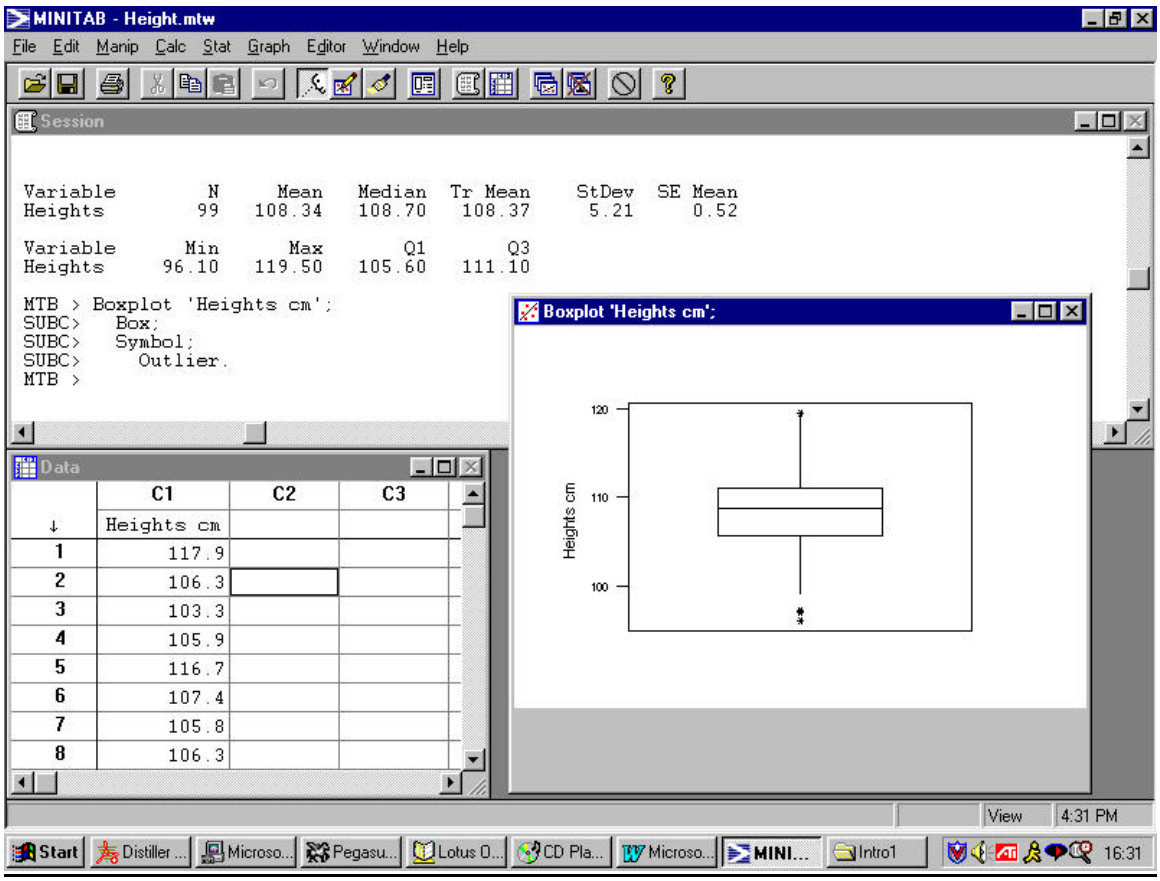
MD/PhD Course in Medical Statistics

Department of Statistics

Practical session on Minitab, descriptive statistics and the Normal distribution: outline solutions

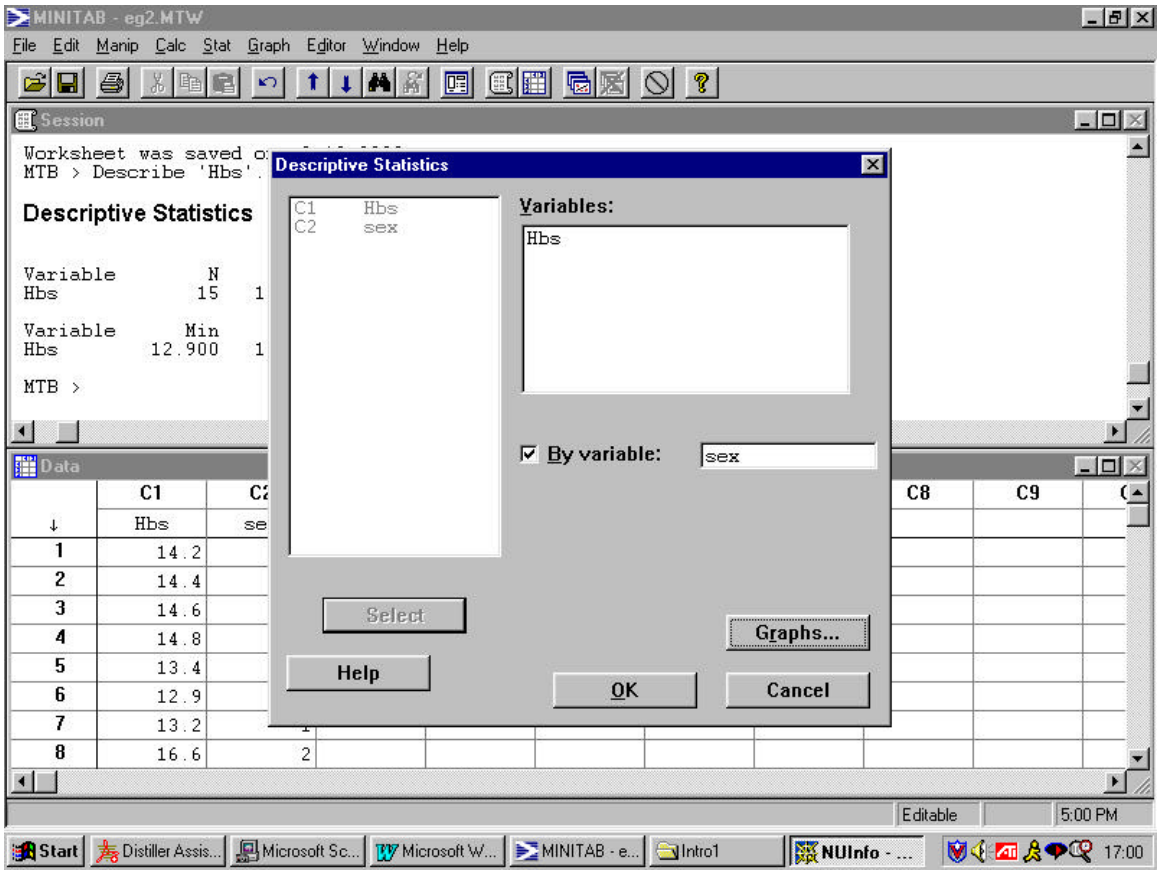
1.

The file HEIGHT.MTW is opened by clicking on File → Open Worksheet and then selecting the file in the usual Windows manner. The screen below shows the results of selecting Stat → Basic Statistics → Descriptive Statistics… and entering C1 (or 'Heights cm') in the Variables: box and clicking on OK. The box and whisker plot comes from Graph → Boxplot… with C1 under Y in the Graph variables:… box.



The mean (labelled Mean) and standard deviation (labelled StDev) appear in the output in the session window. The median is labelled as such and the lower and upper quartiles are labelled Q1 and Q3 respectively. Note that the Describe command gives other information, such as the standard error and the minimum and maximum (`Tr Mean` is a trimmed mean, a quantity we will not use, that attempts to combine the advantages of a mean and a median)

2.
The data can be typed into the data window and the column for haemoglobin concentrations is named as Hbs. Applying the method for finding means used in question 1 gives a mean of 15.253 g/dl and an SD of 1.461 g/dl. If the second column of data is entered in a column which has name 'sex' then the way to obtain separate means and SDs for the sexes is to click OK on the screen as set out below.



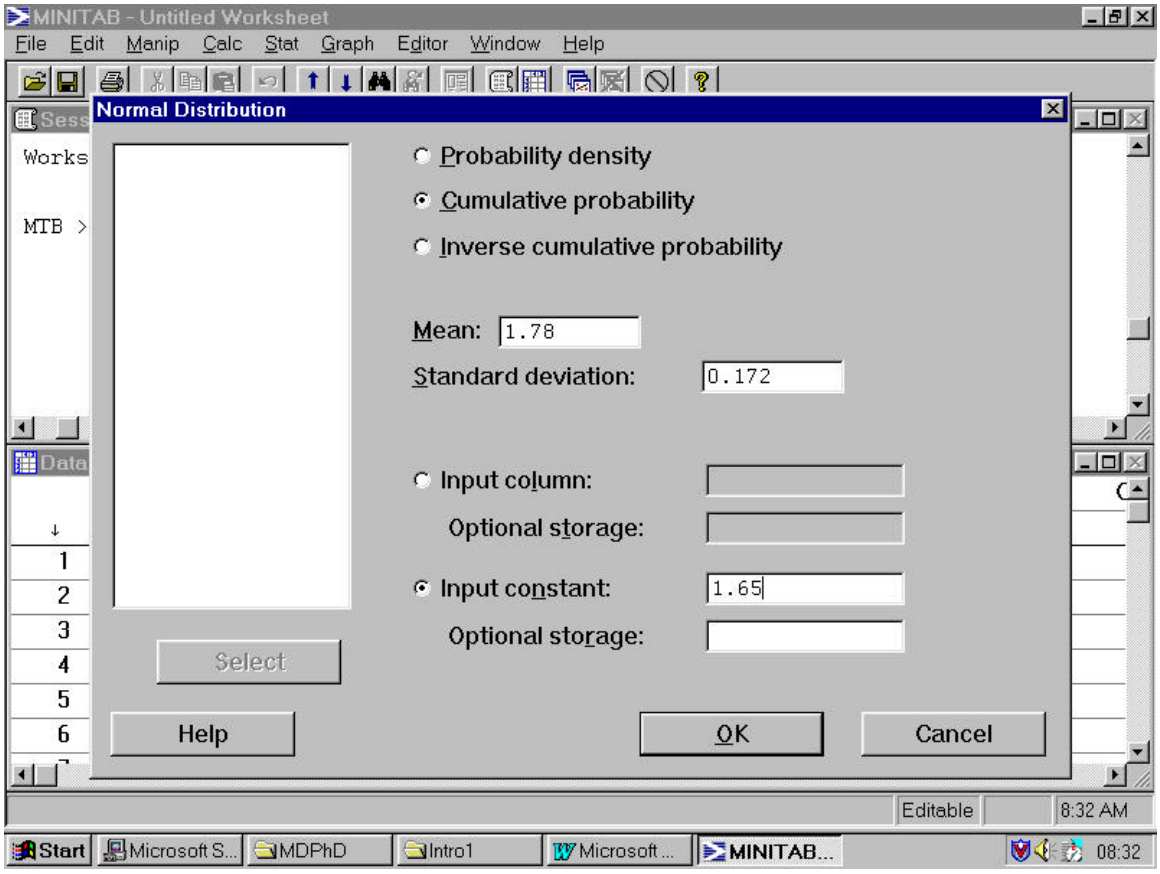
Doing this gives the following output in the session window.

```
MTB > Describe 'Hbs';
SUBC>   By 'sex'.
```

Descriptive Statistics

```
Variable  sex        N     Mean   Median  Tr Mean    StDev  SE Mean
Hbs       1          7   13.929   14.200   13.929    0.750    0.283
          2          8   16.413   16.200   16.413    0.706    0.250

Variable  sex          Min      Max       Q1       Q3
Hbs       1         12.900   14.800   13.200   14.600
          2         15.800   17.800   15.825   16.900
```

3.
Clicking on the path suggested in the question leads to the following dialogue box, which should be filled in as shown.



On clicking OK the following is written to the Session window.

```
MTB > CDF 1.65;
SUBC>   Normal 1.78 0.172.
```

Cumulative Distribution Function

```
Normal with mean = 1.78000 and standard deviation = 0.172000

          x     P( X <= x)
     1.6500         0.2249
```

Which shows that 0.2249, or about 22½% of this population lies below 1.65 log units.

4.
The answer to this question is found in a way that is almost the same as for question 3. The dialogue box is that shown in the answer to question 3 but the Inverse cumulative probability button, rather than the Cumulative probability button should be checked. The value 77.5% must be entered in the Input constant box as 0.775, i.e. as a probability or proportion, not a percentage. Doing this and clicking on OK gives the following output in the Session window:

```
MTB > InvCDF 0.775;
SUBC>   Normal 1.78 0.172.
```

Inverse Cumulative Distribution Function

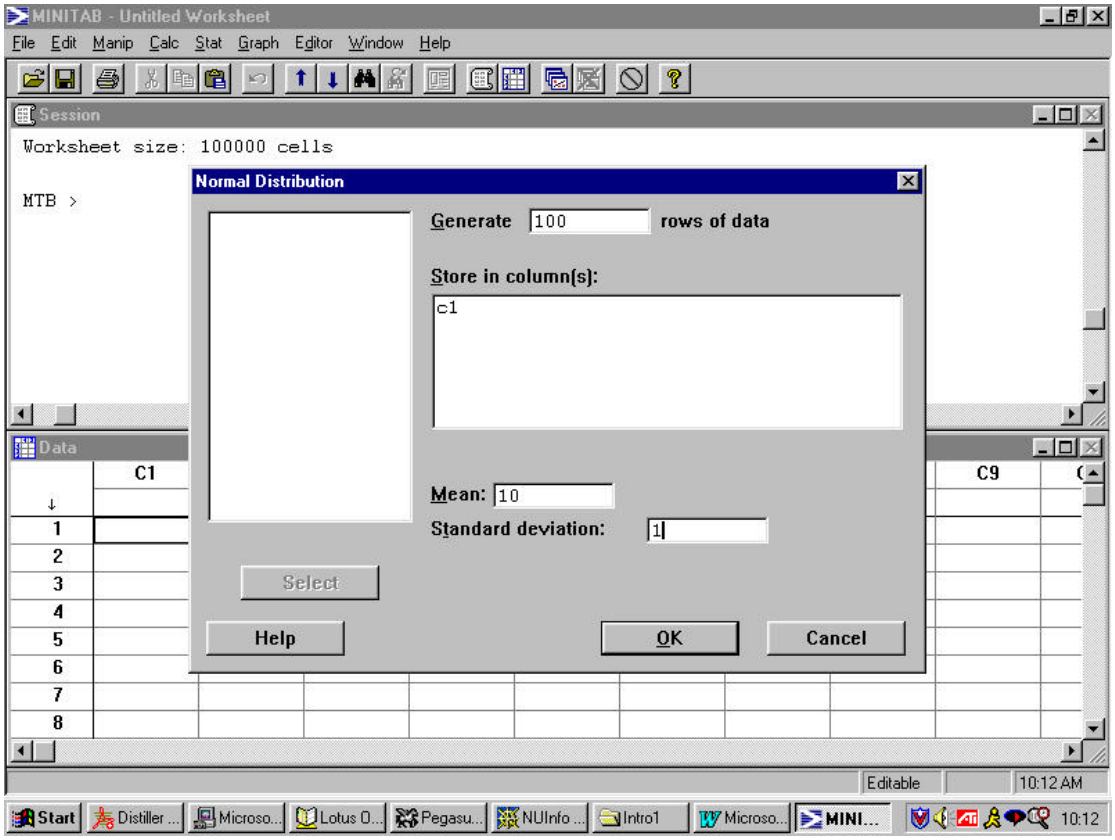
```
Normal with mean = 1.78000 and standard deviation = 0.172000

 P( X <= x)          x
    0.7750       1.9099
```

Thus 77.5% of the population have a value of the variable measuring photo-toxicity that is less than approximately 1.91 log units.

The proportion above this value is 1-0.775 = 0.225, i.e. the same proportion that is below 1.65 (cf. question 3). From the symmetry of the Normal distribution this implies that 1.91 is the same distance above the mean as 1.65 is below it. As the mean is 1.78, this is indeed the case, 1.78-1.65 = 1.91-1.78 = 0.13.

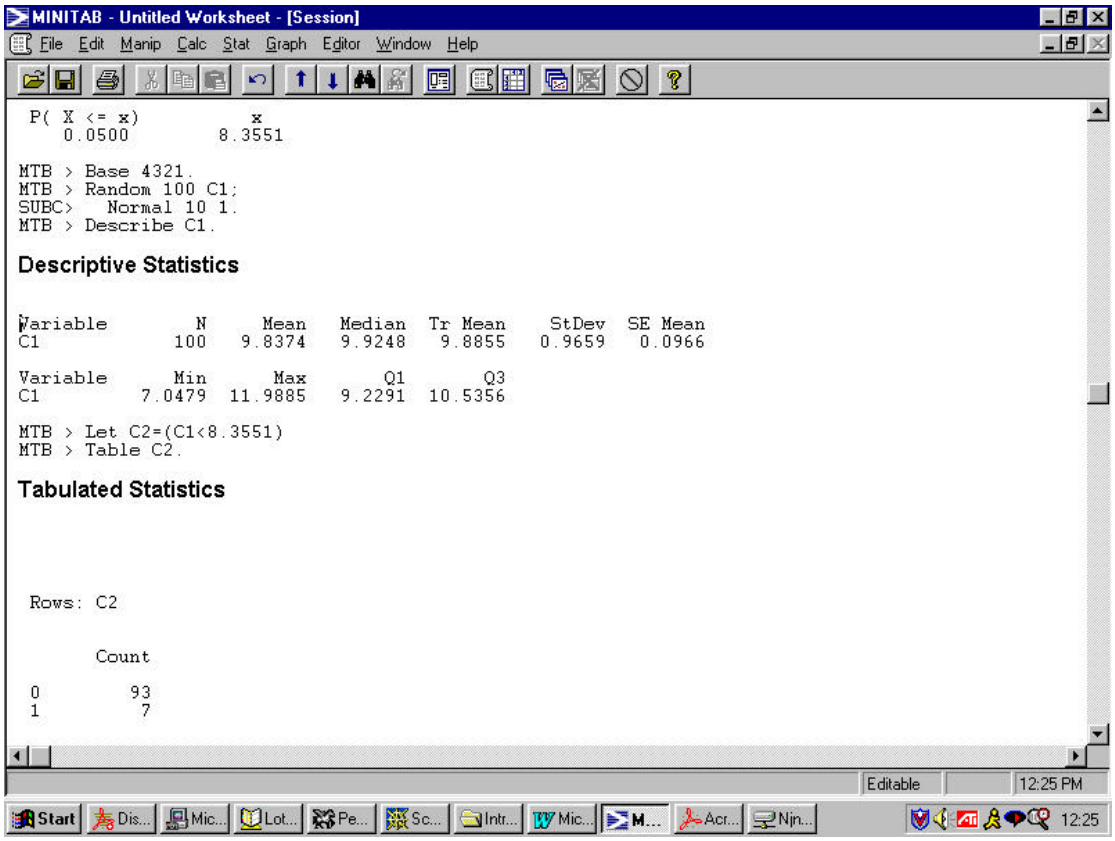
5.

Using the method of question 4, with an Input constant box of 0.05 gives a threshold of 8.3551.
Following the procedure for generating random numbers given in the question leads to the following screen:



Clicking on OK produces 100 Normal numbers from a *population* with mean 10 and SD 1 (i.e. this is a special case where we know m=10, s=1).

Computing descriptive statistics using the methods of questions 1 and 2, and then using the method suggested in the question to determine the number of values less than the threshold and computing a classification table leads to the following session window.



Notice that the *sample* mean and SD, *m* and *s*, are not exactly 10 and 1, but vary from these values by sampling variation (their values are, respectively, 9.8374 and 0.9659).

The table shows that 7 of the 100 values are below 8.3551, so 7% rather than 5% fall below the 5% threshold. This is only to be expected - the number falling below 8.3551 will, *on average*, be 5%, it just will not be guaranteed to be 5% in every sample. This can be appreciated by repeating the instruction to form a column of 100 numbers from a Normal population of mean 10 and SD 1. This gives the sample statistics and number below the threshold shown overleaf: the percentage below the 5% threshold is now 3%.

Descriptive Statistics

```
Variable        N     Mean   Median  Tr Mean    StDev  SE Mean
C1            100   10.037    9.936   10.013    0.904    0.090

Variable      Min      Max       Q1       Q3
C1          7.888   12.680    9.557   10.433
```

```
MTB > Let C2=(C1<8.3551)
MTB > Table C2.
```

Tabulated Statistics

```
Rows: C2

       Count

 0        97
 1         3
 All     100
```

If the instruction to form a column of 100 values is replaced by a similar instruction to form 10000 values then the descriptive statistics obtained are:

```
MTB > Random 10000 C1;
SUBC>   Normal 10 1.
MTB > Describe C1.
```

Descriptive Statistics

```
Variable        N     Mean   Median  Tr Mean    StDev  SE Mean
C1          10000   10.012   10.008   10.012    1.004    0.010

Variable      Min      Max       Q1       Q3
C1          5.693   13.562    9.333   10.674
```

Note how much closer the sample mean and variance, 10.012 and 1.004, are to their population counterparts, 10 and 1, in this much larger sample.

Using exactly the same methods as before shows that 470 of the 10,000 values are below 8.3551, i.e. 4.7% of this sample is below the 5% point.

{note that because this question uses randomly generated values, the answers above will not, in general, be the same as those shown here. However, in the previous copy of the MINITAB screen the Session window contains the line:

MTB > Base 4321.

This sets the random number seed to a specific value and this in turn means that if you type this command before generating the random numbers then the first two sets of 100 numbers and the subsequent 10,000 should be the same as used here. You do not have to type the Base command when using random numbers in MINITAB, it is just that if you do you set the generator to a reproducible state.}