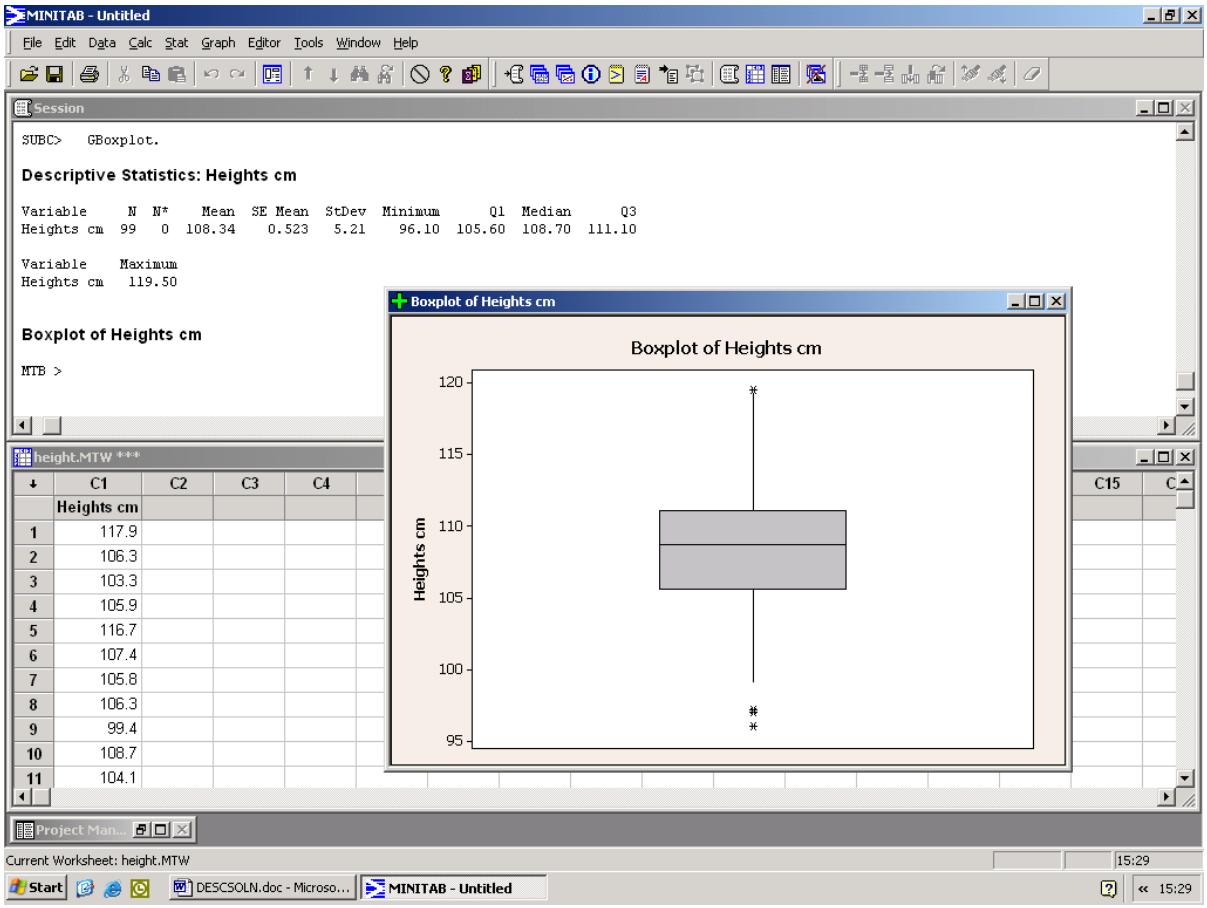
MRes in Medical Statistics
MMB8028

School of Mathematics and 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. When the box stating that a "*copy of the content of this file will be added to the current project*" appears, click OK The screen below shows the results of selecting Stat → Basic Statistics → Display Descriptive Statistics… and entering C1 (or 'Heights cm') in the Variables: box. The boxplot is obtained by clicking on the Graphs… button and then selecting Boxplot of data and then clicking on OK (twice).

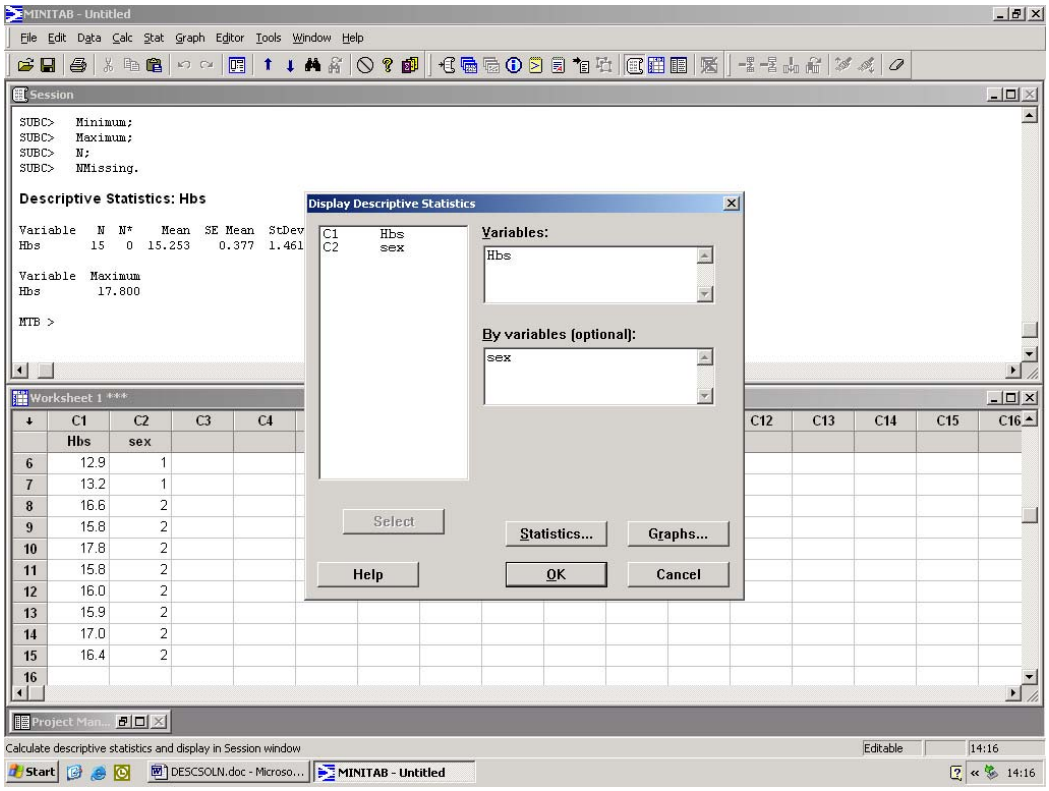


The mean (labelled Mean) and standard deviation (labelled StDev) appear in the output in the session window. The median is labelled as such and the lower and upper quartiles are labelled Q1 and Q3 respectively. Note that this command gives other information, such as the standard error and the minimum 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)

The boxplot could have been obtained independently by clicking on Boxplot… under the Graph item on the menu.

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';
SUBC>   Mean;
SUBC>   SEMean;
SUBC>   StDeviation;
SUBC>   QOne;
SUBC>   Median;
SUBC>   QThree;
SUBC>   Minimum;
SUBC>   Maximum;
SUBC>   N;
SUBC>   NMissing.
```

Descriptive Statistics: Hbs

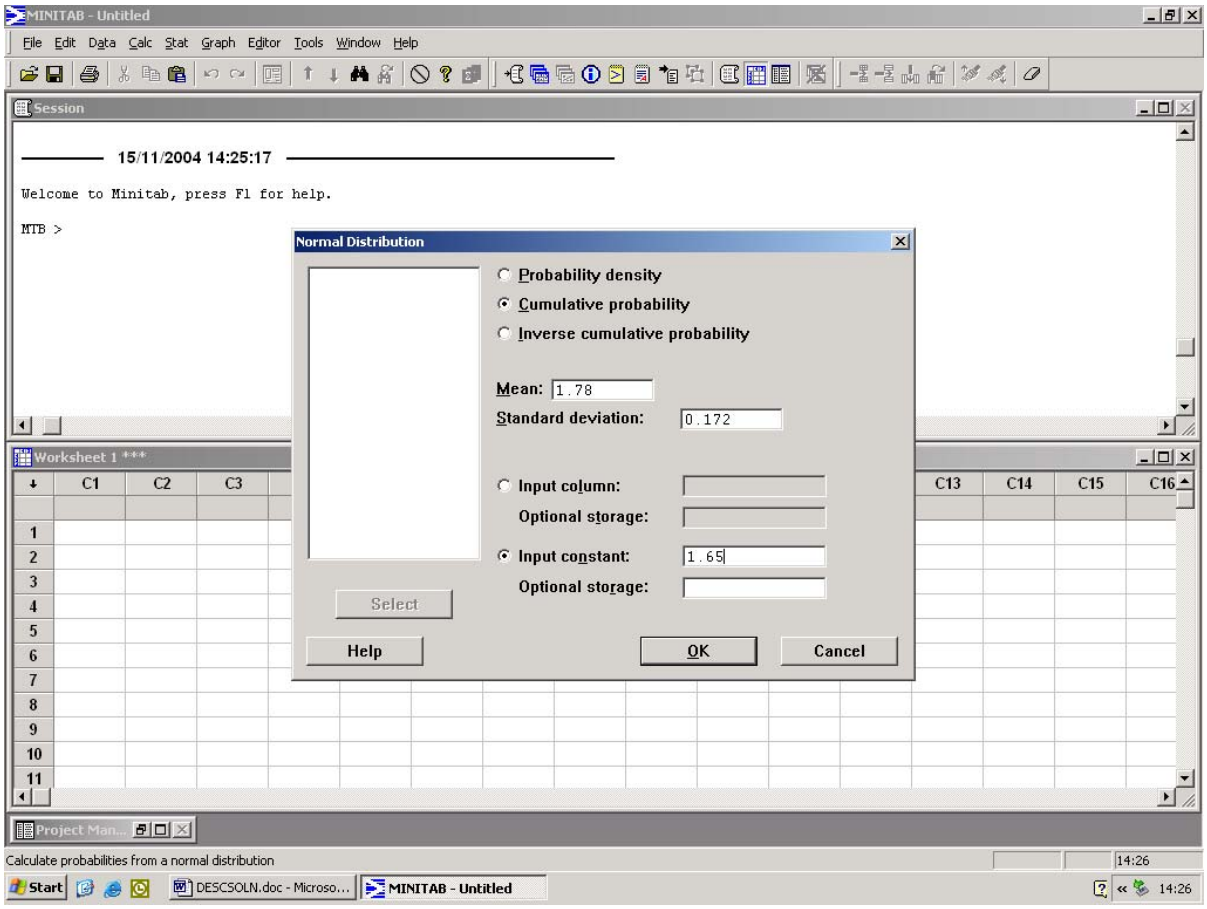
Variable	sex	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Hbs	1	7	0	13.929	0.283	0.750	12.900	13.200	14.200	14.600
	2	8	0	16.413	0.250	0.706	15.800	15.825	16.200	16.900

Variable	sex	Maximum
Hbs	1	14.800
	2	17.800

All the descriptive statistics are now shown separately for the sexes.

[†] Note that the output shown (here and in other examples) is only displayed if the commands prompt is enabled.

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 = 1.78 and standard deviation = 0.172

   x  P( X <= x )
1.65     0.224880
```

This shows that a proportion 0.22488, 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

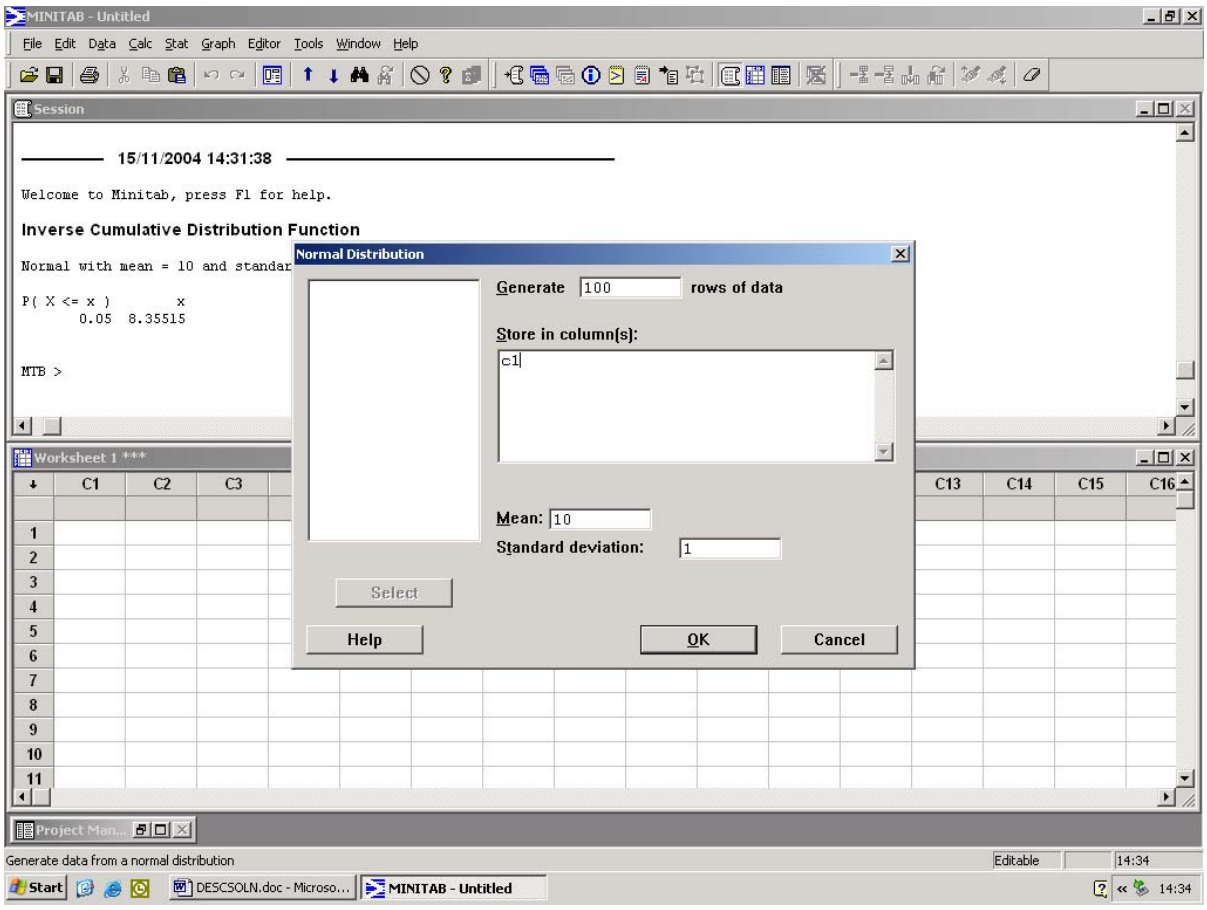
```
Normal with mean = 1.78 and standard deviation = 0.172

P( X <= x )        x
      0.775  1.90993
```

Thus 77.5% of patients in 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.

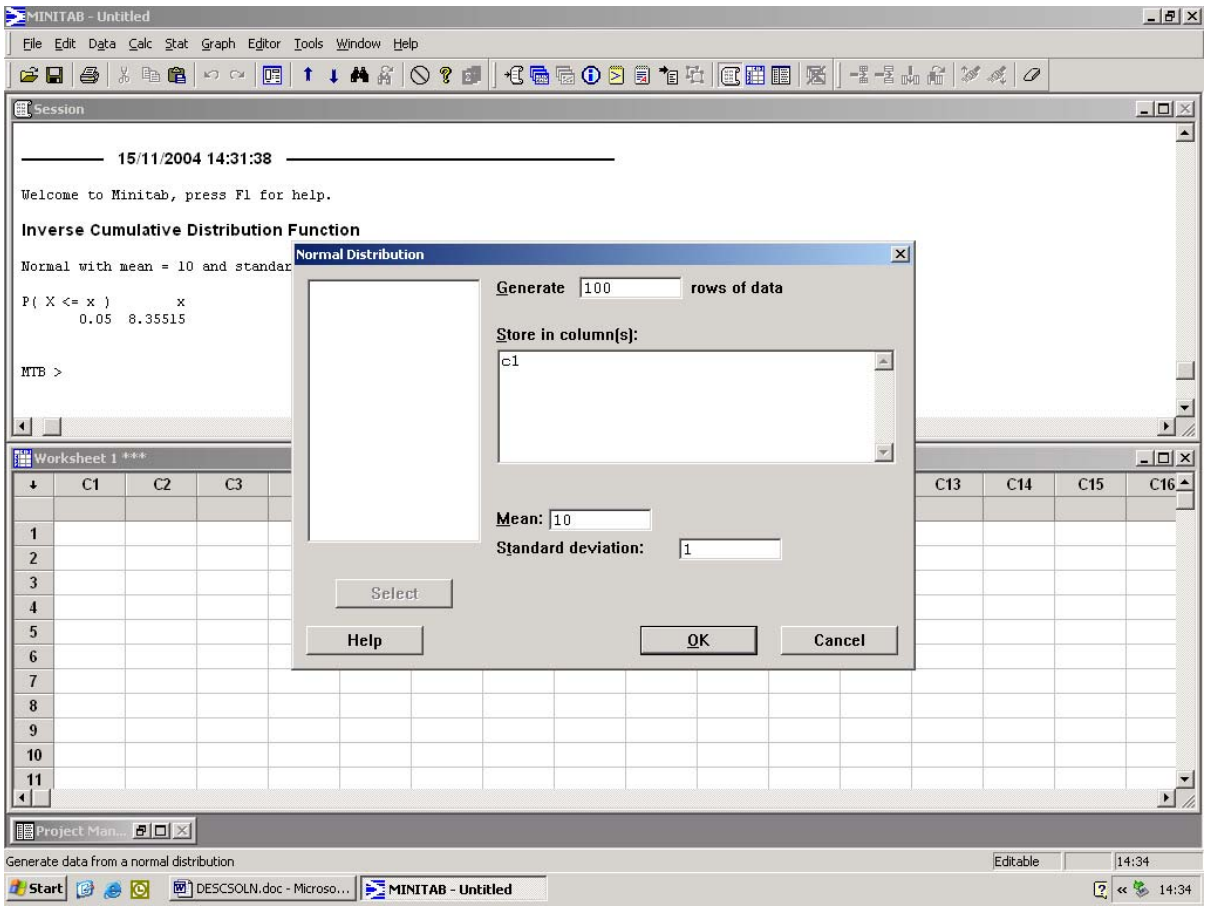
5.

Using the method of question 4, with an Input constant box of 0.05 gives a threshold of 8.35515.
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 μ=10, σ=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 in the sample I generated are, respectively, 10.009 and 0.990 (shown below) but your values will vary from these as everyone will have a different random sample from the population (but see the comment at the end of this answer).

The table shows that 4 of the 100 values are below 8.35515, so 4% rather than 5% fall below the 5% threshold. This is only to be expected - the number falling below 8.35515 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. The proportion below the 5% threshold in this sample will, in general, be different to that obtained from the first sample you drew. For my second sample it was 3.

Descriptive Statistics: C1

```
Variable    N  N*    Mean  SE Mean  StDev  Minimum     Q1  Median      Q3
C1        100   0  10.009   0.0990  0.990    6.327  9.346  10.082  10.672

Variable  Maximum
C1         12.057
```

```
MTB > XTABS C2;
SUBC>   Layout 1 0;
SUBC>   Counts;
SUBC>   DMissing C2.
```

Tabulated statistics: C2

```
Rows: C2

       Count

0         96
1          4
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:

Descriptive Statistics: C1

```
Variable      N  N*    Mean  SE Mean  StDev  Minimum     Q1  Median      Q3
C1        10000   0  10.004   0.0100  1.004    6.398  9.331   9.989  10.680

Variable  Maximum
C1         13.692
```

Note that the sample mean and standard deviation, 10.004 and 1.004, are closer to their population counterparts, 10 and 1, in this much larger sample.

Using exactly the same methods as before shows that 487 of the 10,000 values are below 8.3551, i.e. 4.87% 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, if before you generate any random numbers, you type the following command in the Session window:

MTB > Base 321

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 it sets the random number generator to a reproducible state.}