Paulette Rodriguez

Prof. Loh

Applied Statistics

10 September 2019

Homework 1

 # 1.30

(a).

```
> KPOT40Data <- read.csv(file.choose(), header=T, sep=",")
> attach(KPOT40Data)
> names(KPOT40Data)
[1] "ID"          "Potassium_mg"
[3] "Dose"         "Source"
[5] "X"           "X.1"
[7] "X.2"          "X.3"
[9] "X.4"
> stem(Potassium_mg, scale=2)
```

The decimal point is 2 digit(s) to the right of the |:

```
  26 | 6
  27 | 9
  28 |
  29 | 5688
  30 | 3577
  31 | 02235
  32 | 336689
  33 |
  34 | 9
  35 | 148
  36 | 1
  37 |
  38 |
  39 |
  40 |
  41 |
  42 | 1
```

(b).

As we can conclude from this stem plot, most of the amounts of potassium absorbed by the participants in this study cluster around 2.9K-3.3K with the center being at 3,490 (34|9) and the range, or spread, being 1,549.11.

(c).
Yes, there is an outlier. The outlier is the maximum point 4,213.49, or as we can see from the stem plot created using the program R, (42|1) is the outlier. By computing the five-number summary, I was able to compute the following values.

Minimum = 2664.38
Q1 = 3027.64
Median = 3130.37
Q3 = 3286.95
Maximum = 4213.49

IQR = 259.31

In order to determine if any of the values from the dataset were outliers, I used the following statements: A data value is said to be an outlier if the data value is less than (<) Q1-1.5*(IQR) or greater than (>) Q3+1.5*(IQR). This gave me an answer of 2638.675 after plugging in the values for Q1 and IQR and an answer of 3675.915 after plugging in the values for Q3 and IQR. Hence, there is an outlier if a data value is <2638.675 or >3675.915. Therefore, I was able to conclude that there wasn't a number less than 2638.675, but there was a number greater than 3675.915 which turns out to be the maximum number of the data set, 4213.49 or 42|1.

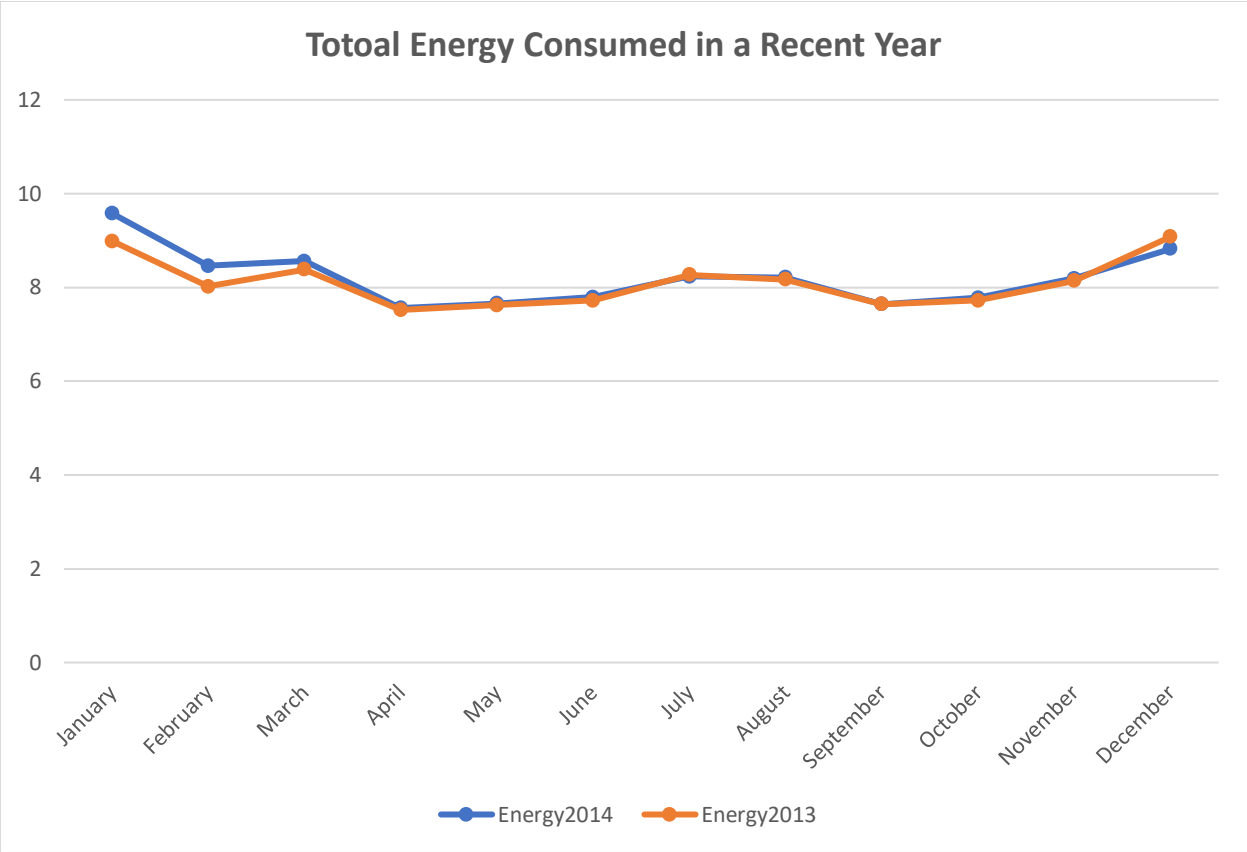
(***d).***
The shape of this stem plot would best be described as having a single peak and being slightly symmetrical. For the center of the distribution, the mean is 3208.437, the median is 3130.37 and there is no mode. Most of the amounts of potassium absorbed by the participants seen on the stem plot appear up to two times but no more than that.

 # 1.32

(a).
We can see that 2013 has the highest usage in December and January. By looking at the table we can describe the energy consumption varies from month to month. From January to June the energy levels are seen to be decreasing. For the next two months, July and August, the energy levels rise up again to around 8.2 quadrillion Btu but fall back down for the following two months, September and October, and then rise up again for the last two months November and December. The energy levels aren't falling or rising at high difference levels but by just seeing the information provided to us in the table we can still see that the energy levels are varying about every two months whether they are rising for two months or falling for two months and so on. Considering the data given we can also assume this variation is caused depending on the month and weather.

(b).

The file provided to make the time plot shows energy levels for years 2013 and 2014. As we can see from the plot, the amounts of energy for both years are relatively the same, but they vary in certain months. The patterns are very similar, but the values for the winter months in 2014 are higher than those in the 2013 winter months. This is most likely due to weather.



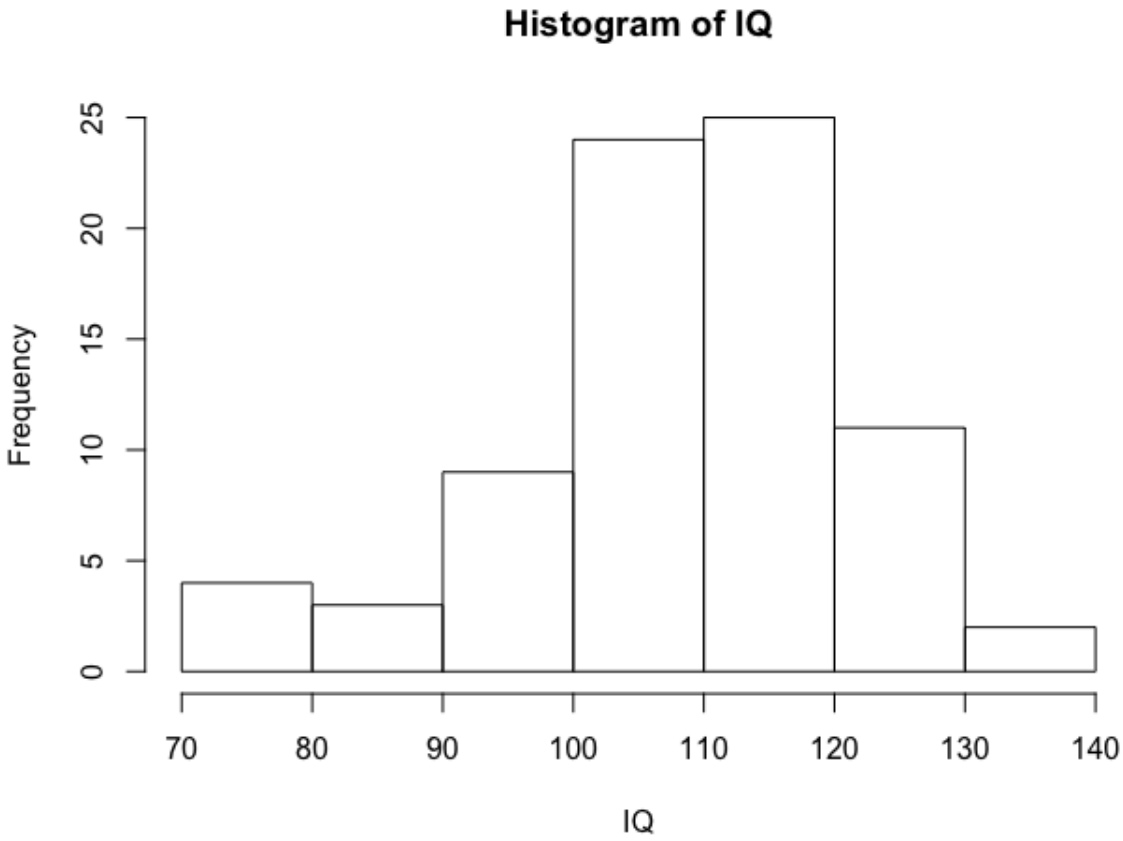
(c.)

A graph display for datasets of this kind are more effective for communicating information about month-to-month variation in energy consumptions. A visual presentation is often most commonly used and the most effective in showing what the data is trying say. For example, the time plot was way more effective than the table because we were able to see the variation in energy consumption for both years with the help of two different colors showing us how both years compared in their month-to-month consumption of energy. Although, it wasn't varying by much, the visuals were helpful in communicating the information that we needed to see. The graphs do a nice job of putting all values of the dataset into a nice picture that depicts the same amount of information seen in the table. When it comes to bigger datasets, graphs are the most effective because it is easier to read what is going on.

#1.40

```
> SevenGrDATA <- read.csv(file.choose(), header=T, sep=",")
> names(SevenGrDATA)
 [1] "ID"         "GPA"        "IQ"         "Gender"     "SelfConcept"
```

```
 [6] "X"         "X.1"       "X.2"       "X.3"       "X.4"
[11] "X.5"       "X.6"
> hist(IQ)
```



The shape of this distribution of IQ scores for the seventh-grade students is pretty normal, or we can say unimodal and skewed mostly to the left. It is centered near 110 and has a mean of about 95.5, the median is 110, and the mode is 103. It is spread from 80 to 140. As we can see from the histogram shown above, there are no outliers.
IQ scores are usually said to be centered at 100. The midpoint for these students is clearly above 100.

**** The same code/dataset will be used for the following questions: #1.58, 1.60, 1.62. The code will be split to answer its corresponding questions****

#1.58

```
> KSUPData <- read.csv(file.choose(), header=T, sep=",")
> names(KSUPData)
[1] "ID"          "Potassium_mg" "Dose"        "Source"
> mean(Potassium_mg)
[1] 3313.376
> median(Potassium_mg)
[1] 3245.62
```

(a).
The mean for these data is 3313.376 as we can see above which was computed using R program.

(b).
The median for these data is 3245.62 as we can see above which was computed in R.

(c.)
The measure that would be most preferred for describing the center of this distribution would be the median. This is because the distribution is right skewed with an outlier, hence, the median is a better measure of center. The median is the value that will most likely stay the same no matter if there are numbers changed around or not. The mean is harder to stay the same since any change in any of the values of the dataset would change the mean.

#1.60

(a).
```
> sd(Potassium_mg)
[1] 280.1192
```

By computing the standard deviation for this dataset using R programming, we got a value of 280.1192.

(b).
```
> quantile(Potassium_mg, 0.25)
   25%
3136.48
> quantile(Potassium_mg, 0.75)
   75%
3481.39
> quantile(Potassium_mg, 0.50)
   50%
3245.62
> IQR(Potassium_mg)
[1] 344.91
```

By computing the quantiles for this dataset using R programming, we got the following values:
Q1 = 3136.48
Q3= 3481.39

(c.)
```
> fivenum(Potassium_mg)
[1] 2680.07 3136.48 3245.62 3481.39 3946.31
```

The above shows the five-number summary of the dataset. The meaning of the five numbers is as follows:

- 2680.07 is the minimum value of the dataset,

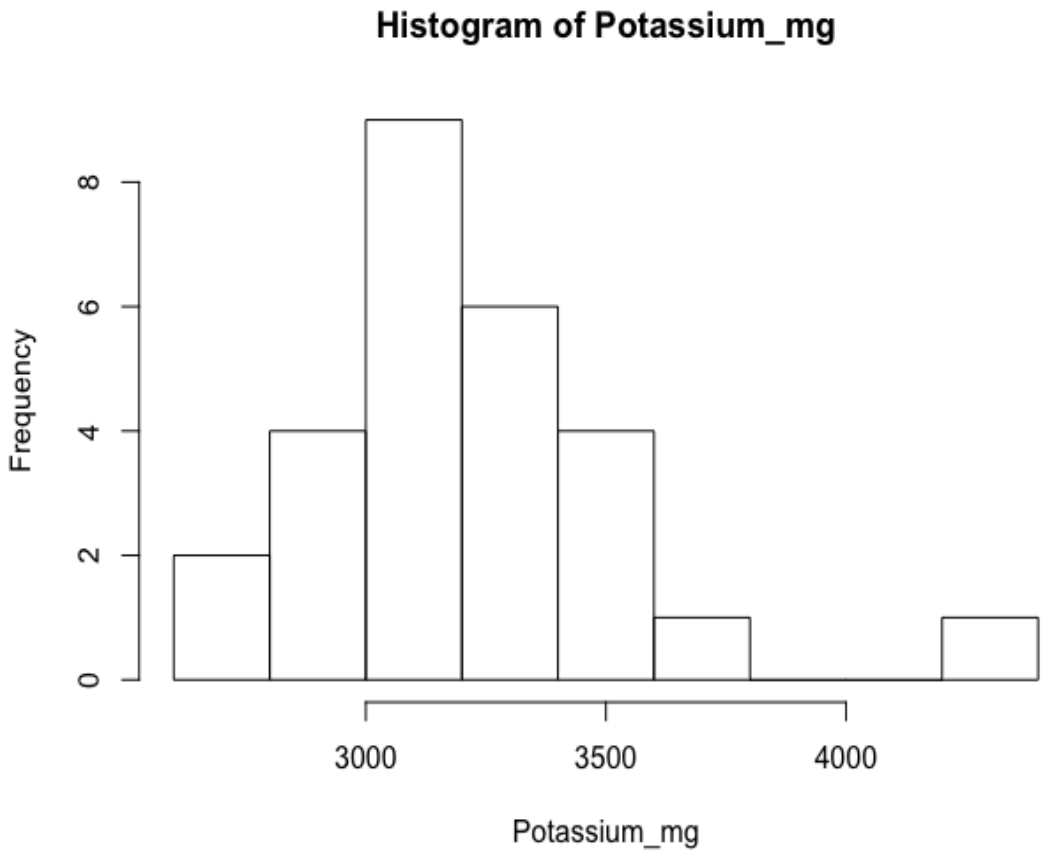
- 3136.48 is Q1 as we saw from the previous question. This number represents 25% of the data below it,
- 3245.62 is the median as we computed from the previous question. This can also be known as Q2 which represents 50% of the observations below or above it,
- 3481.39 is Q3 which presents 75% of the observations, and
- 3946.31 is the maximum value of the dataset.

(d).
The five-number summary would be better for this distribution because it is positively skewed where there might be an outlier.

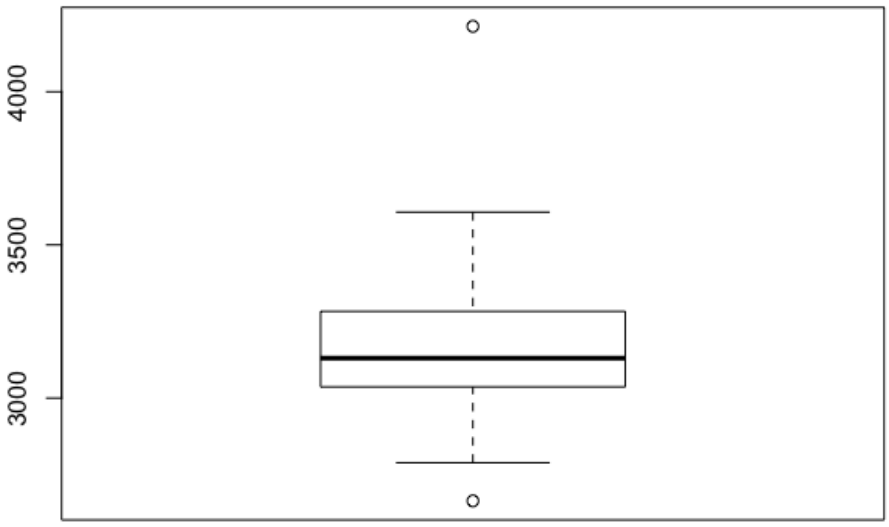
#1.62

(a)
> hist(Potassium_mg)

The distribution of potassium absorption is right skewed with a potential outlier. As we can see from the histogram, there is an outlier after all.



(b)
> boxplot(Potassium_mg)



The distribution of potassium absorption is left skewed as we can see from the boxplot above.

(c)
> stem(Potassium_mg)

The decimal point is 2 digit(s) to the right of the |

26 | 69
28 | 5688
30 | 357702235
32 | 336689
34 | 9148
36 | 1
38 |
40 |
42 | 1

The stem plot might have an advantage since it preserves the data. The boxplot also works but it seems to hide some of the details that the histogram shows. Therefore, the histogram is probably preferred over the stem and box plots.

#1.88

(a).
The five-number summary was found using the following code in R programming.

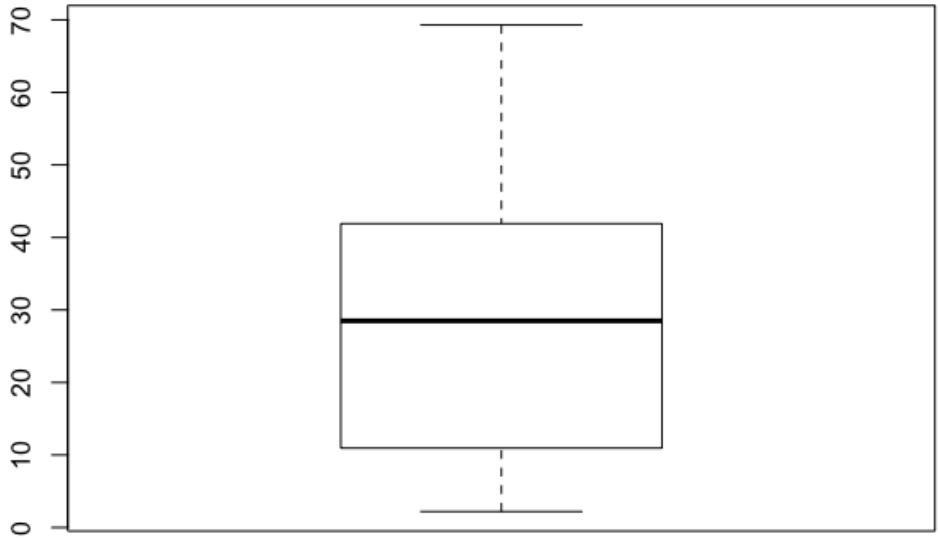
> PinesData <- read.csv(file.choose(), header=T, sep=",")
> attach(PinesData)
> names(PinesData)
[1] "Diameter"
> fivenum(Diameter)

[1] 2.20 10.95 28.50 41.90 69.30

To summarize what was found:

- Min.=2.20cm
- Q1= 10.95cm
- Median = 28.50cm
- Q3 = 41.90cm
- Max.=69.30cm

(b).



(c).