HW2

111078513

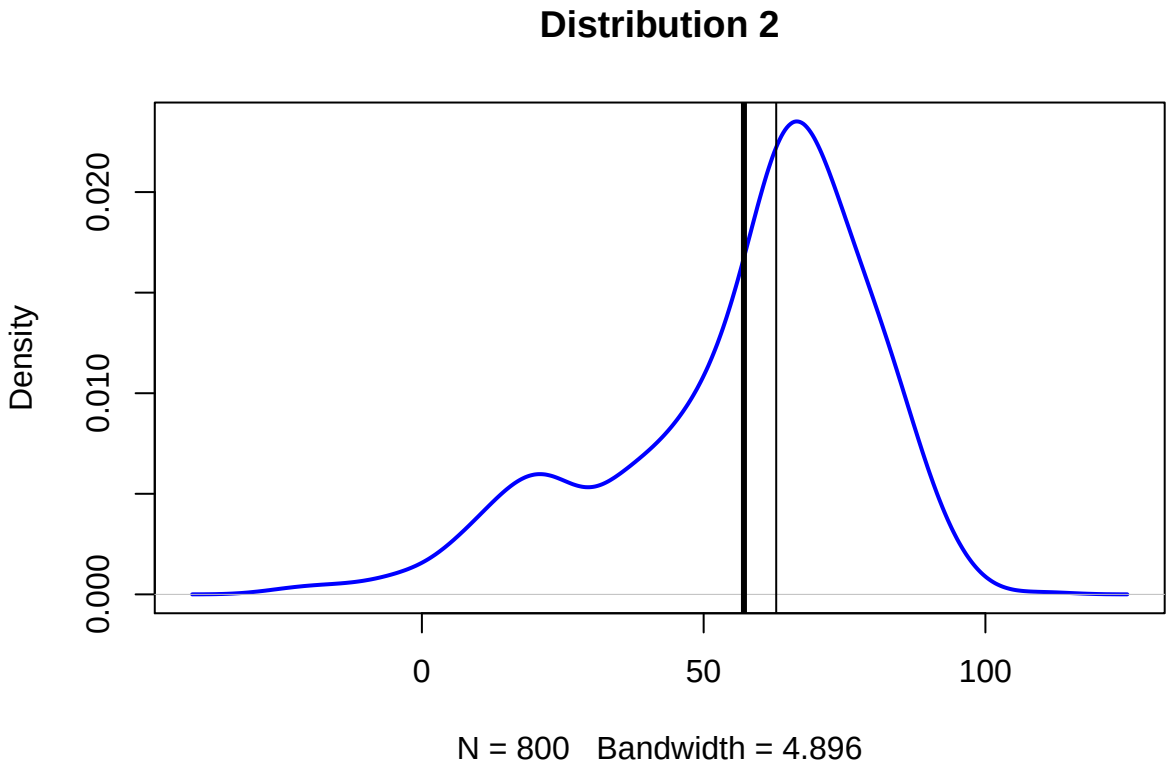
2023-02-25

Question 1

(a) Distributon 2

```
d1 <- rnorm(n=500, mean=70, sd=10)
d2 <- rnorm(n=200, mean=30, sd=20)
d3 <- rnorm(n=100, mean=50, sd=15)
# Combining them into a composite dataset
d123 <- c(d1, d2, d3)
# Let's plot the density function of d123
plot(density(d123), col="blue", lwd=2,
     main = "Distribution 2")

# Add vertical lines showing mean and median
abline(v=mean(d123), lwd = 3)
abline(v=median(d123), lwd = 1)
```

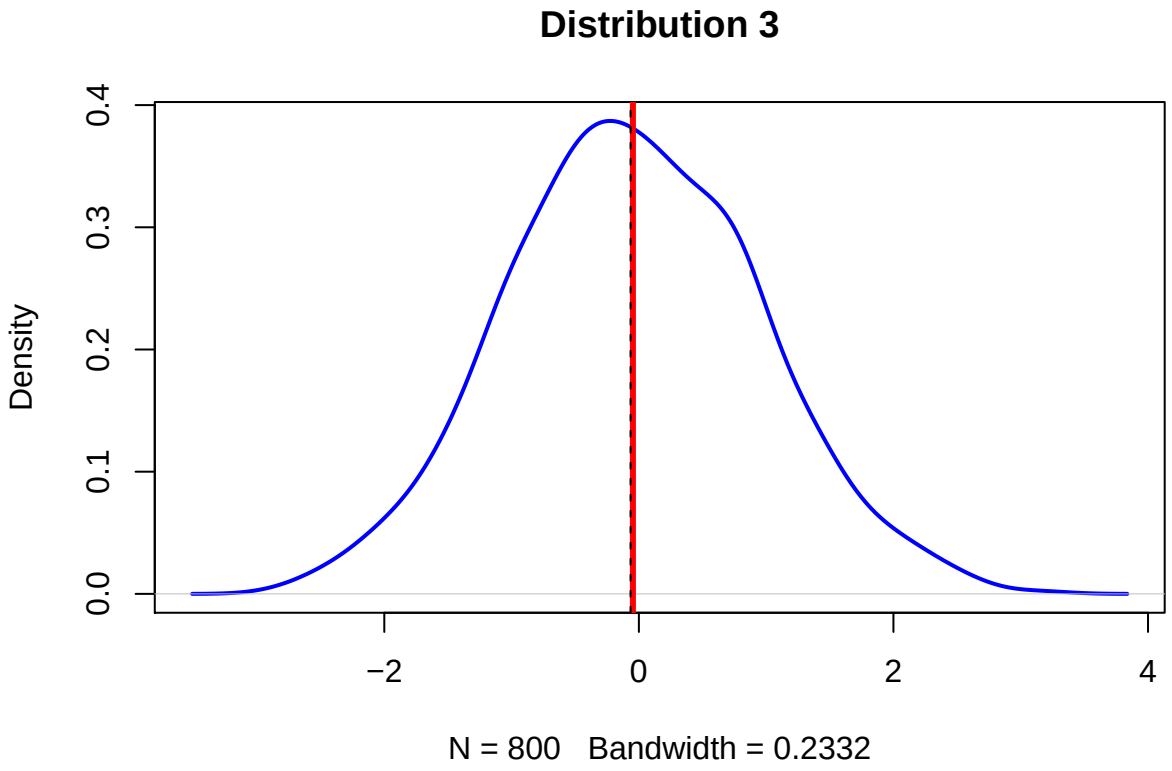


(b) Distribution 3

```
d <- rnorm(n=800)

# Let's plot the density function of d123
plot(density(d), col="blue", lwd=2,
     main = "Distribution 3")

# Add vertical lines showing mean and median
abline(v=mean(d), lwd = 3, col = "red")
abline(v=median(d), lty = 2, lwd = 1)
```



(c) 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?

I think that the "mean" is more sensitive to outliers because if a very large (or very small) number is added to a set of numbers, while the number of numbers only increases by one, the total sum of the numbers changes significantly. According to the calculation method of the mean, when the numerator changes significantly while the denominator only changes slightly, the score itself will change dramatically.

Question 2

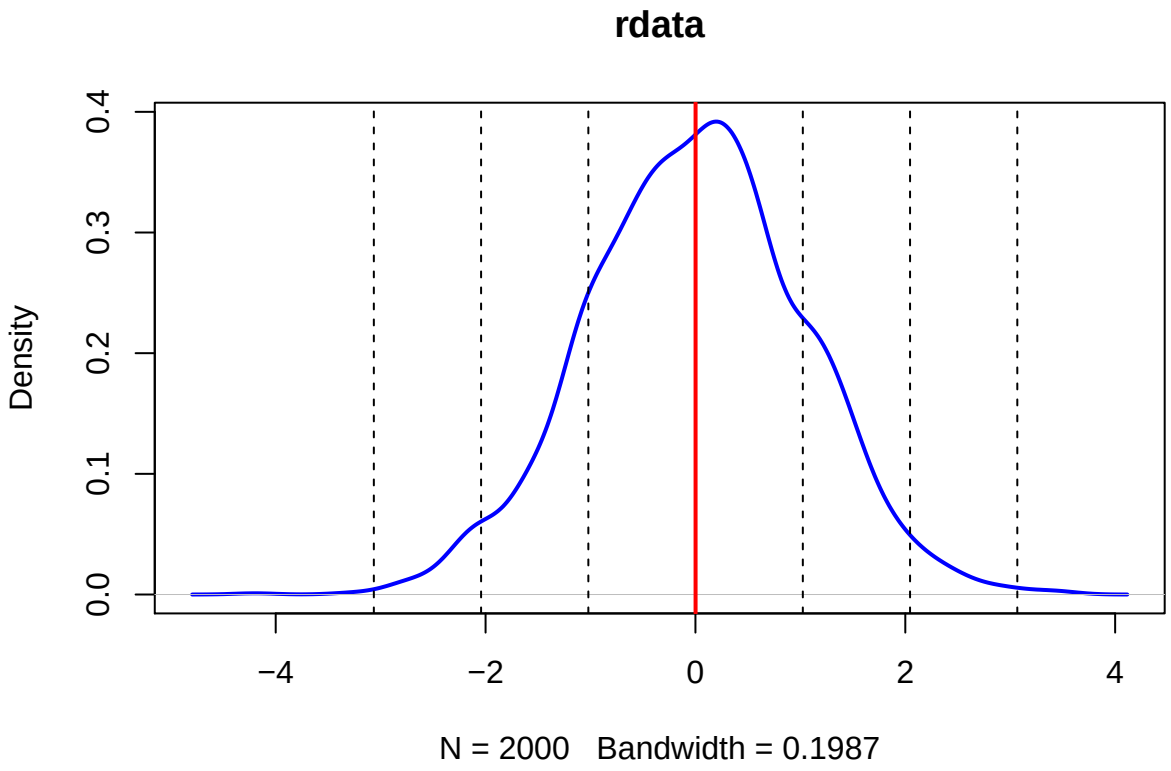
Set the function we will repeatedly use

```
q2 <- function(data_mean, data_sd){
  rdata <- rnorm(n = 2000, mean = data_mean, sd = data_sd)
  plot(density(rdata), col="blue", lwd=2,
       main = "rdata")
```

```
  rdata_mean = mean(rdata)
  rdata_sd = sd(rdata)
  rd_quantile = quantile(rdata, c(0.25, 0.5, 0.75))
  abline(v=rdata_mean, lwd = 2, col = "red")
  abline(v=c(rdata_mean - 3*rdata_sd,
             rdata_mean - 2*rdata_sd, rdata_mean - 1*rdata_sd,
             rdata_mean + 1*rdata_sd,
             rdata_mean + 2*rdata_sd, rdata_mean + 3*rdata_sd),
         lty = 2, lwd = 1)
  result = list()
  result$"rd_quantile" = rd_quantile
  result$"rd_quantile_z" = (rd_quantile-rdata_mean)/rdata_sd
  return(result)
}
```

(a)

```
rdata_mean_a <- 0
rdata_sd_a <- 1
q2_a <- q2(rdata_mean_a, rdata_sd_a)
```



(b)

```
#  Points correspond to the 1st, 2nd, and 3rd quartiles.
q2_a$rd_quantile
```

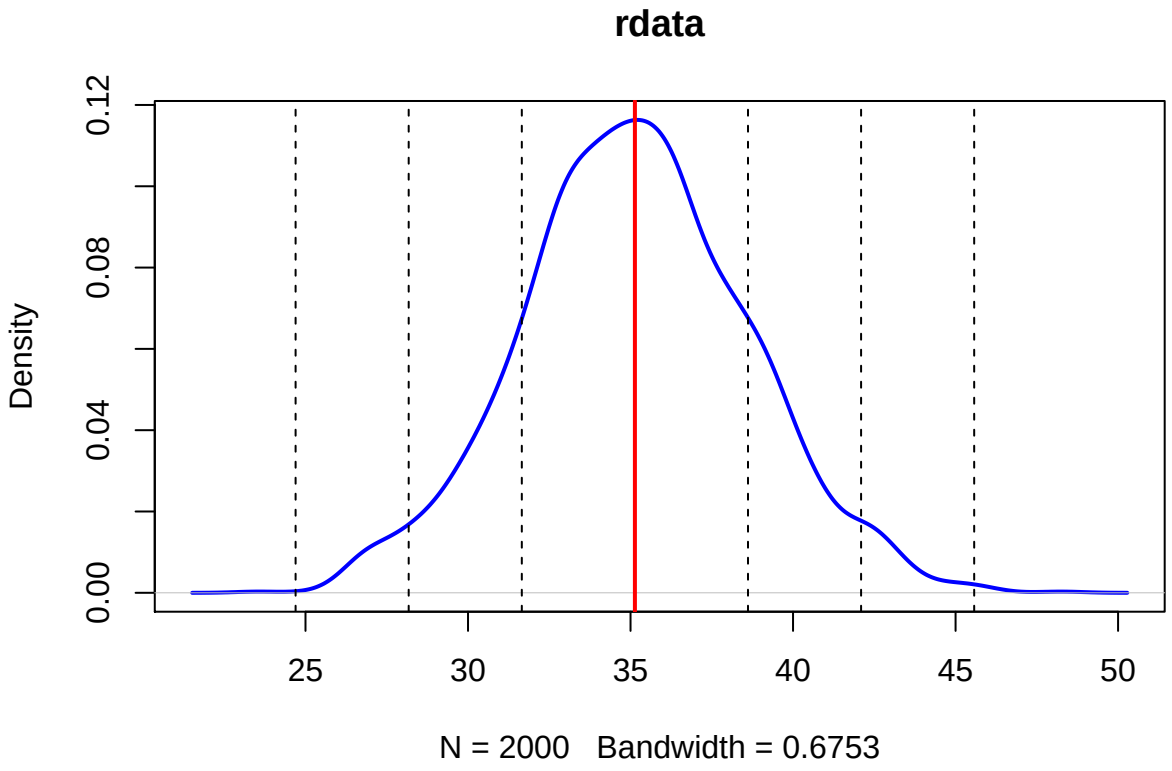
```
##         25%         50%         75%
## -0.69449624  0.01716808  0.65813795
```

```
# The amounts of standard deviations away from the mean.
q2_a$rd_quantile_z
```

```
##         25%         50%         75%
## -0.68087474  0.01546306  0.64262890
```

(c)

```
rdata_mean_c <- 35
rdata_sd_c <- 3.5
q2_c <- q2(rdata_mean_c, rdata_sd_c)
```




```
#  Points correspond to the 1st, 2nd, and 3rd quartiles.
q2_c$rd_quantile
```

```
##      25%      50%      75%
## 32.85795 35.08898 37.45575
```

```
# The amounts of standard deviations away from the mean.
q2_c$rd_quantile_z
```

```
##         25%         50%         75%
## -0.65419931 -0.01301635  0.66717829
```

We can compare the three shifted quartiles of distribution (B) to observe that two different distributions that can be approximated by a bell-shaped curve will have similar quartile results after undergoing the same shift and scaling.

(d)

```
d123_mean <- mean(d123)
d123_sd <- sd(d123)
#  Points correspond to the 1st, 2nd, and 3rd quartiles.
d123_quantile <- quantile(d123, c(0.25, 0.5, 0.75))
d123_quantile
```

```
##      25%      50%      75%
## 45.31116 62.87567 73.06435
```

```
# The amounts of standard deviations away from the mean.
(d123_quantile - d123_mean) / d123_sd
```

```
##        25%        50%        75%
## -0.5143364  0.2485920  0.6911453
```

By comparing the three shifted quartiles of distribution (B), we can observe that the quartiles of a left-skewed distribution will be significantly different from those of a bell-shaped distribution after undergoing the same scaling and shifting.

Question 3

(a)

(a)-1

He suggested to use Freedman-Diaconis.

(a)-2

The benefit of Freedman-Diaconis is that it is less sensitive than the standard deviation to outliers in data.

(b)

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

(b)-i

```
k = ceiling(log2(800))+1
h = (max(rand_data) - min(rand_data))/k
cat("Sturges' formula -> k:", k, ", h:", h)
```

```
## Sturges' formula -> k: 11 , h: 2.945006
```

(b)-ii

```
h = 3.49*sd(rand_data)/(800**(1/3))
k = ceiling((max(rand_data) - min(rand_data))/h)
cat("Scott's normal reference rule -> k:", k, ", h:", h)
```

```
## Scott's normal reference rule -> k: 17 , h: 1.916758
```

(b)-iii

```
h = 2*IQR(rand_data)/(800**(1/3))
k = ceiling((max(rand_data) - min(rand_data))/h)
cat("Freedman-Diaconis' choice -> k:", k, ", h:", h)
```

```
## Freedman-Diaconis' choice -> k: 22 , h: 1.511438
```

(c)

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

(c)-i

```
k = ceiling(log2(810))+1
h = (max(out_data) - min(out_data))/k
cat("Sturges' formula -> k:", k, ", h:", h)
```

```
## Sturges' formula -> k: 11 , h: 5.010726
```

(c)-ii

```
h = 3.49*sd(out_data)/(810**(1/3))
k = ceiling((max(out_data) - min(out_data))/h)
cat("Scott's normal reference rule -> k:", k, ", h:", h)
```

```
## Scott's normal reference rule -> k: 25 , h: 2.22957
```

(c)-iii

```
h = 2*IQR(out_data)/(810**(1/3))
k = ceiling((max(out_data) - min(out_data))/h)
cat("Freedman-Diaconis' choice -> k:", k, ", h:", h)
```

```
## Freedman-Diaconis' choice -> k: 36 , h: 1.538824
```

I found that when using Freedman-Diaconis' choice, my width values varied less, possibly because outliers do not have a significant impact on the IQR, resulting in less disturbance in the calculation of the width.