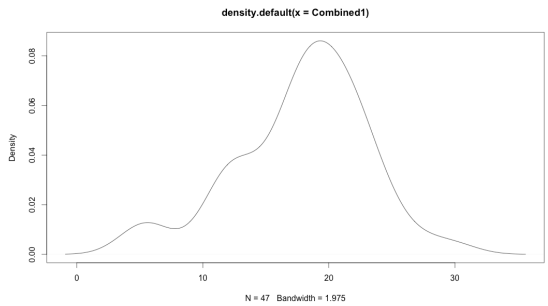
ADA Homework 1

Wenxin Liang UNI:wl2455

Consider the Motivation and Creativity Data in Ramsey &Schafer, Chapter 1, and appended below.

i) Determine whether the mean or the median would be an appropriate measure of location for the combined data. Use graphical methods to support your arguments.

First we using the density plot to observe



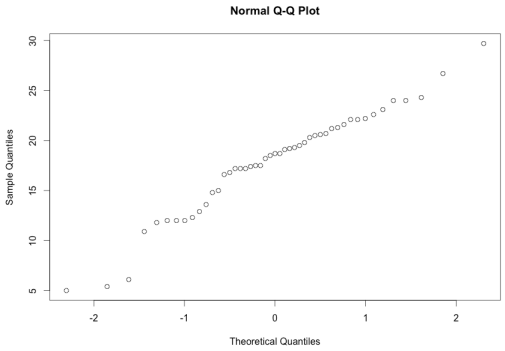

```
> stem(Combined1)

  The decimal point is at the |

   4 | 04
   6 | 1
   8 |
  10 | 98
  12 | 000396
  14 | 80
  16 | 68222455
  18 | 257712358
  20 | 3567236
  22 | 11261
  24 | 003
  26 | 7
  28 | 7
```

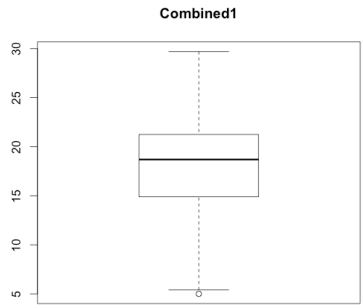
The density graph showed the distribution that the combined data formed kind of like skewness to left. Also from the stem-and-leaf diagram we can observe the skewness a little bit.

We can use the qqplot to make sure,



Since the qq plot we have is not so much close to a line so there is some skewness existing.

We can use the box plot to explain more as followed,



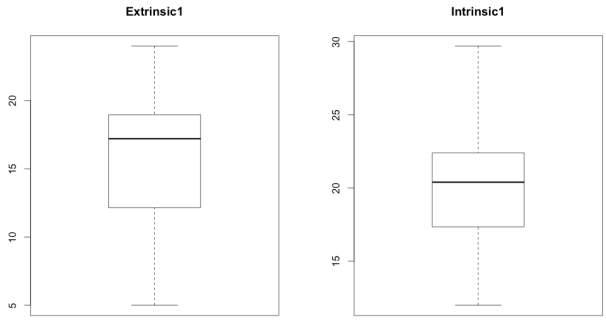
Based on the box plot for combined data we know that there is one extreme point which less than 1.5 box lengths from the box.

Since the density graph and the qqplot graph show the skewness and there is an extreme point existing then the median would be an appropriate measure of location for the combined data.

ii) What would be a reasonable measure of dispersion for the data in each group? Use graphical methods to support your arguments.

There are several measures can measure of dispersion such as Variance, Standard deviation, Interquartile range (IQR), Range or Median Absolute Deviation.

Since the outlier will have effective on the variance, which is $\frac{\sum(X_i-\bar{X})^2}{n}$ and the range, which is max-min. Instead, the standard deviation, which is the square root of the variance, decreases the influence of outlier. The IQR, which is Q3-Q1 does not affect by the outlier and the median absolute deviation, which is a robust measure of the variability of a univariate sample of quantitative data since robust then the influence of outlier will be not counted.



We can use the box plot and the IQR to explain the procedure of finding whether there is an outlier as followed,

Based on the R code to measure the dispersion for the data in extrinsic group, if the observations belong to [1.95,29.15] then we conclude the observations are within the main body of data.

Based on the R code to measure the dispersion for the data in intrinsic group, if the observations belong to [10.1125,29.6125] then we conclude the observations are within the main body of data.

```
> (quan_Extrin <- quantile(Extrinstic1))
   0%   25%   50%   75%  100%
 5.00 12.15 17.20 18.95 24.00
> Q1_Extrin <- 12.15
> Q3_Extrin <- 18.95
> IQR_Extrin <- IQR(Extrinstic1)
> low_Extrin <- Q1_Extrin - 1.5*IQR_Extrin
> (low_Extrin <- Q1_Extrin - 1.5*IQR_Extrin)
[1] 1.95
> (high_Extrin <- Q3_Extrin + 1.5*IQR_Extrin)
[1] 29.15
> (quan_Intrin <- quantile(Intrinstic1))
    0%    25%    50%    75%   100%
12.000 17.425 20.400 22.300 29.700
> Q1_Intrin <- 17.425
> Q3_Intrin <- 22.300
> IQR_Intrin <- IQR(Intrinstic1)
> (low_Intrin <- Q1_Intrin - 1.5*IQR_Intrin)
[1] 10.1125
> (high_Intrin <- Q3_Intrin + 1.5*IQR_Intrin)
[1] 29.6125
```

```
> max(Extrinstic1)
[1] 24
> min(Extrinstic1)
[1] 5
> max(Intrinstic1)
[1] 29.7
> min(Intrinstic1)
[1] 12
```

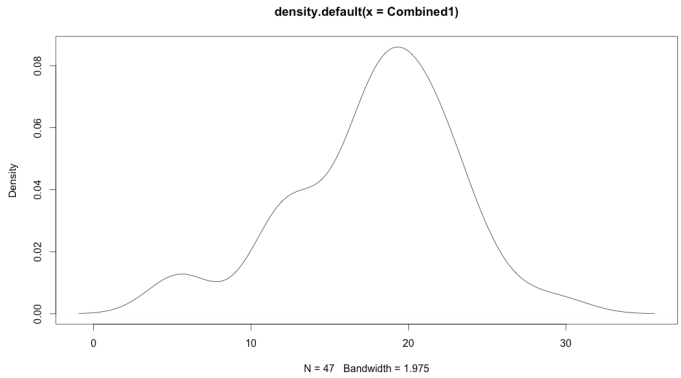
Based on R code, we know that the maximum of intrinsic group is 29.7>29.6125, the upper bound of the interval for the intrinsic group [10.1125,29.6125]. Therefore, there is an extreme point existing in the intrinsic group. Then the reasonable measure of dispersion for the data in intrinsic group is the MAD, the standard deviation and IQR.

Based on R since both the maximum value and the minimum value of extrinsic group are within the interval for the extrinsic group [1.95,29.15], there is no outlier existing. Then the reasonable measure of dispersion for the data in extrinsic group is all the measurements we list before. Generally we choose standard deviation since basically it is the square root of sample variance.

iii) Determine whether the combined data is unimodal, and, if so, estimate the bias and variance of the mode of the distribution for the combined data using each of the following methods:

- Jackknife
- Bootstrap

Based on the density graph and R code, we conclude that the combined data is not unimodal.




```
> (names(table(Combined1)[table(Combined1) == max(table(Combined1))]))
[1] "12"   "17.2"
```

The jackknife is a resampling technique especially useful for variance and bias estimation. Systematically leaving out each observation from a dataset and calculating the estimate and then finding the average of these calculations find the jackknife estimator of a parameter. Let $\widehat{\theta_{(J)}}$ be an estimator computed based on all but X_j,i.e. leaving out the jth observation. Then the jackknife estimator of bias is given by $B_{JACK} = (n-1)\left[\frac{\sum_j^n \hat{\theta}_{(J)}}{n} - \hat{\theta}\right]$ The bias reduced jackknife estimator is given by $\hat{\theta}_{JACK} = \hat{\theta} - B_{JACK}$ and the variance $V_{JACK} = \frac{n-1}{n}\sum_j(\widehat{\theta_{(J)}} - \frac{\sum_j \widehat{\theta_{(J)}}}{n})^2$

The bootstrap is the practice of estimating properties of an estimator (such as its variance by measuring those properties when sampling from an approximating distribution. Let $X_1, \dots, X_n$ be a random sample from F_θ. Suppose an estimator of ϑ is $\widehat{\theta_n}$ when θ is the median, the sample median is approximately $N(\theta, \frac{1}{4nf^2(\theta)})$. For each sample, compute a statistic of interest, say $\widehat{\theta_n^+}$, assess the variability of $\widehat{\theta_n}$ about θ by that of $\widehat{\theta_n^+}$ about $\widehat{\theta_n}$. Estimate the bias $\widehat{\theta_n} - \theta$ by the mean of $\widehat{\theta_n^+} - \hat{\theta}$. Estimate the distribution of $\hat{\theta}$ by the e.d.f. of $\widehat{\theta_n^+}$.

If we really want to do Jack knife and Bootstrap, method 1 to solve the problem of the two modes we average them to show that we know how to use jackknife and bootstrap method. With Jackknife, the variance is 39.69702 and the bias is 0. With bootstrap, the variance is 14.96551 and the bias is 2.455934.

```
> install.packages("bootstrap")
> library("bootstrap")
> beta1 <- function(x){mean(as.numeric(names(table(as.vector(x)))[table(as.v
ector(x))==max(table(as.vector(x)))]))}
> jackknife(Combined1,beta1)
$jack.se
[1] 6.300557

$jack.bias
[1] 0

$jack.values
 [1] 14.6 14.6 14.6 14.6 14.6 17.2 14.6 14.6 14.6 14.6 12.0 12.0 14.6 14.6 14.6
14.6 14.6 14.6 14.6 14.6
[21] 14.6 14.6 14.6 17.2 17.2 14.6 14.6 14.6 12.0 14.6 14.6 14.6 14.6 14.6 14.6
14.6 14.6 14.6 14.6 14.6
[41] 14.6 14.6 14.6 14.6 14.6 14.6 14.6

$call
jackknife(x = Combined1, theta = beta1)

> mode_avg <- jackknife(Combined1,beta1)$jack.values
> sub <- numeric(n)
> for (i in 1:n)
+ {sub[i] <- (mode_avg[i]-sum(mode_avg)/n)^2}
> sub
 [1] 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30
6.760000e+00 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30
[11] 6.760000e+00 6.760000e+00 3.155444e-30 3.155444e-30 3.155444e-30
3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30
[21] 3.155444e-30 3.155444e-30 3.155444e-30 6.760000e+00 6.760000e+00
3.155444e-30 3.155444e-30 3.155444e-30 6.760000e+00 3.155444e-30
[31] 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30
3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30
[41] 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30 3.155444e-30
3.155444e-30 3.155444e-30
> var <- sum(sub)*(n-1)/n
> var
[1] 39.69702
```

```
> jack_knife2 <- jackknife(Combined1,beta1)
> bt_results <- bootstrap(Combined1, 500,beta)
> Boot_mode    <- mean(bt_results$thetastar)
> (Boot_var      <- var(bt_results$thetastar))
[1] 14.96551
>(Boot_bias=Boot_mode-mean(as.numeric(names(table(Combined1))[table(Co
mbined1)==max(table(Combined1))])))
[1] 2.455934
```

Method 2 we can use the maximum mode to do it. With Jackknife, the variance is 74.32634 and the jackknife bias is -15.26809. With bootstrap, the variance is 21.4095 and the bias is 1.4268.

```
> library("bootstrap")
> beta1 <- function(x){max(as.numeric(names(table(as.vector(x)))[table(as.ve
ctor(x))==max(table(as.vector(x)))]))}
> jackknife(Combined1,beta)
$jack.se
[1] 8.621272

$jack.bias
[1] -15.26809

$jack.values
 [1] 17.2 17.2 17.2 17.2 17.2 17.2 17.2 17.2 17.2 17.2 12.0 12.0 17.2 17.2 17.2
17.2 17.2 17.2 17.2 17.2 17.2
[21] 17.2 17.2 17.2 17.2 17.2 17.2 17.2 17.2 12.0 17.2 17.2 17.2 17.2 17.2 17.2
17.2 17.2 17.2 17.2 17.2 17.2
[41] 17.2 17.2 17.2 17.2 17.2 17.2 17.2

$call
jackknife(x = Combined1, theta = beta)

>mode_jack <- jackknife(Combined1,beta)$jack.values
> sub <- numeric(n)
> for (i in 1:n)
+ {sub[i] <- (mode_jack[i]-sum(mode_jack)/n)^2}
> sub
 [1]      0.1101675      0.1101675      0.1101675      0.1101675      0.1101675
0.1101675      0.1101675      0.1101675      0.1101675      0.1101675  23.6982526
23.6982526
[13]      0.1101675      0.1101675      0.1101675      0.1101675      0.1101675
0.1101675      0.1101675      0.1101675      0.1101675      0.1101675      0.1101675
0.1101675
[25]      0.1101675      0.1101675      0.1101675      0.1101675  23.6982526
```

```
0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675
[37] 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675 0.1101675
> var <- sum(sub)*(n-1)/n
> var
[1] 74.32634
> bt_results <- bootstrap(Combined1, 500,beta)
> Boot_mode    <- mean(bt_results$thetastar)
> (Boot_var      <- var(bt_results$thetastar))
[1] 21.4095
>  (Boot_bias=Boot_mode-max(as.numeric(names(table(Combined1))[table(Combined1)==max(table(Combined1))])))
[1]  1.4268
```