HW2 BACS

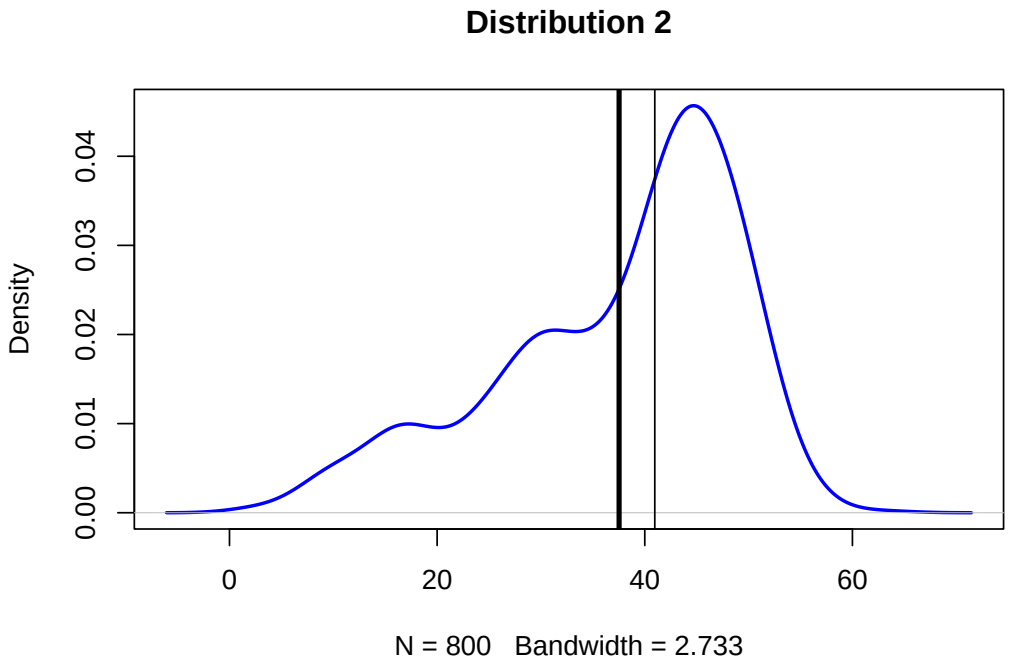
109006206

2023-03-04

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).

```
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),lwd = "3")
abline(v=median(d123))
```



(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).

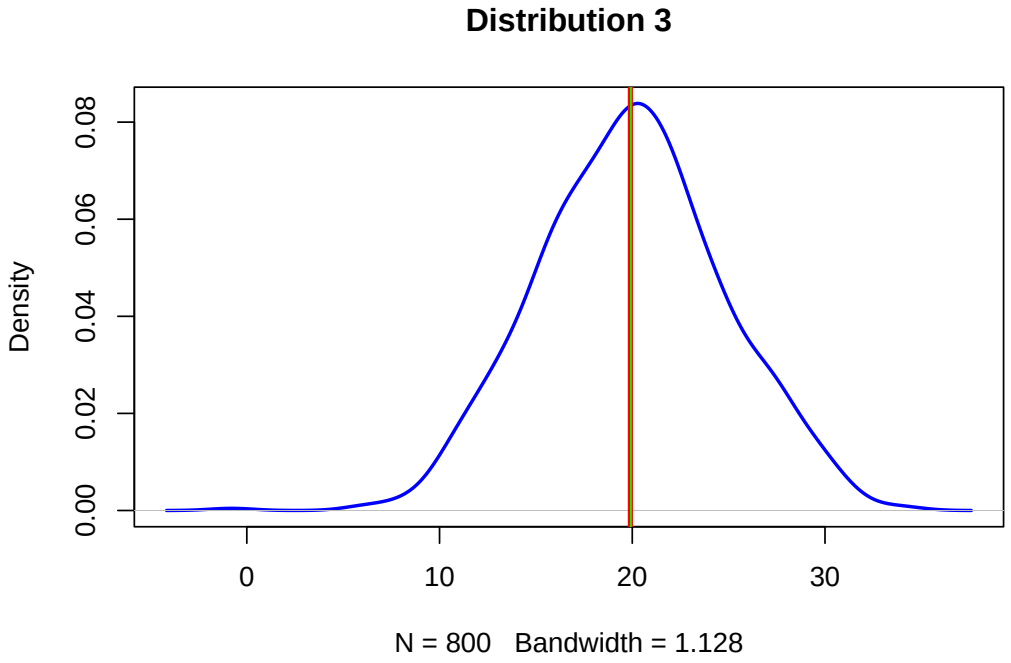
```
d1 <- rnorm(n=800, mean=20, sd=5)
plot(density(d1), col="blue", lwd=2,
     main = "Distribution 3")
cat("Mean is :",mean(d1),"\n")
```

```
## Mean is : 19.90181
```

```
cat("Median is :",median(d1),"\n")
```

```
## Median is : 19.93164
```

```
abline(v=mean(d1),lwd = "3",col = " red")
abline(v=median(d1),col = " green")
```



(Notes : red line is mean & green line is median)

(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?

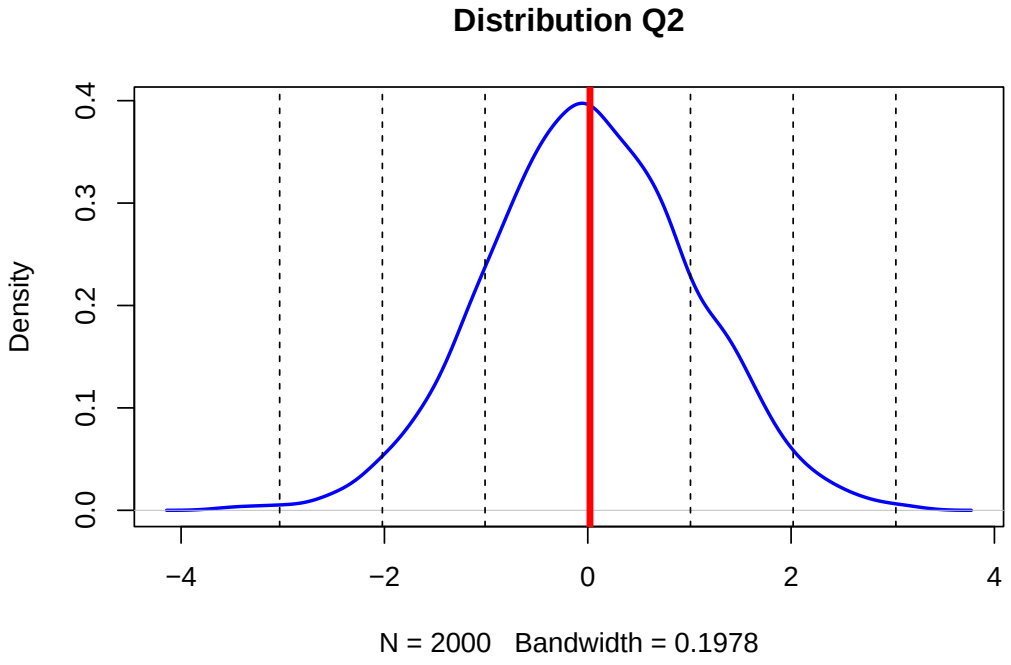
Answer : Mean is more likely to be affected by outliers in the data since we have to sum all the values and divided by the number of data which the data may have outliers

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).

```
rdata <- rnorm(n=2000, mean=0, sd=1)
plot(density(rdata), col="blue", lwd=2, main = "Distribution Q2")
abline(v=mean(rdata),lwd = "4",col = " red")

lines <- seq(-3,-1)
abline(v=sd(rdata)*lines,lty = "dashed")
lines2 <- seq(1,3)
abline(v=sd(rdata)*lines2,lty = "dashed")
```

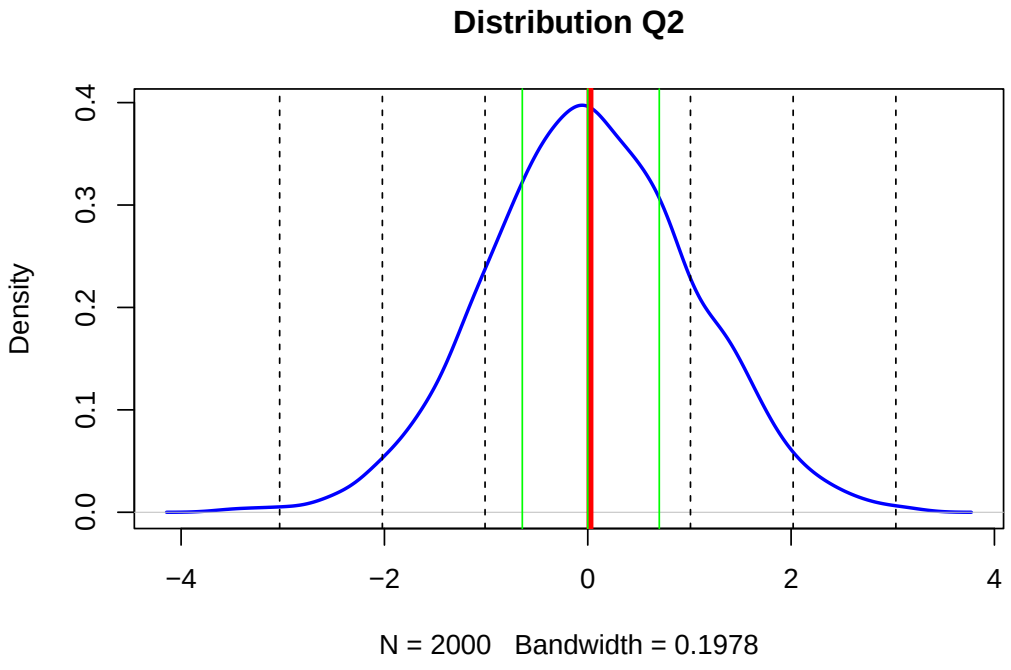


b) Using the quantile() function, which data points correspond to the 1st, 2nd, and 3rd quartiles (i.e., 25th, 50th, 75th percentiles) of rdata? How many standard deviations away from the mean (divide by standard-deviation; keep positive or negative sign) are those points corresponding to the 1st, 2nd, and 3rd quartiles?

Here I create distance of standard deviations and mean function

```
dist<- function(data){
  first = (quantile(data,0.25)-mean(data))/sd(data)
  second = (quantile(data,0.50)-mean(data))/sd(data)
  third = (quantile(data,0.75)-mean(data))/sd(data)
  cat(round(first,5),"standard deviations away from the mean corresponding to 1st quartile\n")
  cat(round(second,5),"standard deviations away from the mean corresponding to 2nd quartile\n")
  cat(round(third,5),"standard deviations away from the mean corresponding to 3rd quartile\n")
}
```

```
q_data<-c(0.25,0.50,0.75)
plot(density(rdata), col="blue", lwd=2, main = "Distribution Q2")
abline(v=mean(rdata),lwd = "4",col = " red")
rdata_q<-abline( v = quantile(rdata,q_data),col = "green")
lines <- seq(-3,-1)
abline(v=sd(rdata)*lines,lty = "dashed")
lines2 <- seq(1,3)
abline(v=sd(rdata)*lines2,lty = "dashed")
```



Answer:

```
quantile(rdata,q_data)
```

```
##           25%           50%           75%
## -6.437387e-01  7.059698e-05  7.033551e-01
```

```
dist(rdata)
```

```
## -0.65882 standard deviations away from the mean corresponding to 1st quartile
## -0.02151 standard deviations away from the mean corresponding to 2nd quartile
## 0.67468 standard deviations away from the mean corresponding to 3rd quartile
```

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)

```
rdata.new<-rnorm(n = 2000,mean = 35,sd = 3.5)
dist(rdata.new)
```

```
## -0.68178 standard deviations away from the mean corresponding to 1st quartile
## 0.03053 standard deviations away from the mean corresponding to 2nd quartile
## 0.66553 standard deviations away from the mean corresponding to 3rd quartile
```

Answer:
Compared to (b), there is no big difference between the number of the 1st and 3rd quartiles.

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)

```
dist(d123)
```

```
## -0.64607 standard deviations away from the mean corresponding to 1st quartile
## 0.29728 standard deviations away from the mean corresponding to 2nd quartile
## 0.73853 standard deviations away from the mean corresponding to 3rd quartile
```

Answer:
Compared to (b), we can see that (d)'s 1st and 3rd quartiles are slightly larger.

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?

Answer:

$$numberofbins = \frac{max - min}{h}$$

$$h = 2 \times IQR \times n^{-1/3}$$

Notes:
1. h : bin-width
2. IQR : Inter Quartile Range
3. n : number of observations

Benefits:
1. Minimize the integral of the squared difference between the histogram
2. Minimize the density of the theoretical probability distribution
3. It is the least sensitive to outliers.

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:

1.Sturges' formula

$$k = [log2n] + 1$$

Answer:

```
rand_data <- rnorm(800, mean=20, sd = 5)
k_sturges<- ceiling(log2(length(rand_data)))+1
h_sturges<- (max(rand_data)-min(rand_data))/k_sturges
cat("The number of bins :",k_sturges)
```

```
## The number of bins : 11
```

```
cat("The bin width :",h_sturges)
```

```
## The bin width : 2.558479
```

2. Scott's normal reference rule

$$h = \frac{3.49\hat{\sigma}}{\sqrt[3]{n}},$$

Note : where σ is the sample standard deviation.

```
rand_data <- rnorm(800, mean=20, sd = 5)
h_scott<- (3.49*sd(rand_data))/length(rand_data)^(1/3)
k_scott<- ceiling((max(rand_data)-min(rand_data))/h_scott)
cat("The number of bins :",k_scott)
```

```
## The number of bins : 17
```

```
cat("The bin width :",h_scott)
```

```
## The bin width : 1.930874
```

3. Freedman-Diaconis' choice

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

```
rand_data <- rnorm(800, mean=20, sd = 5)
h_freed<- (2*IQR(rand_data))/length(rand_data)^(1/3)
k_freed<- ceiling((max(rand_data)-min(rand_data))/h_freed)
cat("The number of bins :",k_freed)
```

```
## The number of bins : 23
```

```
cat("The bin width :",h_freed)
```

```
## The bin width : 1.450378
```

c) Repeat part (b) but let's extend rand_data dataset with some outliers (creating 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?

1.Sturges' formula

```
out_data <- c(rand_data, runif(10, min=40, max=60))
k_sturges<- ceiling(log2(length(out_data)))+1
h_sturges<- (max(out_data)-min(out_data))/k_sturges
cat("The bin width :",h_sturges)
```

```
## The bin width : 5.177849
```

2. Scott's normal reference rule

```
out_data <- c(rand_data, runif(10, min=40, max=60))
h_scott<- (3.49*sd(out_data))/length(out_data)^(1/3)
k_scott<- ceiling((max(out_data)-min(out_data))/h_scott)
cat("The bin width :",h_scott)
```

```
## The bin width : 2.304336
```

3. Freedman-Diaconis' choice

```
out_data <- c(rand_data, runif(10, min=40, max=60))
h_freed<- (2*IQR(out_data))/length(out_data)^(1/3)
k_freed<- ceiling((max(out_data)-min(out_data))/h_freed)
cat("The bin width :",h_freed)
```

```
## The bin width : 1.448241
```

Answer:

After we compared (b) and (c) we can conclude **Freedman-Diaconis' choice** is the least sensitive to outliers. It replaces 3.49 σ of Scott's rule with 2 IQR, which is less sensitive than the standard deviation to outliers in data. Sturges' rule was also found inaccurate for n larger that 200.