HW2_107023058

Student ID: 107023058

Question 1

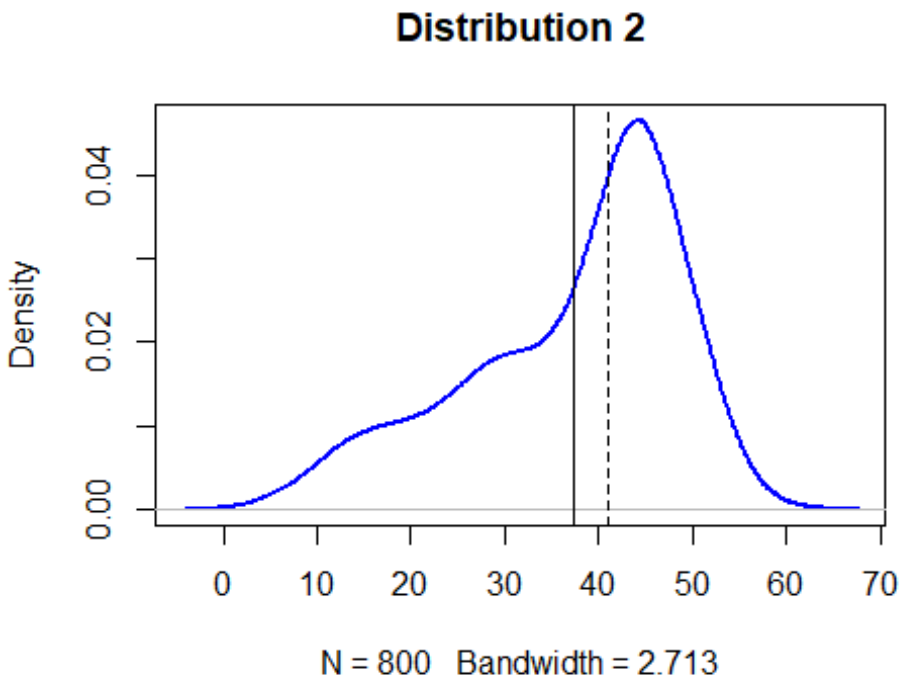
(a).

- Description:
  Create and visualize a new “Distribution 2” that is negatively skewed.

```
d1 <- rnorm(n=500, mean=45, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=15, sd=5)
d123 <- c(d1, d2, d3) # combine them into a single list

plot(density(d123), col="blue", lwd=2, main = "Distribution 2") # plot
the density function

# Add vertical lines showing mean and median
abline(v=median(d123),lty = "dashed")
abline(v=mean(d123))
```

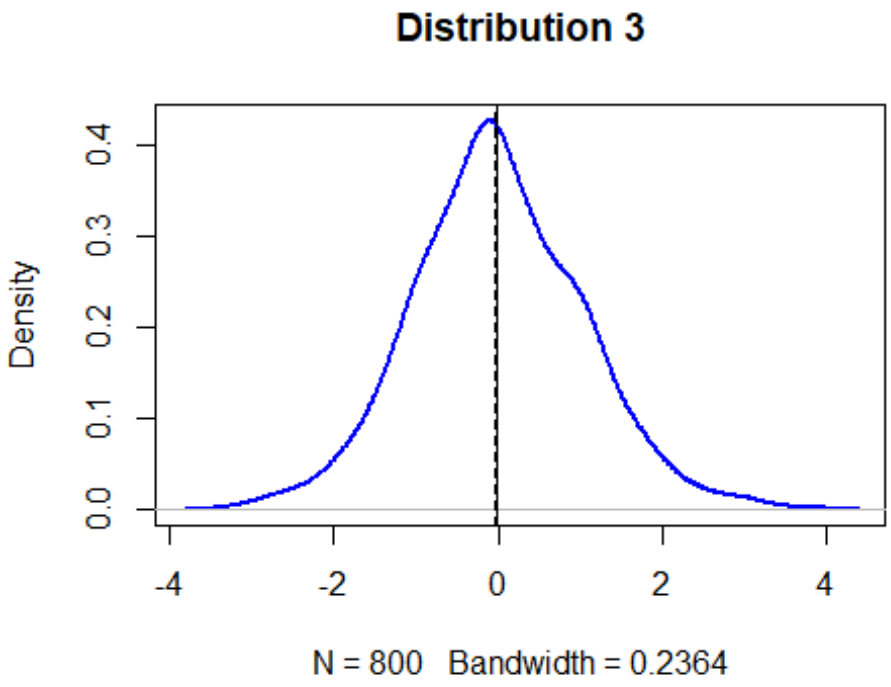


(b)

- Description:

  Create a “Distribution 3” that is normally distributed, compute and draw the mean and median.

```
d4 <- rnorm(n=800)
plot(density(d4),col = "blue",lwd = 2,main = "Distribution 3")
abline(v = median(d4),lty = "dashed")
abline(v = mean(d4))
```



(c)

- Description:

  which measure of central tendency will be more sensitive?

- Ans:

  By the part above, because that the mean move more greatly than median do, I think that mean is more sensitive to the outliers.

question 2

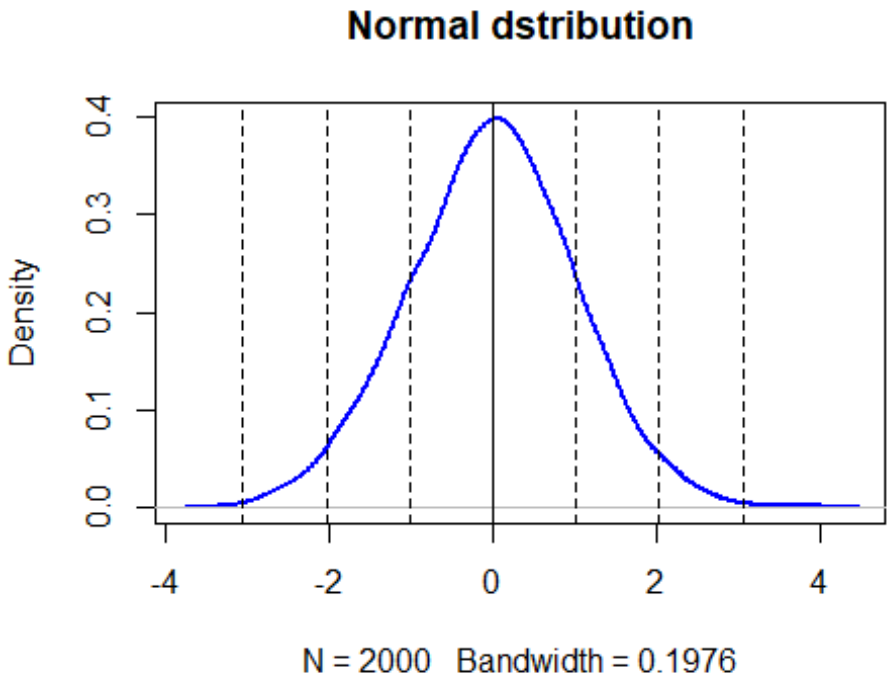
(a)

- Description:
  Create a random normal distribution and plot it with lines on mean and median.

```
rdata <- rnorm(2000,mean = 0,sd = 1)
plot(density(rdata),col= "blue",lwd = 2,main = "Normal distribution")
abline(v = mean(rdata),lty = "solid")

coeff <- c(-3,-2,-1,1,2,3)
for(i in 1:6)
  abline(v = mean(rdata)-coeff[i]*sd(rdata),lty = "dashed")

# draw the line on the point which is 1,2,3 s.d from mean.
```



(b)

- Description:
  With the data set in (a), how many sd away from the mean to the 1st~3rd quantiles?

```
qt <- quantile(rdata) # find the quantile of rdata
data_b <- c()
qt[2:4]

##          25%          50%          75%
## -0.674380862  0.007915793  0.671142262

for(i in 2:4)
  data_b[i-1] <- qt[i] # extract the value of 1st ~ 3rd quantile to the
 data_b list
  data_b[i-1] <- (data_b[i-1] - mean(rdata))/sd(rdata) # calculate how
many sd away from mean
data_b

## [1] -0.674380862  0.007915793  0.665820570
```

(c)

- Description:
  With a new data set,how many sd away from the mean to the 1st and 3rd quantiles?

```
random_dataset <- rnorm(2000,mean = 35 , sd = 3.5)
qt_c <- quantile(random_dataset)
data_c <- c()
data_c[1] <- qt_c[2];data_c[2] <- qt_c[4] # extract the value of 1st an
d 3rd quantile to the data_c list
for(i in 1:2)
  data_c[i] <- (data_c[i]-mean(random_dataset))/sd(random_dataset) # ca
lculate how many sd away from mean
data_c

## [1] -0.6566036  0.6728801
```

- Compare to (b), we can find that the distance from 25% quantile to mean is same with 75% quantile.

(d)

- Description:
  With the data set “d123”,how many sd away from the mean to the 1st and 3rd quantiles?

```r
qt_d <- quantile(d123)
data_d <- c()
data_d[1] <- qt_d[2];data_d[2] <- qt_d[4] # extract the value of 1st an
d 3rd quantile to the data_d list
for(i in 1:2)
  data_d[i] <- (data_d[i]-mean(d123))/sd(d123) # calculate how many sd
away from mean
data_d
```

```
## [1] -0.6558264  0.7343363
```

- Compare to (b), we can find that the distance from 25% quantile to mean is not same with 75% quantile. I think it is because that d123 is added with some outliers.

question 3

(a)

- Description:

  What is the formula and its benefit?

- Ans:

  He suggests to use the Freedman-Diaconis's choice,which is "h = 2 * IQR(x) / (n^(1/3))". The benefit of the formula is that it's less sensitive than the standard deviation to outliers in data

(b)

- Description:
  Compute the k and h in three ways.

```r
rand_data <- rnorm(800,mean = 20,sd = 5)
#(i) k = ceiling(log2n)+1
sturges_k <- ceiling(log2(800))+1
h1 <- (max(rand_data)-min(rand_data))/sturges_k
h1
```

```
## [1] 2.855481
```

```r
#(ii) k = Scott's normal  reference rule
h2 <- (3.49*sd(rand_data))/(800^(1/3))
h2
```

```
## [1] 1.830913
```

```
#(iii) Freedman-Diaconis' choice
h3 <- 2*IQR(rand_data)/(800^(1/3))
h3

## [1] 1.423996
```

(c)

- Description:
  Use a new dataset with an outlier, and repeat (b).

```
#(c)
out_data <- c(rand_data, runif(10,min = 40,max = 60))
#(i) k = ceiling(log2n)+1
sturges_k <- ceiling(log2(800))+1
h1 <- (max(out_data)-min(out_data))/sturges_k
h1

## [1] 4.81835

#(ii) k = Scott's normal  reference rule
h2 <- (3.49*sd(out_data))/(800^(1/3))
h2

## [1] 2.2825

#(iii) Freedman-Diaconis' choice
h3 <- 2*IQR(out_data)/(800^(1/3))
h3

## [1] 1.434799
```

(d)

- Description:

  Which of the three method change the least when outliers are added?

- Ans:

  According to the result of (b) and (c), I think the third method will change the k and h the leasst. Because method 3 only depends on the IQR of the data set, adding outliers will not have a big impact on its value.So I think that is why the method change the least.