BACS_HW2_106022113

Problem1

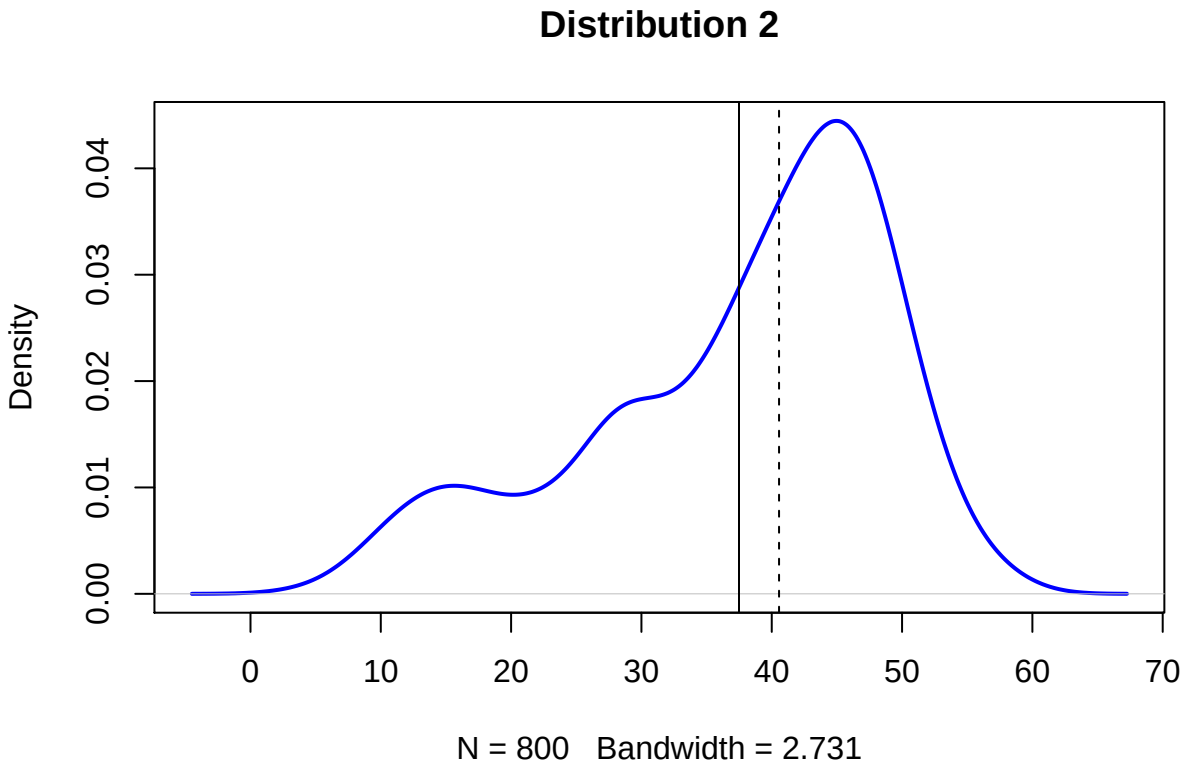
(a) Distribution 2 (left skewed)

```
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)

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

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



```r
# mean
mean(d123)
```

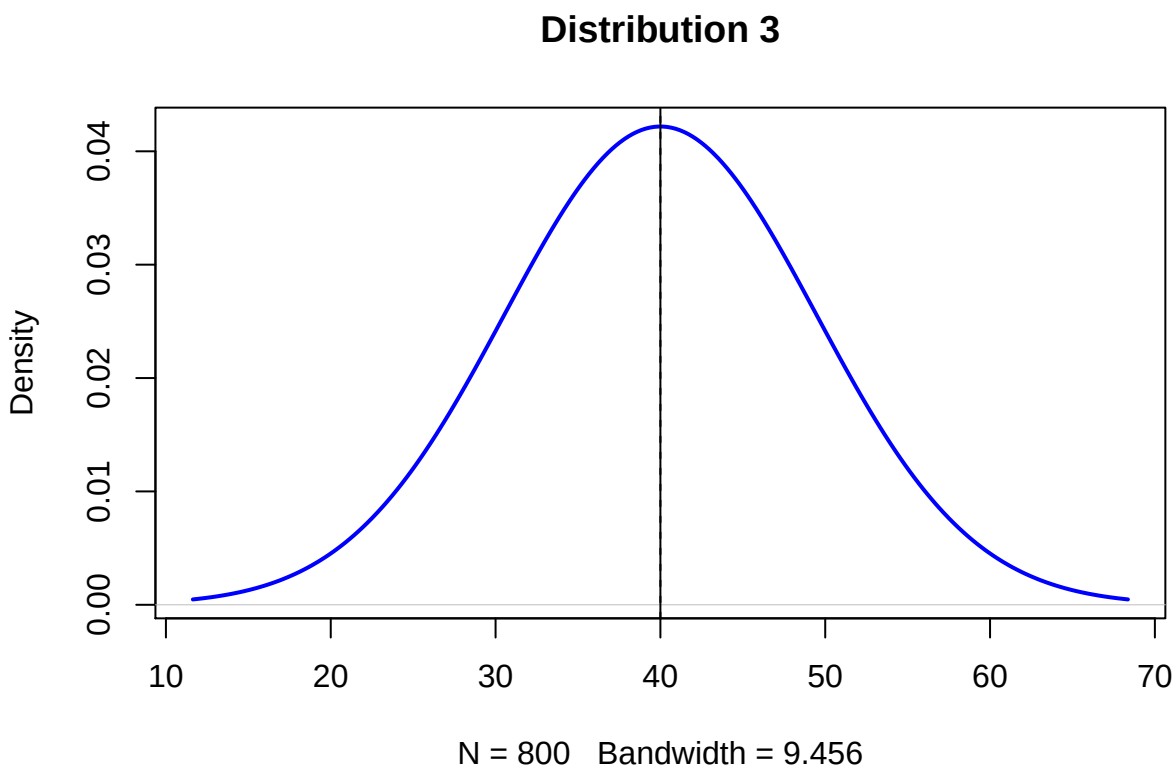
```
## [1] 37.48606
```

```r
# median
median(d123)
```

```
## [1] 40.56072
```

(b) Distribution 3 (Normally Distributed)

```r
d4 <- rnorm(n=800, mean=40, sd=0)
plot(density(d4), col="blue", lwd=2, main = "Distribution 3")
# Add vertical lines showing mean and median
abline(v=mean(d4))
abline(v=median(d4), lty="dashed")
```




```r
# mean
mean(d4)
```

```
## [1] 40
```

```r
# median
median(d4)
```

```
## [1] 40
```

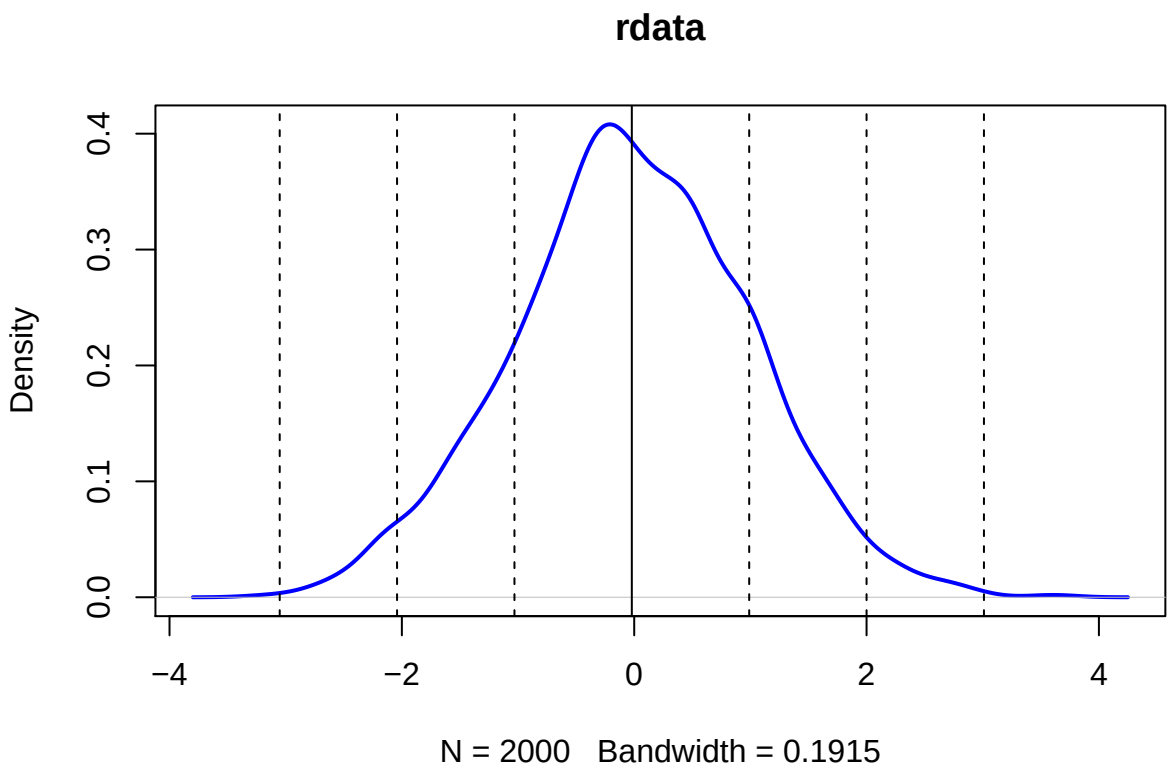
(c) mean or median more sensitive to outliers?

Ans: Mean is more sensitive to outliers since it is calculated within it. While medians are not distorted by the extreme values of outliers.

Problem 2

(a) rdata (7 vertical lines)

```r
rdata <- rnorm(n=2000, mean=0, sd=1)
plot(density(rdata), col="blue", lwd=2, main = "rdata")
sd <- sd(rdata)
#Add vertical lines
abline(v=mean(rdata))
abline(v=mean(rdata)-sd, lty="dashed")
abline(v=mean(rdata)-2*sd, lty="dashed")
abline(v=mean(rdata)-3*sd, lty="dashed")
abline(v=mean(rdata)+sd, lty="dashed")
abline(v=mean(rdata)+2*sd, lty="dashed")
abline(v=mean(rdata)+3*sd, lty="dashed")
```



(b) Calculate quantiles and sds away from the mean

```
q1 <- quantile(rdata, 0.25)
q1 # first quantile
```

```
##        25%
## -0.6636254
```

```
q2 <- quantile(rdata, 0.5)
q2 # second quantile
```

```
##         50%
## -0.03241373
```

```
q3 <- quantile(rdata, 0.75)
q3 # third quantile
```

```
##       75%
## 0.6402251
```

```r
#sds away from q1
(q1-mean(rdata))/sd(rdata)
```

```
##        25%
## -0.6364463
```

```r
#sds away from q2
(q2-mean(rdata))/sd(rdata)
```

```
##         50%
## -0.01167951
```

```r
#sds away from q3
(q3-mean(rdata))/sd(rdata)
```

```
##       75%
## 0.6540913
```

(c) new dataset compare with (b)

```r
d5 <- rnorm(n=2000, mean=35, sd=3.5)
#sds away from q1
(quantile(d5, 0.25)-mean(d5))/sd(d5)
```

```
##        25%
## -0.6725129
```

```r
#sds away from q3
(quantile(d5, 0.75)-mean(d5))/sd(d5)
```

```
##      75%
## 0.663895
```

Compared with (b), since the standard deviation is larger, the proportion of the distance will be larger too.

(d) dataset d123 compare with (b)

```r
#sds away from q1
(quantile(d123, 0.25)-mean(d123))/sd(d123)
```

```
##        25%
## -0.6387593
```

```r
#sds away from q3
(quantile(d123, 0.75)-mean(d123))/sd(d123)
```

```
##       75%
## 0.7652424
```

Ans: Compared with (b), since the distribution is not normal(skewed), the standard deviation distance between mean will not be the same for q1 and q3. And since the standard deviaiton varies for d123, the proportion of the distance will not be the same too.

Problem 3

(a) Formula suggested to calculate bins for histograms

Ans: The Freedman–Diaconis rule

$$h = 2\frac{IQR(x)}{\sqrt[3]{n}}$$

Benefits:It can minimize the difference between the area under the empirical probability distribution and the area under the theoretical probability distribution.

(b) Compute Bin width and number of bins using the formulas

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

```
n <- 800
# Number of Bins
k <- round(1 + log2(800))
k
```

i. Sturges' formula

```
## [1] 11
```

```
# Min Bin Width
h = (max(rand_data) - min(rand_data)) / k
h
```

```
## [1] 2.804417
```

```
sd <- sd(rand_data)
# Bin Width
h <- 3.49*sd/(n^(1/3))
h
```

ii. Scott's Normal Reference

```
## [1] 1.875739
```

```r
# Number of bins
k <- ceiling((max(rand_data) - min(rand_data))/h)
k
```

```
## [1] 17
```

```r
# Bin Width
h <- 2*IQR(rand_data)/n^(1/3)
h
```

iii. Freedman-Diaconis' choice

```
## [1] 1.496917
```

```r
# Number of bins
k <- ceiling((max(rand_data) - min(rand_data))/h)
k
```

```
## [1] 21
```

(c) compute new dataset bins with the three formulas

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

```r
n <- 810
# Number of Bins
k <- round(1 + log2(n))
k
```

i. Sturges' formula

```
## [1] 11
```

```r
# Min Bin Width
h = (max(out_data) - min(out_data)) / k
h
```

```
## [1] 5.203431
```

```
sd <- sd(out_data)
# Bin Width
h <- 3.49*sd/(n^(1/3))
h
```

ii. Scott's Normal Reference

```
## [1] 2.264728
```

```
# Number of bins
k <- ceiling((max(out_data) - min(out_data))/h)
k
```

```
## [1] 26
```

```
# Bin Width
h <- 2*IQR(out_data)/n^(1/3)
h
```

iii. Freedman-Diaconis' choice

```
## [1] 1.470456
```

```
# Number of bins
k <- ceiling((max(out_data) - min(out_data))/h)
k
```

```
## [1] 39
```

(d) Compare formula with least change

Ans:

The Freedman-Diaconis' Formula changed least in the bin width because it depends on the IQR, which is not affected by outliers.