HW3

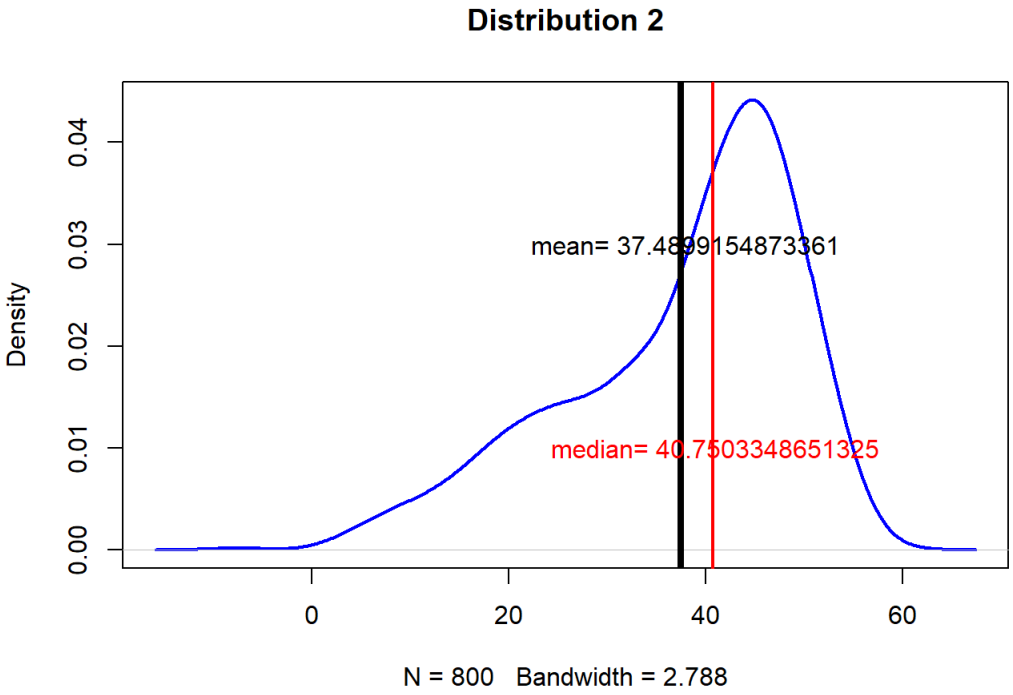
My student number: 106070020, student that help me: 106070038

2021/03/13

Question 1

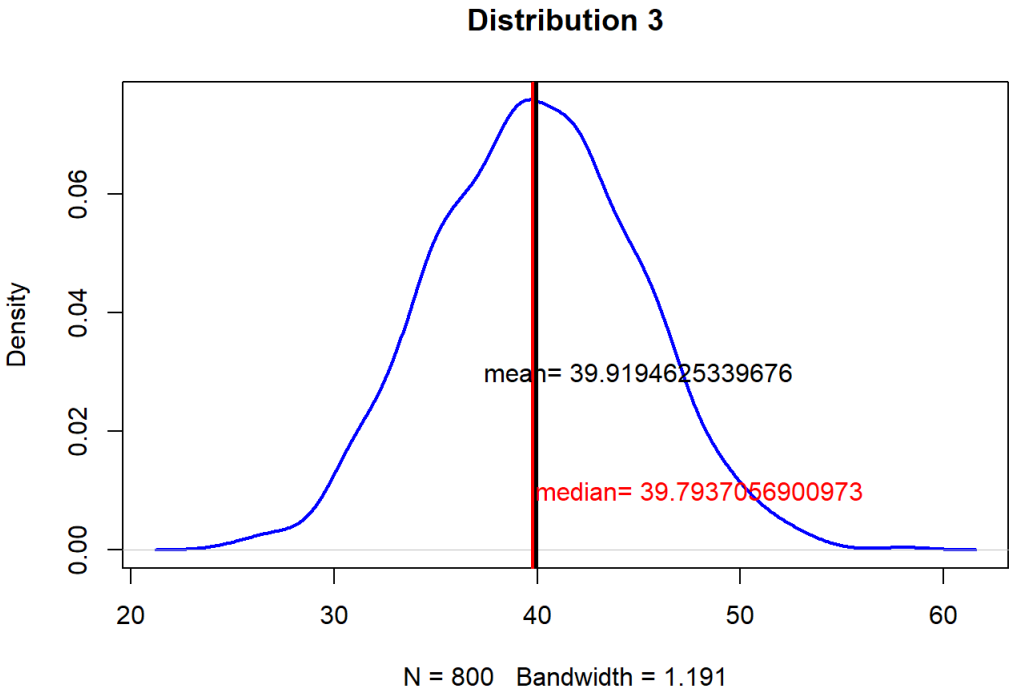
(a) Create and visualize a new “Distribution 2”: a combined dataset (n=800) that is negatively skewed (tail stretches to the left). Change the mean and standard deviation of d1, d2, and d3 to achieve this new distribution. Compute the mean and median, and draw lines showing the mean (thick line) and median (thin line).

```
d1 <- rnorm(n=500, mean=45, sd=5)
d2 <- rnorm(n=200, mean=30, sd=6)
d3 <- rnorm(n=100, mean=15, sd=7)

# Let’s combine them into a single dataset
d123 <- c(d1, d2, d3)
a<-paste("mean=",mean(d123))
b<-paste("median=",median(d123))
# Let’s plot the density function of abc
{plot(density(d123), col="blue", lwd=2,
      main = "Distribution 2")

  # Add vertical lines showing mean and median
  abline(v=mean(d123),lwd = 4)
  abline(v=median(d123), lwd = 2, col="red")
  text(x=38, y=0.03,a)
  text(x=41, y=0.01,b, col="red")}
```



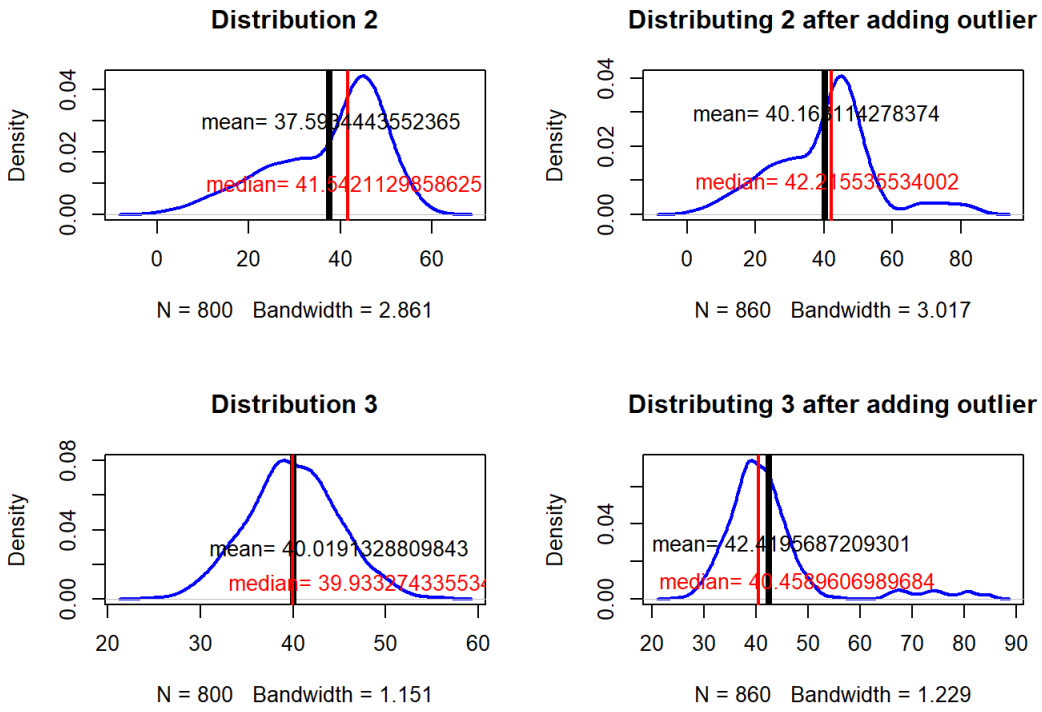
(b) Create a “Distribution 3”: a single dataset that is normally distributed (bell-shaped, symmetric) -- you do not need to combine datasets, just use the rnorm function to create a single large dataset (n=800). Show your code, compute the mean and median, and draw lines showing the mean (thick line) and median (thin line).

```
d1 <- rnorm(n=800, mean=40, sd=5)
a<-paste("mean=",mean(d1))
b<-paste("median=",median(d1))
# Let's plot the density function of abc
{plot(density(d1), col="blue", lwd=2,
      main = "Distribution 3")

  # Add vertical lines showing mean and median
  abline(v=mean(d1),lwd=4)
  abline(v=median(d1),lwd=2, col="red")
  text(x=45, y=0.03,a)
  text(x=48, y=0.01,b, col="red")}
```



(c) In general, which measure of central tendency (mean or median) do you think will be more sensitive (will change more) to outliers being added to your data?

```
op=par(mfrow=c(2,2))
d1 <- rnorm(n=500, mean=45, sd=5)
d2 <- rnorm(n=200, mean=30, sd=6)
d3 <- rnorm(n=100, mean=15, sd=7)

# Let’s combine them into a single dataset
d123 <- c(d1, d2, d3)
a<-paste("mean=",mean(d123))
b<-paste("median=",median(d123))
# Let’s plot the density function of abc
{plot(density(d123), col="blue", lwd=2,
      main = "Distribution 2")

  # Add vertical lines showing mean and median
  abline(v=mean(d123),lwd=4)
  abline(v=median(d123),lwd=2, col="red")
  text(x=38, y=0.03,a)
  text(x=41, y=0.01,b, col="red")}

d4 <- runif(60, min=65, max=85)
# Let’s combine them into a single dataset
d123 <- c(d1, d2, d3, d4)
a<-paste("mean=",mean(d123))
b<-paste("median=",median(d123))
# Let’s plot the density function of abc
{plot(density(d123), col="blue", lwd=2,
      main = "Distributing 2 after adding outlier")

  # Add vertical lines showing mean and median
  abline(v=mean(d123),lwd=4)
  abline(v=median(d123),lwd=2, col="red")
  text(x=38, y=0.03,a)
  text(x=41, y=0.01,b, col="red")}

d1 <- rnorm(n=800, mean=40, sd=5)
a<-paste("mean=",mean(d1))
b<-paste("median=",median(d1))
# Let’s plot the density function of abc
{plot(density(d1), col="blue", lwd=2,
      main = "Distribution 3")

  # Add vertical lines showing mean and median
  abline(v=mean(d1),lwd=4)
  abline(v=median(d1),lwd=2, col="red")
  text(x=45, y=0.03,a)
  text(x=48, y=0.01,b, col="red")}


# Let’s combine them into a single dataset
d14 <- c(d1, d4)
a<-paste("mean=",mean(d14))
b<-paste("median=",median(d14))
# Let’s plot the density function of abc
{plot(density(d14), col="blue", lwd=2,
      main = "Distributing 3 after adding outlier")

  # Add vertical lines showing mean and median
  abline(v=mean(d14),lwd=4)
  abline(v=median(d14),lwd=2,col="red")
  text(x=45, y=0.03,a)
  text(x=48, y=0.01,b, col="red")}
```



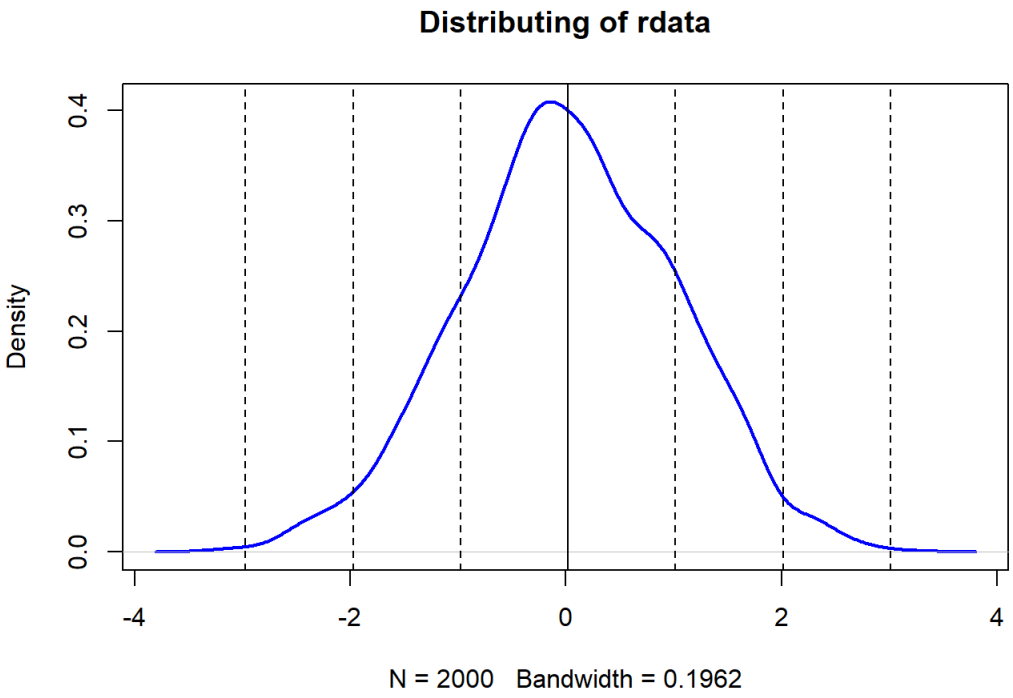

```
par(op)
```

Conclusion: It is obvious that the mean will be more sensitive than the median when the outliers are added.

Question 2

(a) Create a random dataset (call it 'rdata') that is normally distributed with: n=2000, mean=0, sd=1. Draw a density plot and put a solid vertical line on the mean, and dashed vertical lines at the 1st, 2nd, and 3rd standard deviations to the left and right of the mean. You should have a total of 7 vertical lines (one solid, six dashed).

```
rdata <- rnorm(n=2000, mean=0, sd=1)
{plot(density(rdata), col="blue", lwd=2,
      main = "Distributing of rdata")

  # Add vertical lines showing mean and median
  abline(v=mean(rdata))
  abline(v=mean(rdata)+1*sd(rdata),lty="dashed")
  abline(v=mean(rdata)+2*sd(rdata),lty="dashed")
  abline(v=mean(rdata)+3*sd(rdata),lty="dashed")
  abline(v=mean(rdata)-1*sd(rdata),lty="dashed")
  abline(v=mean(rdata)-2*sd(rdata),lty="dashed")
  abline(v=mean(rdata)-3*sd(rdata),lty="dashed")
}
```



(b) Using the quantile() function, which data points correspond to the 1st, 2nd, and 3rd quartiles (i.e., 25th, 50th, 75th percentiles)? How many standard deviations away from the mean (divide by standard-deviation; keep positive or negative sign) are those points corresponding to the 1st, 2nd, and 3rd quartiles?

```
rdata <- rnorm(n=2000, mean=0, sd=1)
q1<-quantile(rdata, prob=.25)
sdq1<- (q1-mean(rdata))/sd(rdata)
q2<-quantile(rdata, prob=.5)
sdq2<-(q2-mean(rdata))/sd(rdata)
q3<-quantile(rdata, prob=.75)
sdq3<-(q3--mean(rdata))/sd(rdata)
quantile1<-c(q1,q2,q3)
std1<-c(sdq1,sdq2,sdq3)
quantile1
```

```
##         25%         50%         75%
## -0.66153147  0.01583093  0.68317996
```

```
std1
```

```
##          25%          50%          75%
## -0.680485149  0.000342852  0.702244402
```

```
quantile1-std1
```

```
##         25%         50%         75%
##  0.01895368  0.01548808 -0.01906444
```

Conclusion: The points of 1st, 2nd, and 3rd quartiles become sligtly different after minusing the mean and dividing by standard-deviation.

(c)Now create a new random dataset that is normally distributed with: n=2000, mean=35, sd=3.5. In this distribution, how many standard deviations away from the mean (use positive or negative) are those points corresponding to the 1st and 3rd quartiles? Compare your answer to (b).

```
rdata2 <- rnorm(n=2000, mean=35, sd=3.5)
q21<-quantile(rdata2, prob=.25)
sdq21<- (q21-mean(rdata2))/sd(rdata2)
q22<-quantile(rdata2, prob=.5)
sdq22<-(q22-mean(rdata2))/sd(rdata2)
q23<-quantile(rdata2, prob=.75)
sdq23<-(q23-mean(rdata2))/sd(rdata2)
quantile2<-c(q21,q22,q23)
std2<-c(sdq21,sdq22,sdq23)
quantile2
```

```
##      25%      50%      75%
## 32.77250 34.97884 37.47388
```

```
std2
```

```
##         25%         50%         75%
## -0.66145549 -0.02782291  0.68871739
```

```
quantile1-std1
```

```
##         25%         50%         75%
##  0.01895368  0.01548808 -0.01906444
```

```
quantile2-std2
```

```
##      25%      50%      75%
## 33.43395 35.00667 36.78516
```

Conclusion: Compare to the results of (b), the points after minusing the mean and dividing by standard-deviation are very different to those of 1st and 3rd quartiles.

(d) Finally, recall the dataset d123 shown in the description of question 1. In that distribution, how many standard deviations away from the mean (use positive or negative) are those data points corresponding to the 1st and 3rd quartiles? Compare your answer to (b).

```
d1 <- rnorm(n=500, mean=15, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)
d123 <- c(d1, d2, d3)
q31<-quantile(d123, prob=.25)
sdq31<- (q31-mean(d123))/sd(d123)
q32<-quantile(d123, prob=.5)
sdq32<-(q32-mean(d123))/sd(d123)
q33<-quantile(d123, prob=.75)
sdq33<-(q33-mean(d123))/sd(d123)
quantile3<-c(q31,q32,q33)
std3<-c(sdq31,sdq32,sdq33)
quantile3
```

```
##      25%      50%      75%
## 13.92972 19.31844 30.33141
```

```
std3
```

```
##        25%        50%        75%
## -0.7429626 -0.2832861  0.6561582
```

```
quantile1-std1
```

```
##         25%         50%         75%
##  0.01895368  0.01548808 -0.01906444
```

```
quantile3-std3
```

```
##      25%      50%      75%
## 14.67268 19.60173 29.67525
```

Conclusion: Compare to the results of (b), the points after minusing the mean and dividing by standard-deviation are very different to those of 1st and 3rd quartiles.

Qusetion 3

(a)From the question on the forum, which formula does Rob Hyndman's answer (1st answer) suggest to use for bin widths/number? Also, what does the Wikipedia article say is the benefit of that formula?

Answer: Rob Hyndman suggests to use Freedman-Diaconis method. The bin-width is set to h=2×IQR×n−1/3. So the number of bins is (max−min)/h, where n is the number of observations, max is the maximum value and min is the minimum value. The benefit of this method are minimizing the difference between the area under the empirical probability distribution and the area under the theoretical probability distribution.

(b)

Given a random normal distribution: rand_data <- rnorm(800, mean=20, sd = 5)

Compute the bin widths (h) and number of bins (k) according to each of the following formula:

i. Sturges' formula

ii. Scott’s normal reference rule (uses standard deviation)

iii. Freedman-Diaconis’ choice (uses IQR)

```
rand_data <- rnorm(800, mean=20, sd = 5)
#(i)
k1 <- log2(800)+1
h1 <- (max(rand_data)-min(rand_data))/k1
#(ii)
h2 <- 3.49*5/800^(1/3)
k2 <- (max(rand_data)-min(rand_data))/h2
#(iii)
h3 <- 2*IQR(rand_data)/800^(1/3)
k3 <- (max(rand_data)-min(rand_data))/h3
bin_widths <- c(h1, h2, h3)
bins <- c(k1 ,k2 ,k3)
bin_widths
```

```
## [1] 2.974152 1.879744 1.424492
```

```
bins
```

```
## [1] 10.64386 16.84083 22.22297
```

(c) Repeat part (b) but extend the rand_data dataset with some outliers (use a new dataset out_data):

out_data <- c(rand_data, runif(10, min=40, max=60))

```
out_data <- c(rand_data, runif(10, min=40, max=60))
x <- length(out_data)
#(i)
k1 <- log2(x)+1
h1 <- (max(out_data)-min(out_data))/k1
#(ii)
h2 <- 3.49*sd(out_data)/x^(1/3)
k2 <- (max(out_data)-min(out_data))/h2
#(iii)
h3 <- 2*IQR(out_data)/x^(1/3)
k3 <- (max(out_data)-min(out_data))/h3
bin_widths <- c(h1, h2, h3)
bins <- c(k1 ,k2 ,k3)
bin_widths
```

```
## [1] 5.027999 2.195518 1.445359
```

```
bins
```

```
## [1] 10.66178 24.41675 37.08933
```

(d) From your answers above, in which of the three methods does the bin width (h) change the least when outliers are added (i.e., which is least sensitive to outliers), and (briefly) WHY do you think that is?

Answer: Freedman-Diaconis’ choice is least sensitive to outliers, as it is based on the interquartile range(a lind of robust statistics), which is less affected by the outliers.