BACS HW Week3 106071041

Question 1

(a) Distribution 2

new distribution negatively skewed
changing 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)

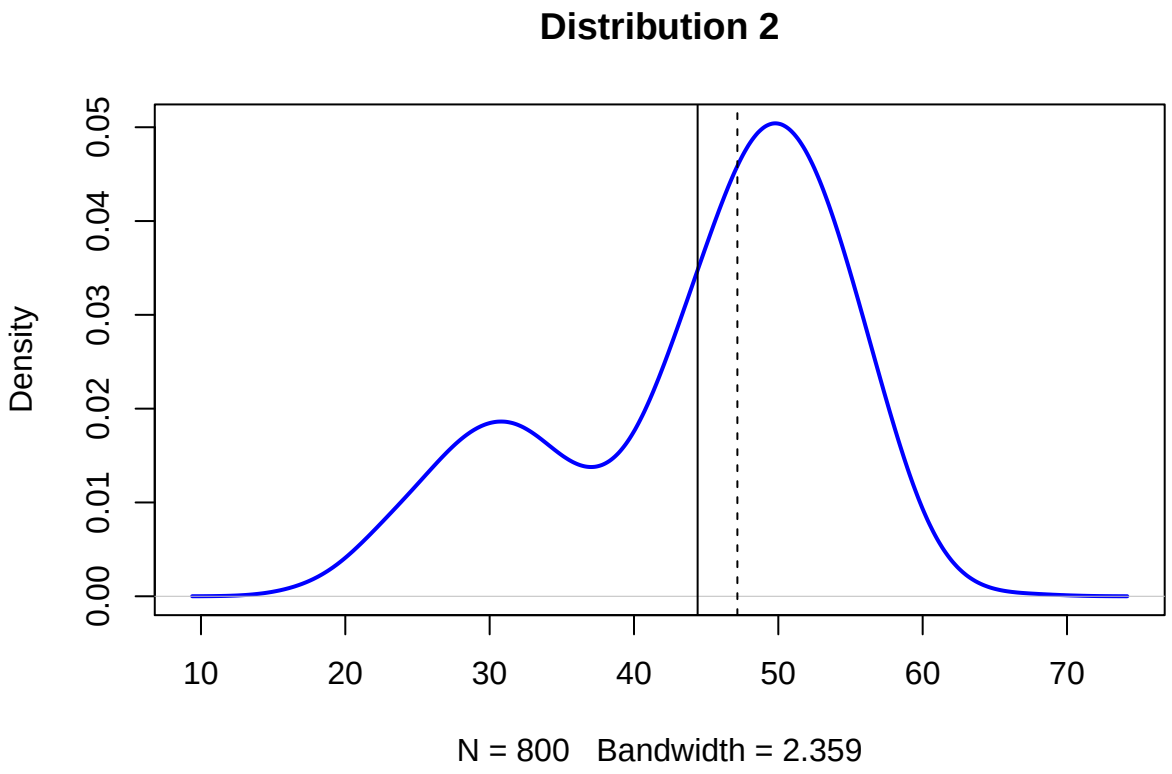
```
# Three normally distributed data sets (d1 changed)
d1 <- rnorm(n=500, mean=50, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)

# Let's combine them into a single dataset
d123 <- c(d1, d2, d3)

# 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
m1 <- mean(d123)
m2 <- median(d123)
cat("Mean:", m1, "\n")
```

```
## Mean: 44.40856
```

```
cat("Median:", m2)
```

```
## Median: 47.16248
```

```
abline(v=mean(d123))
abline(v=median(d123), lty="dashed")
```



(b) Distribution 3

normally distributed(`rnorm`,n=800)
compute the mean and median, and draw lines.

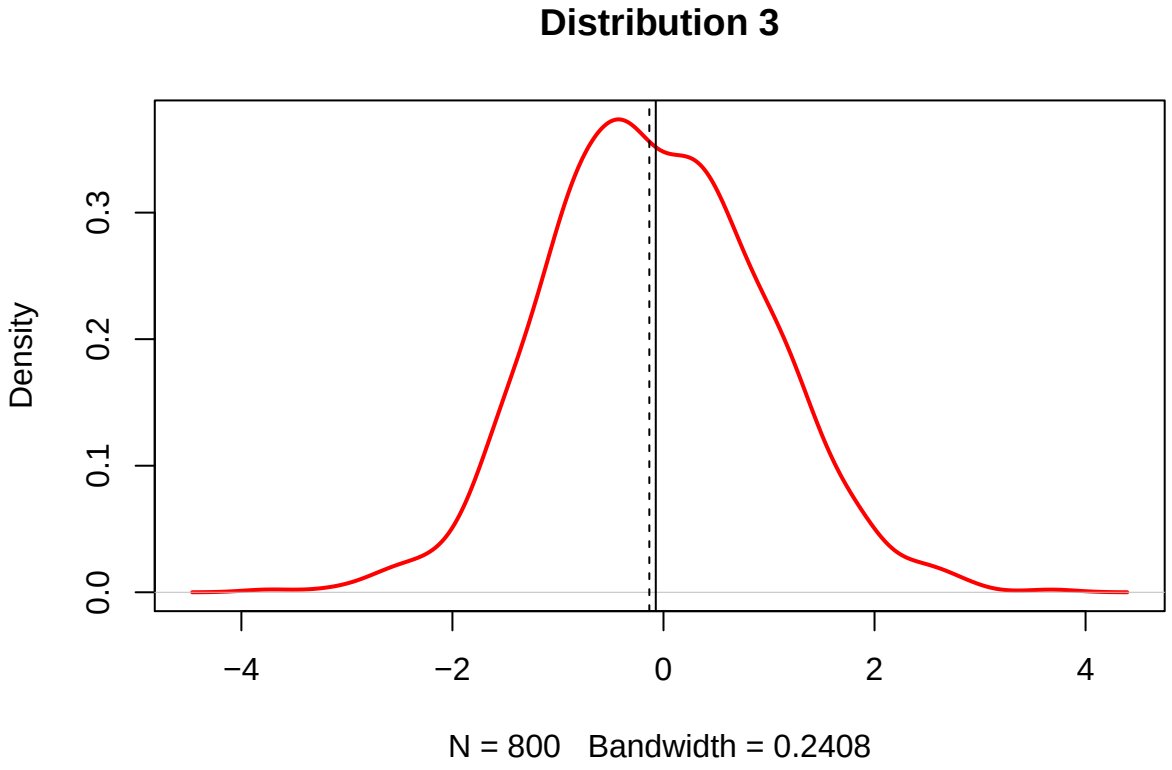
```
Distribution3 <- rnorm(800)
# Mean
mean(Distribution3)
```

```
## [1] -0.07281781
```

```
#Median
median(Distribution3)
```

```
## [1] -0.1340782
```

```
#Plot
plot(density(Distribution3), col="red", lwd=2,
     main = "Distribution 3")

#Add vertical lines showing mean and median
abline(v=mean(Distribution3))
abline(v=median(Distribution3), lty="dashed")
```



(c) More-sensitive measure, Mean or Median?

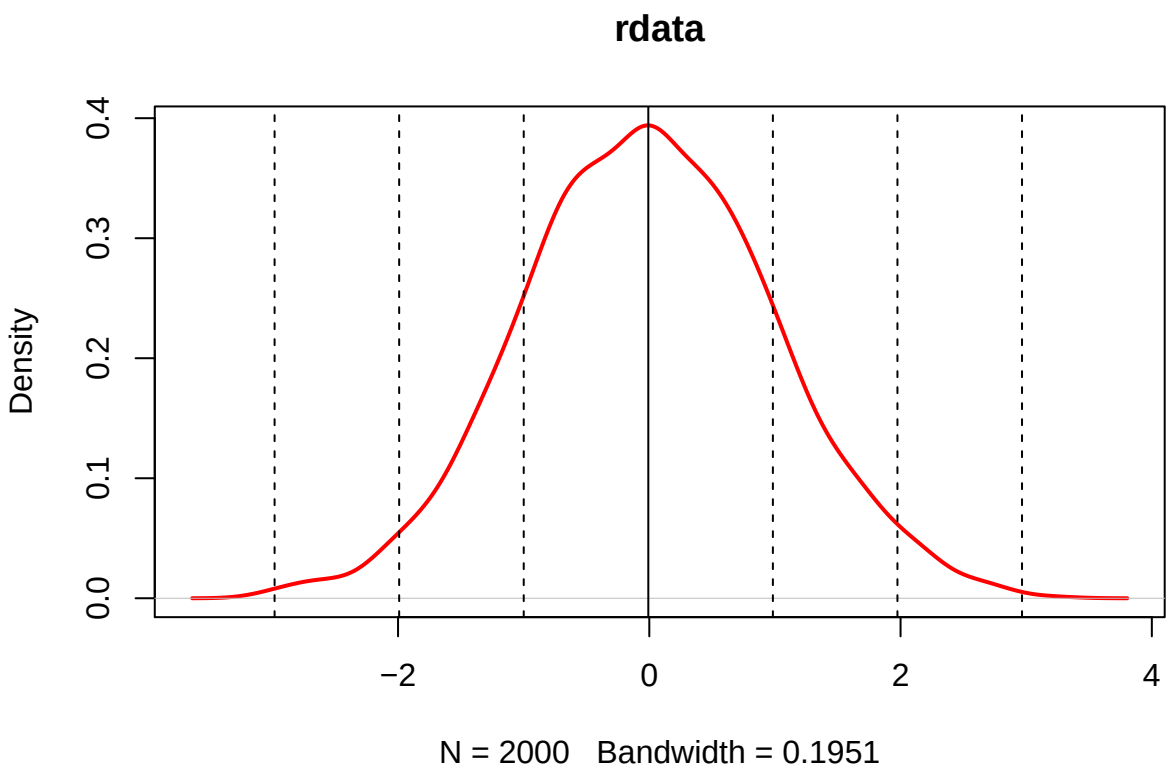
Mean. If Bill Gates come into the room, the average income will suddenly skyrocket while the Median of the income will be more or less the same.

Question 2

(a) normally-distributed random dataset

n=2000, mean=0, sd=1
Draw a density plot with 7 vertical lines(1 solid 6 dashed)

```
rdata <- rnorm(2000, 0, 1)
plot(density(rdata), col="red", lwd=2,
     main = "rdata")

avg <- mean(rdata)
std <- sd(rdata)

abline(v = avg)
abline(v=avg+std*(-3), lty="dashed")
abline(v=avg+std*(-2), lty="dashed")
abline(v=avg+std*(-1), lty="dashed")
abline(v=avg+std*(1), lty="dashed")
abline(v=avg+std*(2), lty="dashed")
abline(v=avg+std*(3), lty="dashed")
```



(b) Quantiles and distance away from mean

1st, 2nd, and 3rd quartiles corresponding distance with mean (keep +/-)

```
q1 <- quantile(rdata, 0.25)
q2 <- quantile(rdata, 0.5)
q3 <- quantile(rdata, 0.75)

q1_distance <- (q1-avg)/std
q2_distance <- (q2-avg)/std
q3_distance <- (q3-avg)/std

q1_distance
```

```
##      25%
## -0.68837
```

```
q2_distance
```

```
##          50%
## 1.856808e-06
```

```
q3_distance
```

```
##       75%
## 0.6654391
```

(c) New normally-distributed random dataset

n=2000, mean=35, sd=3.5
corresponding distance with mean for the 1st and 3rd quartiles
Compare your answer to (b)

```
rdata2 <- rnorm(2000, 35, 3.5)
avg2 <- mean(rdata2)
std2 <- sd(rdata2)

q12 <- quantile(rdata2, 0.25)
q32 <- quantile(rdata2, 0.75)

q12_distance <- (q12-avg2)/std2
q32_distance <- (q32-avg2)/std2

q12_distance
```

```
##        25%
## -0.6741787
```

```
q32_distance
```

```
##       75%
## 0.6697687
```

Compare your answer to (b) 1st quantile and 3rd quantile are both larger than those of (b)

(d) Dataset d123

corresponding distance with mean for the 1st and 3rd quartiles
Compare your answer to (b)

```
avg123 <- mean(d123)
std123 <- sd(d123)

q1_123 <- quantile(d123, 0.25)
q3_123 <- quantile(d123, 0.75)

q1_123_distance <- (q1_123-avg123)/std123
q3_123_distance <- (q3_123-avg123)/std123

q1_123_distance
```

```
##        25%
## -0.6808668
```

```
q3_123_distance
```

```
##       75%
## 0.7507834
```

Compare your answer to (b) 1st quantile is smaller than that of (b)
3rd quantile is larger than that of (b)

Question 3

(a) Rob Hyndman's suggested formula on the forum and the benefit of that formula

Freedman–Diaconis' choice
less sensitive to outliers in the data than the standard deviation

(b) Normally-distribution random dataset

`rand_data <- rnorm(800, mean=20, sd = 5)`
Compute the bin widths (h) and number of bins (k) with 3 methods

```
rand_data <- rnorm(800, mean=20, sd=5)
n <- 800
```

```
k1 = floor(log2(n))+1
h1 = (max(rand_data) - min(rand_data)) / k1
cat("the number of bins:", k1,"\n")
```

i. Sturges' formula

```
## the number of bins: 10
```

```
cat("the width of bins:", h1)
```

```
## the width of bins: 2.888545
```

```
std_rd <- sd(rand_data)
h2 = 3.49*std_rd/(n**(1/3))
k2 = ceiling((max(rand_data) - min(rand_data))/h2)
cat("the number of bins:", k2,"\n")
```

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

```
## the number of bins: 16
```

```
cat("the width of bins:", h2)
```

```
## the width of bins: 1.867398
```

```
q1_rd <- quantile(rand_data, 0.25)
q3_rd <- quantile(rand_data, 0.75)
iqr_rd <- q3_rd - q1_rd
h3 = 2*iqr_rd/(n**(1/3))
k3 = ceiling((max(rand_data) - min(rand_data))/h3)
cat("the number of bins:", k3,"\n")
```

iii. Freedman-Diaconis' choice (uses IQR)

```
## the number of bins: 21
```

```
cat("the width of bins:", h3)
```

```
## the width of bins: 1.410517
```

(c) Extend the rand_data dataset with some outliers

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

```
out_data <- c(rand_data, runif(10, min=40, max=60))
n_add_out <- 800+10
```

```
k12 = floor(log2(n_add_out))+1
h12 = (max(out_data) - min(out_data)) / k12
cat("the number of bins:", k12,"\n")
```

i. Sturges' formula

```
## the number of bins: 10
```

```
cat("the width of bins:", h12)
```

```
## the width of bins: 5.371185
```

```
std_out <- sd(out_data)
h22 = 3.49*std_out/(n_add_out**(1/3))
k22 = ceiling((max(out_data) - min(out_data))/h22)
cat("the number of bins:", k22,"\n")
```

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

```
## the number of bins: 25
```

```r
cat("the width of bins:", h22)
```

```
## the width of bins: 2.226278
```

```r
q1_rd2 <- quantile(out_data, 0.25)
q3_rd2 <- quantile(out_data, 0.75)
iqr_rd2 <- q3_rd2 - q1_rd2
h32 = 2*iqr_rd2/(n_add_out**(1/3))
k32 = ceiling((max(out_data) - min(out_data))/h32)
cat("the number of bins:", k32,"\n")
```

iii. Freedman-Diaconis' choice (uses IQR)

```
## the number of bins: 38
```

```r
cat("the width of bins:", h32)
```

```
## the width of bins: 1.420682
```

(d) The least sensitive one among the three measures

```r
sturge_diff <- h12-h1
scott_diff <- h22-h2
freeman_diff <- h32-h3

diff_list <- c(sturge_diff, scott_diff, freeman_diff)
names(diff_list) <- c("sturge_diff","scott_diff","freeman_diff")

diff_list
```

```
##  sturge_diff   scott_diff freeman_diff
##   2.48264042   0.35888022   0.01016471
```

```r
sort(diff_list)[1]
```

```
## freeman_diff
##   0.01016471
```

Brief Why while Sturge using only the number of the sample to compute, Scott and Freeman use some distribution index like standard deviation and IQR so they can have a better whole picture of the data. While standard deviation is more sensitive to outliers than IQR, which contains only 50% in the middle, the width(h) derived from the Freeman methods will change the least after some outliers added in.