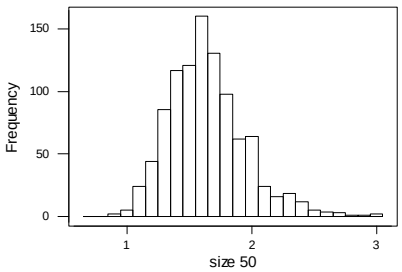
Running the macro three times as described in the question, and saving the three sets of means in columns C1, C2 and C3 (which have been names 'size 10', 'size 50' and 'size 250' respectively) gives the following screen.

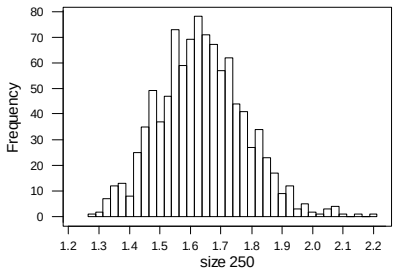


	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
	size 10	size 50	size 250							
1	2.55084	1.87330	1.68669							
2	1.17187	1.64760	1.74379							
3	1.59003	1.55393	1.70170							
4	1.75261	1.41372	1.59691							
5	1.39769	1.30627	1.45943							

Histograms of these three sets of means are shown below.







The means are based on samples of size 10 (top left) , 50 (above) and 250 (left). As you would expect, the larger the sample size, the more concentrated the distribution. However, the important observation is that as the sample size gets larger, the shape of the distribution of means gets closer and closer to that of a Normal distribution.

In the study document and in the previous questions, the underlying population was always Normal. If this is the case then distribution of the sample means will also follow a Normal distribution. This is fine if the population is Normal, but what about other forms of population, such as that which is illustrated by the histogram in this question?

This question illustrates an important feature of the distribution of sample means which was not covered in the study document. This is the fact that the distribution of sample means is approximately Normal *whatever the shape of the distribution of the individual observations.* The approximation gets closer as the size of the sample gets bigger, as illustrated by the three histograms above.

A consequence is that methods for statistical inference based on assumptions of Normality have wider applicability than might be thought.

End of solution sheet