Lab Session 1 Solutions

Here are solutions to the Lab Session 1 exercises. Also on Canvas you will find example .R files for each of the questions.

1. (a) Minimum=1.1
   Maximum=12.9
   Median=$\frac{5.6+9.2}{2} = 7.4$
   $Q_1 = 1.1 + 0.75(3.4 - 1.1) = 2.825$
   $Q_3 = 11.7 + (0.25(12.9 - 11.7) = 12.0$
   $\bar{x} = 7.317$
   s = 4.709

   (b) Minimum=10
   Maximum=61
   Median= 28
   $Q_1 = 16.5$
   $Q_3 = 48.5$
   $\bar{x} = 31.444$
   s = 17.960

2. Note that copying the data to the Jupyter Notebook led to an entry, 138, being left out of the initial data set for the mileage, so you would have seen 40 entries in the pdf version of the questions and 39 in the notebook version. On the left column are the answers for n=40, on the right n=39. Interesting to compare. Does it make much difference?

   n=40

   (i) Stem and leaf plot with number of leafs $\sqrt{n} \simeq 6$

11	9
12	5 6 8
13	2 5 5 5 6 8 8
14	0 0 2 2 4 4 5 5 6 6 7 7 8 9
15	0 0 2 3 4 6 7 8
16	1 3 4 5 8
17	3 6

   (ii) $\bar{x} = 146.8$
   Median = 146
   The mean and the median are very close suggesting the distribution is symmetric. This is reflected in the shape of the stem & leaf.

   (iii) Range = 57
   IQR = 155.5-138 = 17.5
   Quartile deviation = IQR/2 = 8.75
   s = 13.05

   n=39

   (i) Stem and leaf plot with number of leafs $\sqrt{n} \simeq 6$

11	9
12	5 6 8
13	2 5 5 5 6 8
14	0 0 2 2 4 4 5 5 6 6 7 7 8 9
15	0 0 2 3 4 6 7 8
16	1 3 4 5 8
17	3 6

   (ii) $\bar{x} = 147.03$
   Median = 146
   The mean and the median are very close suggesting the distribution is symmetric. This is reflected in the shape of the stem & leaf.

   (iii) Range = 57
   IQR = 156-138 = 18
   Quartile deviation = IQR/2 = 9
   s = 13.14

(iv) (a) All the values are greater than the previous maximum so mean and median will both increase.
$\bar{x} = 162.22$
Median = 149.5
Range = 171
Quartile Deviation = 14.625
s = 36.81

(b) The mean is now greater than the median, suggesting that the data is now skewed.

11	9
12	5 6 8
13	2 5 5 5 6 8 8
14	0 0 2 2 4 4 5 5 6 6 7 7 8 9
15	0 0 2 3 4 6 7 8
16	1 3 4 5 8
17	3 6 9
18	
19	8 8 9
20	4
21	0
22	
23	4
24	
25	
26	0 7
27	
28	
29	0

(iv) (a) All the values are greater than the previous maximum so mean and median will both increase.
$\bar{x} = 162.71$
Median = 149
Range = 171
Quartile Deviation = 14.75
s = 37.02

(b) The mean is now greater than the median, suggesting that the data is now skewed.

11	9
12	5 6 8
13	2 5 5 5 6 8
14	0 0 2 2 4 4 5 5 6 6 7 7 8 9
15	0 0 2 3 4 6 7 8
16	1 3 4 5 8
17	3 6 9
18	
19	8 8 9
20	4
21	0
22	
23	4
24	
25	
26	0 7
27	
28	
29	0

3. Again, a strange character must have toggled my insert key when copying into the notebook for this so there should have been an initial value 11.7 in this data. Something to look out for when copying between the different formats. With the extra value get: $\bar{x} = 11.43$ and s = 0.495; without the extra value get $\bar{x} = 11.4$ and $s = 0.52$

   The subjective impression is that the sample mean is less than 12.1 and the sample standard deviation is small suggesting a difference. Furthermore, there is only one data point greater than 21.1.

4. Group 1: $\bar{x} = 3.62 \quad s = 0.294$
   Group 2: $\bar{x} = 3.25 \quad s = 0.259$

   The subjective impression is that the sample means differ by approximately 0.37 and standard deviations are relatively small, suggesting a possible difference in the mean response.

   Did you consider how you might assess whether the data was normally distributed?

For each group did you try to compare the distribution of the data with a normal distribution with the same mean and standard deviation?

5. This is paired or matched data, father/son pairings from 8 families. So, the first step is to take differences between each pair and calculate the numerical summaries for the differences. So, the differences are:
   -2.5 7.1 1.1 1.8 1.4 -1.0 -2.5 3.2
   For the differences:
   $\bar{x} = 1.075 \quad s = 3.192$
   The subjective impression is that the mean of the differences is fairly close to zero with the standard deviation larger, suggesting that there may be no difference. We will return to this when we have more statistical tests to apply.

6. Don't worry if you didn't find anything. hist() provides all you need. You may have found the package ggplot2 (also comes in the tidyverse package that we installed last week). I've given a short example using geom_histogram() and geom_freqpoly(), but I barely scratch the surface. They are worth having a play with and you can get some really great plots using functions from this library.