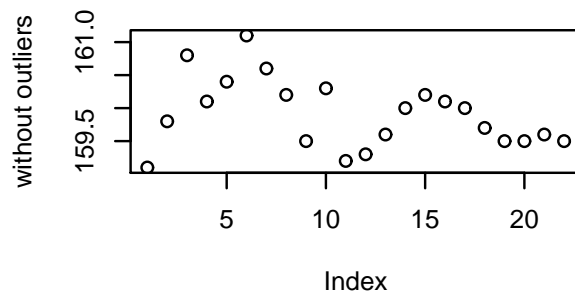
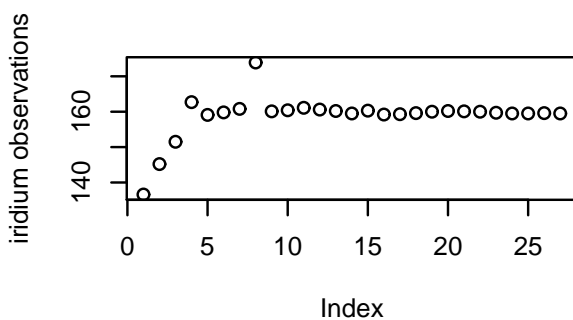
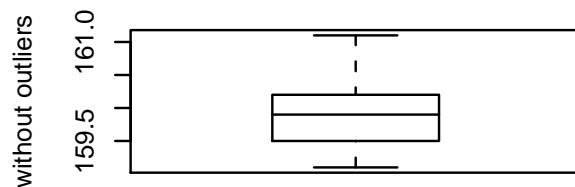
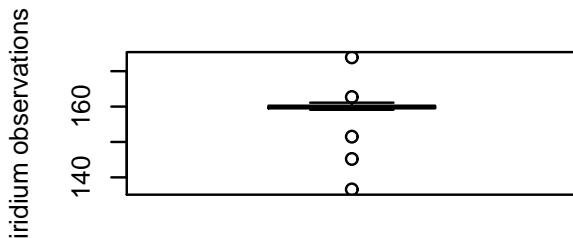
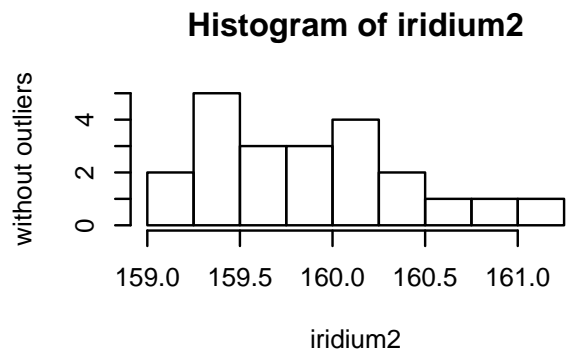
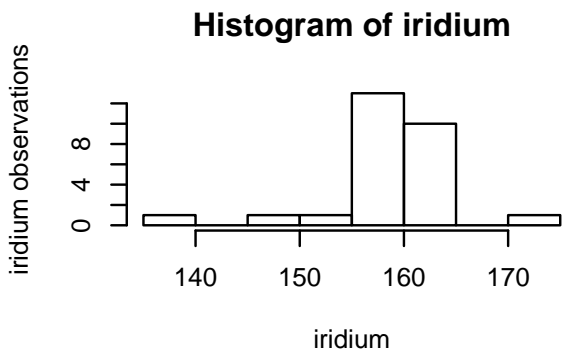


Solution Sheet 1

1. (a). The plots indicate that the values of five of the observations are strikingly different from the rest, and plotting the observations in the order taken identifies these as the first four observations and the eighth. In what follows I have included two sets of analyses – one with the full data set (data set *iridium*) and one, for comparison, with the five outliers omitted (referred to in the plot titles as *iridium2*).

R gives the following histogram, boxplot and plot in order taken (in each case the plot for the full data set is on the left). Note how the observations bunch in the full plots, and much easier it is to get an overview of the remainder of the observations once the outliers are removed.



R gives stem-and-leaf plots as follows (full set on the left):

```

13 | 7
14 |
14 | 5
15 | 2
15 | 999
16 | 00000000000000001113
16 |
17 | 4
159 | 123
159 | 55556678
160 | 00112234
160 | 68
161 | 1

```

(b,c,d).

From the plot, it appears that the first four observations are unrepresentative of typical values from the experiment, and are due to some initial aspect of the way the experiment was set up (equipment warming up, operator getting callibrated etc.). The eighth observation also appears atypical, but it is not clear why. If it is just a mis-recording, then there is no harm in omitting it, but if it is a genuine observation representing real but occasional variability, then it should really be left in. The two sets of analyses below are one with the full data set and one with the five outliers omitted.

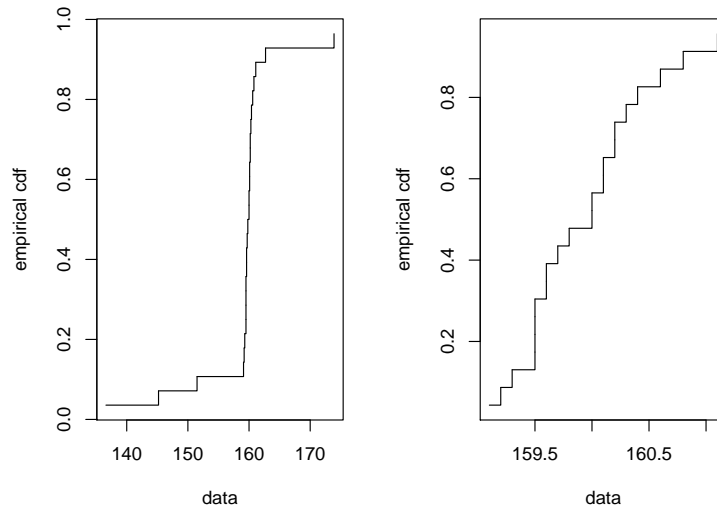
	Full set	Set with outliers removed
median	159.8	159.9
mean	158.81	159.91
10% trimmed mean	159.55	159.88
20% trimmed mean	159.84	159.86
variance	38.75	0.2755
standard deviation	6.225	0.5249
Lower hinge H_1	159.5	159.5
Upper hinge H_3	160.25	160.2
Lower quartile Q_1	159.5	159.5
Upper quartile Q_3	160.3	160.2
IQR	0.75	0.70

For the full data set, the effect of the outliers can be seen in the difference between the median (which is resistant to the effect of outliers), the mean (which is substantially affected by outliers), and the trimmed mean (where the effects are removed by trimming off values including the outliers). For the data set with the outliers removed, there is, as one would expect, very little difference between the median and the mean, and there is no need to bother computing the trimmed means.

The different ways the outliers affect the variance and the IQR is even more marked. The variance is about 100 times greater with the outliers than without then, and the standard deviation is thus about 10 times greater. However, the upper and lower hinges and quartiles, and hence the IQR, are only slightly affected by the outliers and give a consistent view.

One slight point is that observations appear to become slightly less variable over time, and there appears to be a slight hint of systematic sinusoidal variation towards the end of the experiment.

2. Below are plots of the empirical (or sample) cumulative distribution function for both the full data set and for the data set with the outliers omitted (see the solution to question 1 for a discussion of these data sets). The plot for the full data set is on the left. Again, it is much easier to get a feel for the remaining data once the outliers are removed.



3,4,5 – the solutions will appear on the 2nd solution sheet, next week.