Homework 2

Miles Tweed

9/12/2020

Problem 1

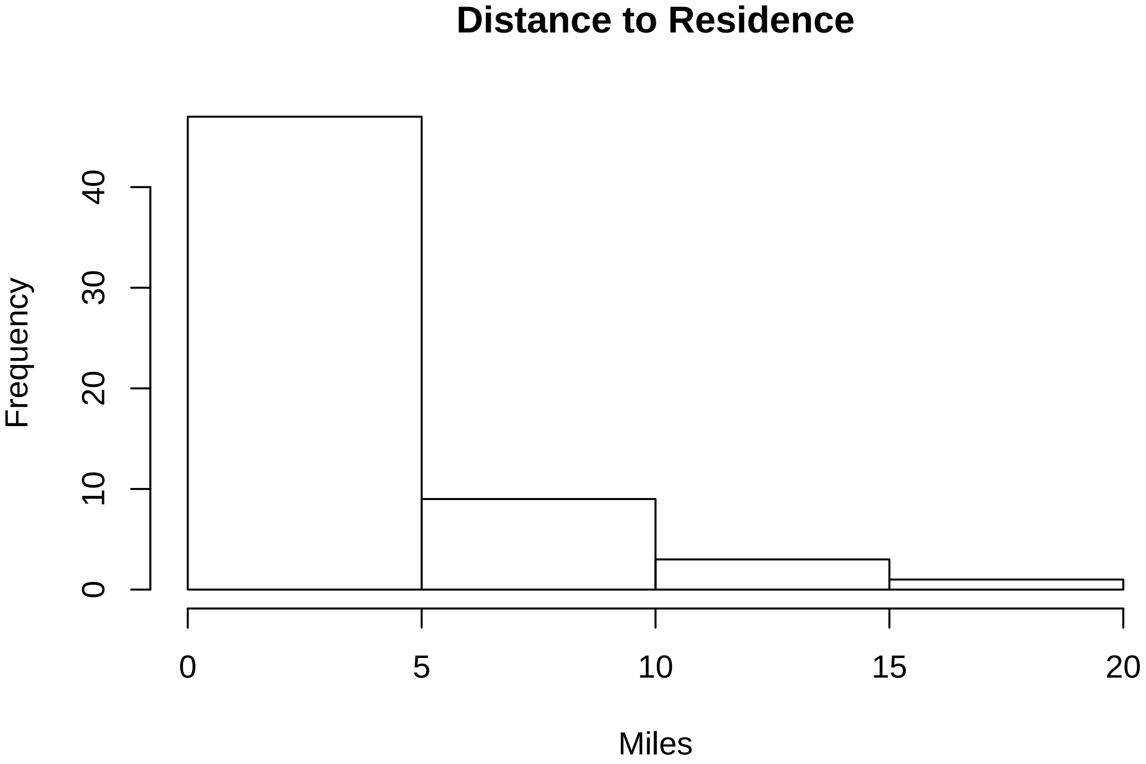
I will choose the quantitative variable distance_residence to analyse for this problem. For this analysis I am assuming that the distance to the student's residence is measured in miles.

```
df <- read.csv('http://sites.williams.edu/bklingen/files/2015/07/fl_student_survey.csv')
```

Part 1

Looking at the distribution of this data graphically, most of the data seems to be accumulated in the lower range with a few data points at the upper end of the range. This indicates a right skewed distribution.

```
hist(df$distance_residence, xlab = "Miles", main = "Distance to Residence")
```



I propose that using the median will be the most appropriate measure of the center of this distribution. This is because, unlike the mean, the median is not affected by outlier values. This means that the median is more representative of the center of the more common values.

```
median(df$distance_residence)
```

```
## [1] 2
```

Looking at the mean, it is obvious that it has been pulled towards the right tail end of the distribution. While, mathematically this represents the average of all numerical values it does not accurately portray the middle of the distribution of values.

```
mean(df$distance_residence)
```

```
## [1] 3.818333
```

Part 2

The first measure of variability that does not use measures of position is the range of the data. This measure looks at the total spread of the data from highest to lowest, in otherwords, the maximum value minus the minimum value.

```
dist_range <- max(df$distance_residence) - min(df$distance_residence)
dist_range
```

```
## [1] 19.8
```

This measure tells us that all of the values in the dataset occur, at most, within 19.8 units of each other (in this case we are assuming the units to be in miles). The range does not tell us more than how far apart the lowest and highest values are.

Another measure of variability is standard deviation (s). Standard deviation is calulated using the arithmetic mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ which is subracted from each observation and squared. The average of these values is called the variance and the square root of variance is the standard deviation and is in the same units as the observations. The formula for standard deviation is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

This measure of variability is much more robust than range because it say much more about the distribution of the values. A small standard deviation indicates that the observations occur more densly around the mean and the peak will be sharp. A large standard deviation inbicates that the observations are more spread throughout the distribution and the peak will be broad. The standard deviation is used to describe confidence intervals, for example, 65% of the observations will fall between one standard deviation on either side of the mean while 95% will fall between two standard devitions in a normally distributed sample. Additionally, standard deviation can be used to identify outlier values. If a values is beyond 3 standard deviations from the mean, it is likely an outlier. The standard deviation of the distance_residence data is:

```
dist_sd <- sd(df$distance_residence)
dist_sd
```

```
## [1] 4.117902
```

The 65% confidence interval of the data is:

```
paste(mean(df$distance_residence)-sd(df$distance_residence), 'to',
      mean(df$distance_residence)+sd(df$distance_residence))
```

```
## [1] "-0.299568537669549 to 7.93623520433622"
```

However, it is impossible to have negative distance so this range would actually start at 0. Similarly the 95% confidence interval would start at 0 and extend to:

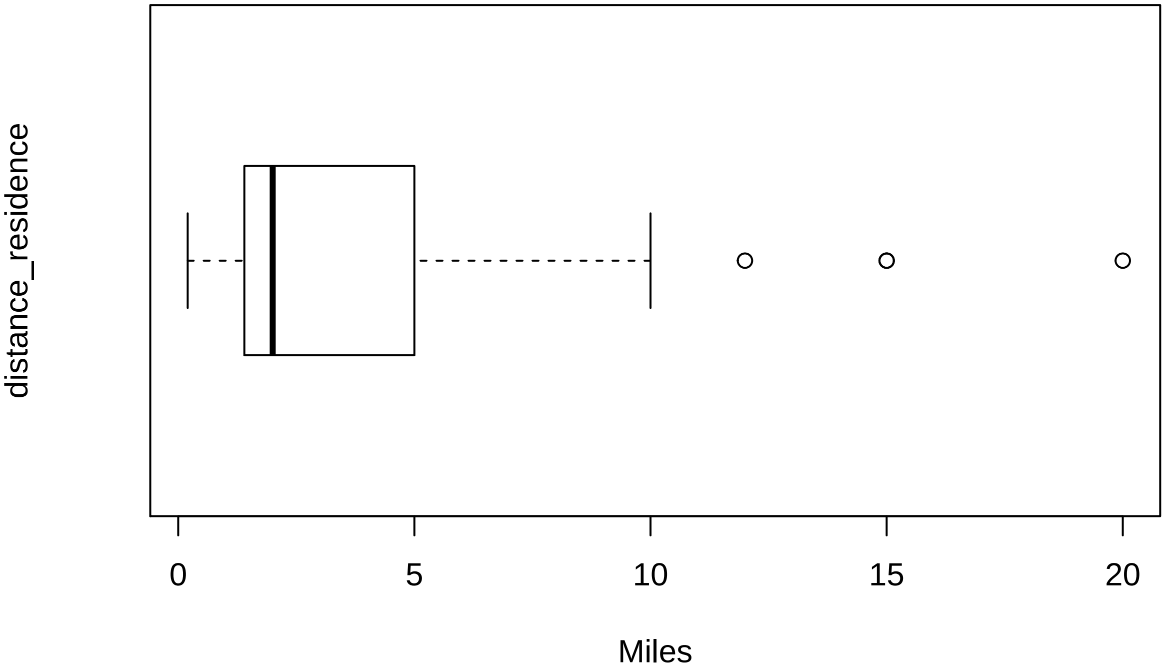
```
mean(df$distance_residence)+(2*sd(df$distance_residence))
```

```
## [1] 12.05414
```

Part 3

A box plot, as below, summarizes the data based on the five-number measures of position.

```
boxplot(df$distance_residence, horizontal = T,
        xlab = 'Miles', ylab ='distance_residence')
```



It is better to characterize this particular distribution using measures that are associated with an observation's position, since this data is right skewed and it was already determined that the median was a better measure for the center of the distribution.

The positional measures of a distribution are the quantiles and more specifically the quartiles. Like confidence intervals, these measures describe the amount of information that they contain. The first quartile contains 25% of the data, the second quartile is the median and it contains 50% of the observations. The third quartile contains 75% of the data and the first quartile, median, and third quartile are included in this summary (along with the minimum value, the mean, and the maximum value):

```
summary(df$distance_residence)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.200   1.450   2.000   3.818   5.000  20.000
```

The interquartile range or IQR is the range of values between the third quartile and the first quartile, which in the case of our data is $5 - 1.45 = 3.55$. This values can be used to identify any potential outliers using the 1.5*IQR rule. The wiskers on the above graph extend to the first values just within 1.5*IQR above the third quartile and below 1st quartile. Observation that extend beyond these whiskers are potential outliers Looking at the graph, there are three such values (17.0, 17,0, 20.0).

In general, these positional measures are advantageous to the standard deviation because they are not reactive to outliers but can still be used to represent the spread of the data as well as highlight any potential outliers.

Problem 2

2.21

a) The assessed value of houses in a large city should be right skewed because there would be a few homes with an extremely high assessed value wereas as the majority of homes would, likely, be grouped together in the lower range of value.

b) The number of time a faculty member overdraws their account would be a right skewed distribution. This is because it should be a rare occurance unless the faculty is severely underpaid. However, some people may be very bad with money and overdraw a lot.

c) The IQ for the general population is by definition a symmetric disribution since the mean score is the defined as 100 in order to make the distribution normal.

d) The heights of female college students may be normally distributed or right skewed. This is because the majority of student will fall into one range while there will be a smaller amount of taller and shorter individuals. The distribution os morelikely to be right skewed if the college has a womens basketball team.

2.34

The management would use the mean because it would, likely, be higher than the median. This is because the mean would react to the very high salary of the CEO and other executives and be pulled closer to the tail. The median, however, would not be affected by the inclusion of these outliers and would, therefore, be more representative of the average workers salary.

2.37

a)

```
travel <- c(0, 0, 4, 0, 0 , 0, 10, 0, 6, 0)
travel_mean <- sum(travel)/length(travel)
paste('The mean is', travel_mean)
```

```
## [1] "The mean is 2"
```

The mean indicates that on average the owner's employee's travel 2 miles on public transportation per day.

```
sort_travel <- sort(travel)
travel_median <- (sort_travel[5]+sort_travel[6]) /2
paste('The median value is:', travel_median)
```

```
## [1] "The median value is: 0"
```

The median indicates that the owner's employees have a tendancy to not use public transportation at all. this indicated that most of the contribution to the mean values are from a smaller group who use public transportation a lot. In other words, at least half of the owners employees do not use public transportation.

The table below is a count of the possible values. This shows that the mode or most common value is 0 with 7 instances.

```
table(travel)
```

```
## travel
##  0  4  6 10
##  7  1  1  1
```

b)

```
travel2 <- append(travel, 90)
travel2_mean <- sum(travel2)/length(travel2)
paste('The new mean is:', travel2_mean)
```

```
## [1] "The new mean is: 10"
```

The addition of this new value greatly increases the mean values which now indicated that the owners employees now travel and average of 10 miles per day on public transportation.

```r
sort_travel2 <- sort(travel2)
travel2_median <- (sort_travel2[6])
paste('The new median is:', travel2_median)
```

```
## [1] "The new median is: 0"
```

However, the median still indicates that it is more common for (or at least half of) her employees do not use public transportation at all.

Additionally, the table below shows that the modal value is still 0 with a count of 7 instances.

```r
table(travel2)
```

```
## travel2
##  0  4  6 10 90
##  7  1  1  1  1
```

2.46

a)

```r
sick <- c(0,0,4,0,0,0,6,0)
sick.range = max(sick)-min(sick)
paste('The range of the data is', sick.range)
```

```
## [1] "The range of the data is 6"
```

The range indicates that there is a 6 day difference between someone who takes the leaset amount of sick days and someone who takes the most.

b)

```r
n = length(sick)
sick.mean = sum(sick)/n
val_dif = sick-sick.mean
sick_sd <- sqrt(sum(val_dif^2)/(n-1))
paste('The standard deviation is:', sick_sd)
```

```
## [1] "The standard deviation is: 2.37546987833084"
```

The standard deviation indicates the spread of the data on either side of the mean. This measure can be used to generate a confidence interval, for example, we could be 65% confident that someone would take a number of sick days that correspond to the range of one standard deviation below to one standard deviation above the mean.

c)

```r
sick2 <- c(0,0,4,0,0,0,60,0)
sick2.range = max(sick2)-min(sick2)
paste('The range of the data is', sick2.range)
```

```
## [1] "The range of the data is 60"
```

The range of the data has been increased 10 fold.

```r
n = length(sick2)
sick2.mean = sum(sick2)/n
val_dif2 = sick2-sick2.mean
sick2_sd <- sqrt(sum(val_dif2^2)/(n-1))
paste('The standard deviation is:', sick2_sd)
```

```
## [1] "The standard deviation is: 21.0577437402152"
```

The standard deviation as well increased by a lot. This is because standard deviation, and similarly mean, are highly affected by outliers in the data. The mean gets moved towards the tail on which the outlier resides. The variance, which is the sum of the squared difference between the values and the mean, is used to calulate the standard deviation and increases rapidly.

2.54

a) The 95% confidence interval is represented by two standard deviation above and below the mean.

```r
paste('95% of the weights will fall between',133 - (2*17), 'and', 133 + (2*17))
```

```
## [1] "95% of the weights will fall between 99 and 167"
```

b)

```r
paste('An athlete whose weight is 3 sd above the mean would weigh', 133 + (3*17)
      ,'lbs.')
```

```
## [1] "An athlete whose weight is 3 sd above the mean would weigh 184 lbs."
```

This values would be considered an outlier for the female college atheletes data. This is because 3 standard deviations from the mean is generally considered a metric for identifying outliers. Additionally, if this data is right skewed the median is more representative of the common weight for the atheletes. The mean will get pulled towards the tail and the discrepancy between this value and other more common values will be greater.

2.62

```r
NGT <- list(c("Italy", "France", "Germany", "Brazil", "Britain", "Canada", "Japan", "US"),
            c(42, 37, 35, 34, 28, 26, 25, 13))
NGTdf <- data.frame(NGT)
names(NGTdf) <- c("countries","vac_days")

# barplot(height = NGTdf$vac_days, names.arg = NGTdf$countries,
#         col = c('lightblue', 'lightgreen'), space = 1.5,
#         ylab = 'Vacation Days', xlab = 'Countries',
#         main = "Average Vacation Days Per Country")
```

a)

```r
#length(NGTdf$vac_days)
sorted.vac.days <- sort(NGTdf$vac_days)
vac.median <- (sorted.vac.days[4] + sorted.vac.days[5])/2
paste('The median values is:', vac.median)
```

```
## [1] "The median values is: 31"
```

b)

```r
vac.quart1 <- (sorted.vac.days[2] + sorted.vac.days[3])/2
paste('The first quartile is:',vac.quart1)
```

```
## [1] "The first quartile is: 25.5"
```

c)

```r
vac.quart3 <- (sorted.vac.days[6] + sorted.vac.days[7])/2
paste('The third quartile is:',vac.quart3)
```

```
## [1] "The third quartile is: 36"
```

d)

- The median value indicates that, in this sample, exactly half of the observation have less than 31 sick days while the other half have more than 31 sick days.
- The first quartile value indicates that, in this sample, 25% of the observation have less than 25.5 sick days while the other 75% have more than 25.5 sick days.
- The Third quartile value indicates that, in this sample, 75% of the observation have less than 36 sick days while the other half have more than 36 sick days.

2.66

a) The standard deviation provides a lot more information about the distribution of values than the range does. The range only describes the distance between the minimum value and the maximum value. The standard deviation can be used to describe the density of the values around the mean as well as to help identify any outliers.

b) IQR is sometimes prefferable to standard deviation because is is less affected by outliers. This is because s is calulated using the mean while IQR is calculated aroud the median. Additionally, IQR can be used to detect potential outliers using the 1.5 * IQR rule which creates whiskers that extend to the first value that occurs within 1.5*IQR above the third quartile and below the first quartile.

c) Standard defiation is prefferable to IQR because it can be used to describe the spread of the distribution and the shape of the peaks, whereas IQR doesn't contain this infomation. If the data is normally distributed, the standard deviation is a more mathematiacally rigourous measure. Additionally, the standard deviation is used in calculating confidence intervals which are important tools of statistical analysis.

2.76

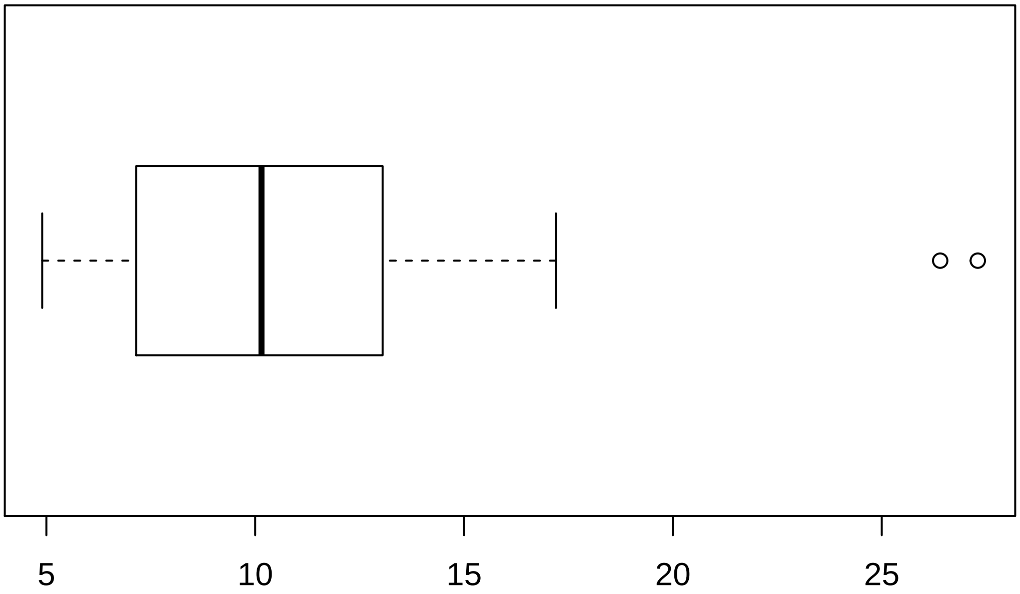
```
unemp.data <- list(c("Belgium", "Bulgaria", "Czech.Rep", "Denmark", "Germany", "Estonia", "Ireland", "G
                    "Spain", "France" ,"Croatia", "Italy", "Cyprus", "Lativia", "Lithuania", "Luxembourg
                    "Hungary", "Malta", "Netherlands", "Austria", "Poland", "Portugal", "Romania", "Slov
                    "Slovikia", "Finland", "Sweden", "UK"),
                    c(8.4, 13.0, 7.0, 7.0, 5.3, 8.6, 13.1, 27.3, 26.4, 10.3, 17.2, 12.2, 15.9, 11.9, 11.
                      5.8, 10.2, 6.5, 6.7, 4.9, 10.3, 16.5, 7.3, 10.1, 14.2, 8.2, 8.0, 7.5))
unemp.df <- data.frame(unemp.data)
names(unemp.df) <- c("country", "unemployment_rate")
head(unemp.df)
```

```
##     country unemployment_rate
## 1   Belgium               8.4
## 2  Bulgaria              13.0
## 3 Czech.Rep               7.0
## 4   Denmark               7.0
## 5   Germany               5.3
## 6   Estonia               8.6
```

```
summary(unemp.df)
```

```
##       country   unemployment_rate
##  Austria  : 1   Min.   : 4.900
##  Belgium  : 1   1st Qu.: 7.225
##  Bulgaria : 1   Median :10.150
##  Croatia  : 1   Mean   :11.129
##  Cyprus   : 1   3rd Qu.:13.025
##  Czech.Rep: 1   Max.   :27.300
##  (Other)  :22
```

```
boxplot(unemp.df$unemployment_rate, horizontal = T)
```



a)

- The lower edge of the box represents the value 7.225. This will include the values for Czech Republic(7.0), Denmark (7.0), Romania (7.3), UK (7.5) and Sweden (8.0) if we consider values between 7 and 8 to be on (or near) this edge. The upper edge of the box represents the value 13.025. This will include the value for Italy (12.2), Bulgaria (13.0), and Ireland (13.1) if we consider values between 12 and 13.1 to be on or near this edge. No values exactly coincide with the edges of these quartiles.
- The whiskers extend to 1.5*IQR above the third quartile and below the first quartile which would represent the value 1.45 at the lower limit and the value 18.85 at the upper limit. The values to which these wiskers extend corrospond to Austia (4.9) at the lower limit, and Croatia (17.2) at the upper limit

b) There are no outlier at the lower range of unemployment but Spain (26.4) and Greece (27.3) are the outliers for the upper range. This is because they extend beyond the threshold of 1.5*IQR (8.7) above the third quratile value (13.025), in other words, their unemplyment rate is greater that .

c) The Greek unemployment rate would not be an outlier according to the 3 standard deviation criteria. Three standard deviations above the mean corrosponds to the value of 27.88 and Greece's unemployment rate is 27.3 which is still with in this range.

```
unemp.sd <- sd(unemp.df$unemployment_rate)
unemp.mean <- mean(unemp.df$unemployment_rate)
paste("Three standard deviations above the mean is:",unemp.mean + (3 * unemp.sd))
```

```
## [1] "Three standard deviations above the mean is: 27.8800603460492"
```

Problem 3

Function Definition

I decided to write a function 'sortvect' that will sort a vector of numeric values passed in as an argument. By default, the function sorts the vector in ascending order unless the ascending argument is passed in as FALSE.

```
sortvect <- function(vect, ascending = T){
  if (length(vect) == 1) {
    return(vect)
  }
  if (ascending) {
    for (i in 1:(length(vect)-1)) {
      for (j in length(vect):(i+1)) {
        if (j > 1 & vect[j] < vect[(j-1)]) {
          vect[c(j-1,j)] <- vect[c(j,j-1)]
        }
      }
    }
  } else {
    for (i in 1:(length(vect)-1)) {
      for (j in length(vect):(i+1)) {
        if (j > 1 & vect[j] > vect[(j-1)]) {
          vect[c(j-1,j)] <- vect[c(j,j-1)]
        }
      }
    }
  }
  return(vect)
}
```

Testing

I will first define some test vectors of varying length. Some will be in ascending order, some will be in decending order, and some will be in random order.

```
vec1_desc <- c(10,9,8,7,6,5,4,3,2,1)
vec1_asc <- c(1,2,3,4,5,6,7,8,9,10)
vec1_rand <- c(4,3,7,10,8,6,2,5,9,1)
vec2 <- c(1,2,3,4,10,9,8,7,6)
vec3 <- c(3,6,1,3)
vec4 <- c(64,200,6,3,8,123,66,83,99)
vec.one <- c(7)
```

Next, I will test the function with the test vectors.

```
sortvect(vec1_desc)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
sortvect(vec1_asc)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
sortvect(vec1_rand)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
sortvect(vec2)
```

```
## [1]  1  2  3  4  6  7  8  9 10
```

```
sortvect(vec3)
```

```
## [1] 1 3 3 6
```

```
sortvect(vec4)
```

```
## [1]   3   6   8  64  66  83  99 123 200
```

```r
sortvect(vec.one)
```

```
## [1] 7
```

Finally, I will change the default of ascending to FALSE in order to get a vector in descending order.

```r
vec4 <- c(64,200,6,3,8,123,66,83,99)
sortvect(vec4, ascending = FALSE)
```

```
## [1] 200 123  99  83  66  64   8   6   3
```