BACS-hw02-107070004

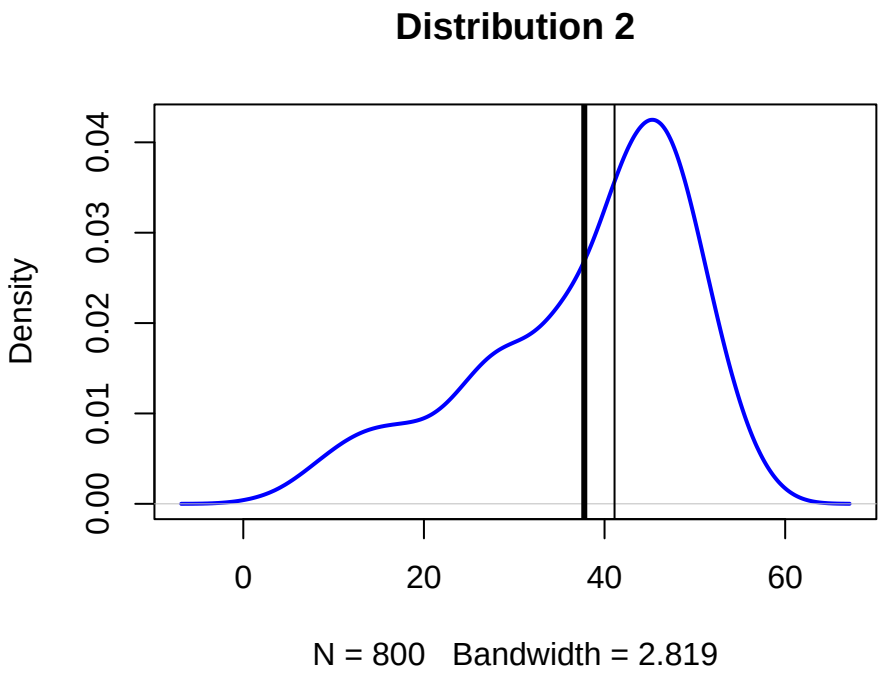
Question 1)

(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).

```
# New 3 normally distributed data sets
d1 <- rnorm(n=500, mean=45, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=15, sd=5)

# Composite dataset (n=800) that is negatively skewed
d123 <- c(d1, d2, d3)
plot(density(d123), col="blue", lwd=2,
     main = "Distribution 2")

# Add thick line for mean
abline(v=mean(d123), lwd=3)
# Add thin line for median
abline(v=median(d123), lwd=1)
```



(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).

```
# New a single dataset normally distributed (bell-shaped, symmetric)
d4 <- rnorm(800)

# Show the plot
plot(density(d4), col="blue", lwd=2,
     main = "Distribution 3")


# Add thick line for mean
mean(d4)
```

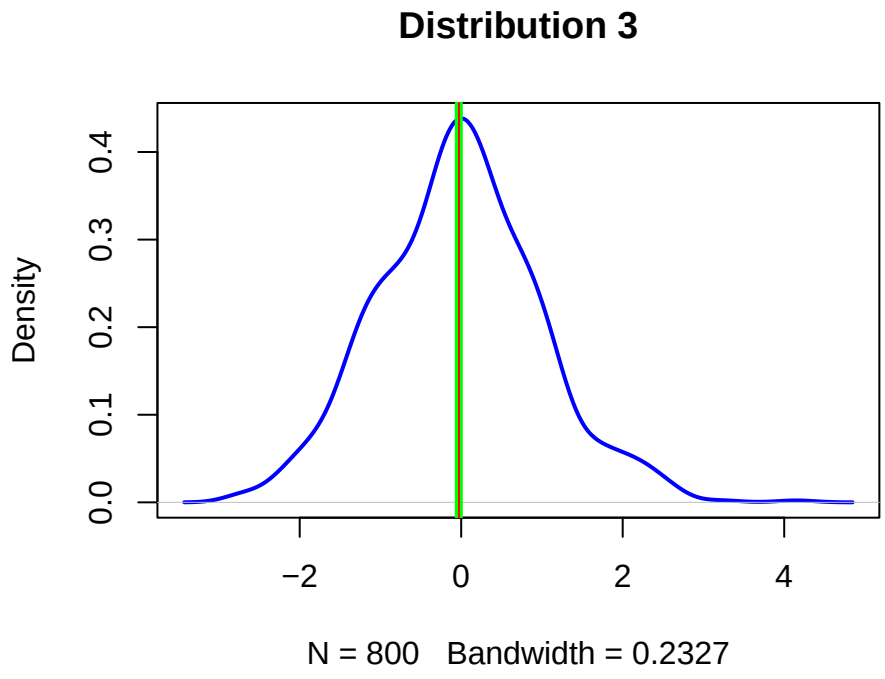
```
## [1] -0.02986477
```

```
abline(v=mean(d4), lwd=4, col="green")

# Add thin line for median
median(d4)
```

```
## [1] -0.02744064
```

```
abline(v=median(d4), lwd=1, 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?

The mean is more sensitive to the existence of outliers than **the median**.

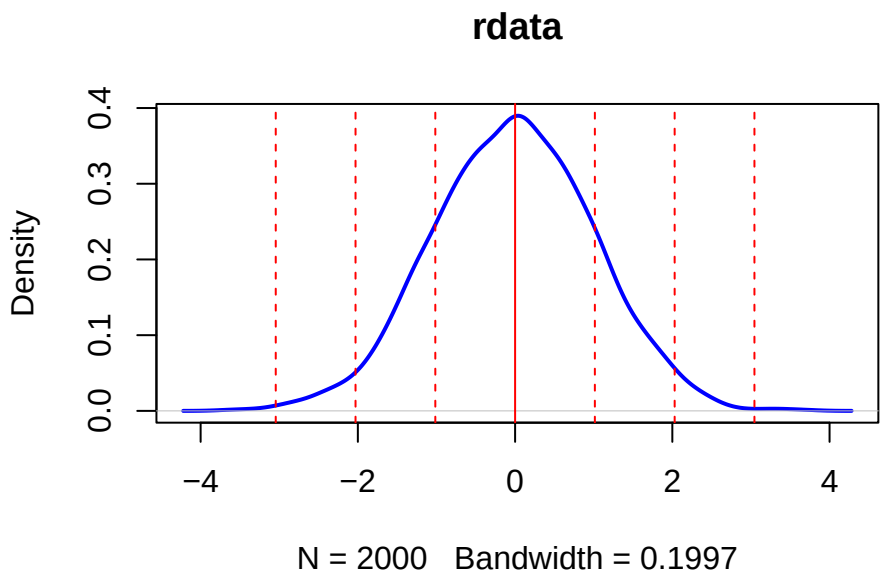
Explanation: Consider the initial retirement age dataset again, with one difference; the last observation of 60 years has been replaced with a retirement age of 81 years. This value is much higher than the other values, and could be considered an outlier. However, it has not changed the middle of the distribution, and therefore the median value is still 57 years. 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 81 As the all values are included in the calculation of the mean, the outlier will influence the mean value. (54+54+54+55+56+57+57+58+58+60+81 = 644), divided by 11 = 58.5 years In this distribution the outlier value has increased the mean value.

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).

```
# New a single dataset normally distributed (n=2000, mean=0, sd=1)
rdata <- rnorm(n=2000, mean=0, sd=1)
# Show the plot
plot(density(rdata), col="blue", lwd=2,
     main = "rdata")

# Add lines for mean and Q1, Q2, Q3 of the left and right of the mean
abline(v=seq(from=-3,to = 3)*sd(rdata), lwd=1, col="red", lty=c(2,2,2,1,2,2,2))
```



b) Using the quantile() function, which data points correspond to the 1st, 2nd, and 3rd quartiles (i.e., 25th, 50th, 75th percentiles) of rdata? 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?

```
quantile <- quantile(x = rdata, probs = c(1/4,2/4,3/4))
(quantile-mean(rdata))/sd(rdata)
```

```
##         25%         50%         75%
## -0.68545489  0.01728829  0.68637094
```

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)

```
nrdata <- rnorm(n = 2000,mean = 35,sd = 3.5)
nquantile <- quantile(x = nrdata, probs = c(1/4,3/4))
(nquantile-mean(nrdata))/sd(nrdata)
```

```
##        25%        75%
## -0.6554264  0.6526375
```

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)

```
d123_quantile<-quantile(x = d123,probs = c(1/4,3/4))
(d123_quantile-mean(d123))/sd(d123)
```

```
##        25%        75%
## -0.6506507  0.7362464
```

There are no huge difference between (b) and (d).

Question 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?

He suggests to use Freedman-Diaconis rule which is very robust and works well in practice. It's based on the interquartile range, denoted by IQR. It replaces sd of Scott's rule with IQR, which is less sensitive than the standard deviation to outliers in data.

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. Sturges' formula
ki <- nclass.Sturges(rand_data)
hi <- (max(rand_data)-min(rand_data))/ki
cat("Number of bins is", ki,",and bin widths is", hi)
```

```
## Number of bins is 11 ,and bin widths is 3.016784
```

```
# ii. Scott's normal reference rule (uses standard deviation)
hii <- 3.49*sd(rand_data)/(length(rand_data)^(1/3))
kii <- ceiling((max(rand_data)-min(rand_data))/hii)
cat("Number of bins is", kii,",and bin widths is", hii)
```

```
## Number of bins is 18 ,and bin widths is 1.859827
```

```
# iii. Freedman-Diaconis' choice (uses IQR)
hiii <- 2*IQR(rand_data)/(length(rand_data)^(1/3))
kiii <- ceiling((max(rand_data)-min(rand_data))/hiii)
cat("Number of bins is", kiii,",and bin widths is", hiii)
```

```
## Number of bins is 24 ,and bin widths is 1.411638
```

c) Repeat part (b) but extend the rand_data dataset with some outliers (create a new dataset out_data): out_data <- c(rand_data, runif(10, min=40, max=60)) 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?

```
out_data <- c(rand_data, runif(10, min=40, max=60))
# i. Sturges' formula
nki <- nclass.Sturges(out_data)
nhi <- (max(out_data)-min(out_data))/nki
cat("Number of bins is", nki,",and bin widths is", nhi)
```

```
## Number of bins is 11 ,and bin widths is 5.081545
```

```
cat("Change for bin widths is", nhi - hi)
```

```
## Change for bin widths is 2.064761
```

```
# ii. Scott's normal reference rule (uses standard deviation)
nhii <- 3.49*sd(out_data)/(length(out_data)^(1/3))
nkii <- ceiling((max(out_data)-min(out_data))/nhii)
cat("Number of bins is", nkii,",and bin widths is", nhii)
```

```
## Number of bins is 26 ,and bin widths is 2.209219
```

```
cat("Change for bin widths is", nhii - hii)
```

```
## Change for bin widths is 0.3493916
```

```
# iii. Freedman-Diaconis' choice (uses IQR)
nhiii <- 2*IQR(out_data)/(length(out_data)^(1/3))
nkiii <- ceiling((max(out_data)-min(out_data))/nhiii)
cat("Number of bins is", nkiii,",and bin widths is", nhiii)
```

```
## Number of bins is 39 ,and bin widths is 1.453798
```

```r
cat("Change for bin widths is", nhiii - hiii)
```

```
## Change for bin widths is 0.0421605
```

The Freedman-Diaconis' choice has least sensitive to outliers. Because IQR is more robust to quantify the amount of variation than the others.