Homework2_BACS

109090035

2/27/2023

Q1

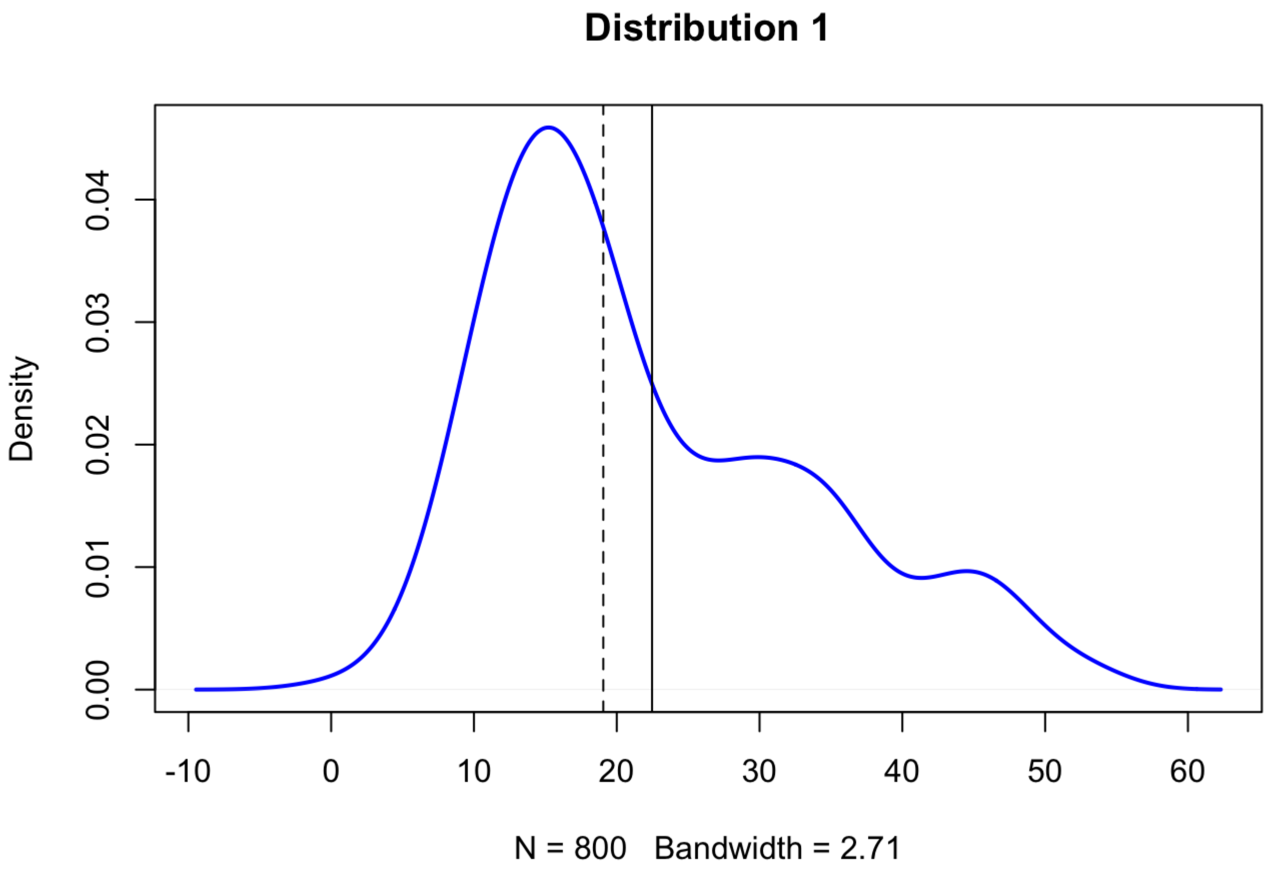
In the box below, we have created a composite distribution by combining three normal distributions, and drawn a density plot. The mean (thick line) and median (dashed line) are drawn as well. Two important things to observe: first, the distribution is positively skewed (tail stretches to the right); second, the mean and median are different.

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

# 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 1")

# Add vertical lines showing mean and median
abline(v=mean(d123))
abline(v=median(d123), lty="dashed")
```



## (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).

```
# Three 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)

# Combining them into a composite dataset
d123 <- c(d1, d2, d3)
mean(d123)
```

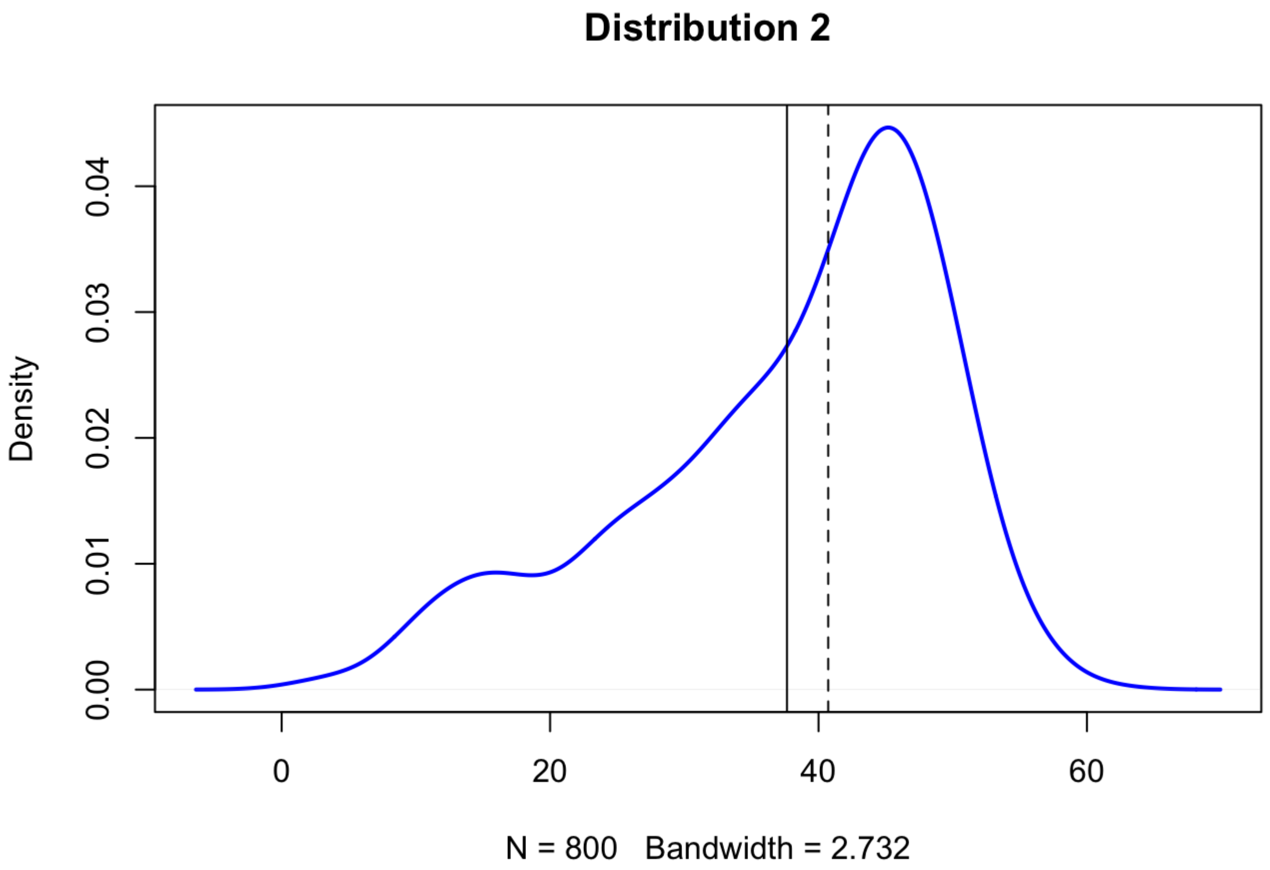
```
## [1] 37.65201
```

```
sd(d123)
```

```
## [1] 11.65669
```

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



## (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).

```
# Three normally distributed data sets
d1 <- rnorm(n=800, mean=0, sd=1)
mean(d1)
```

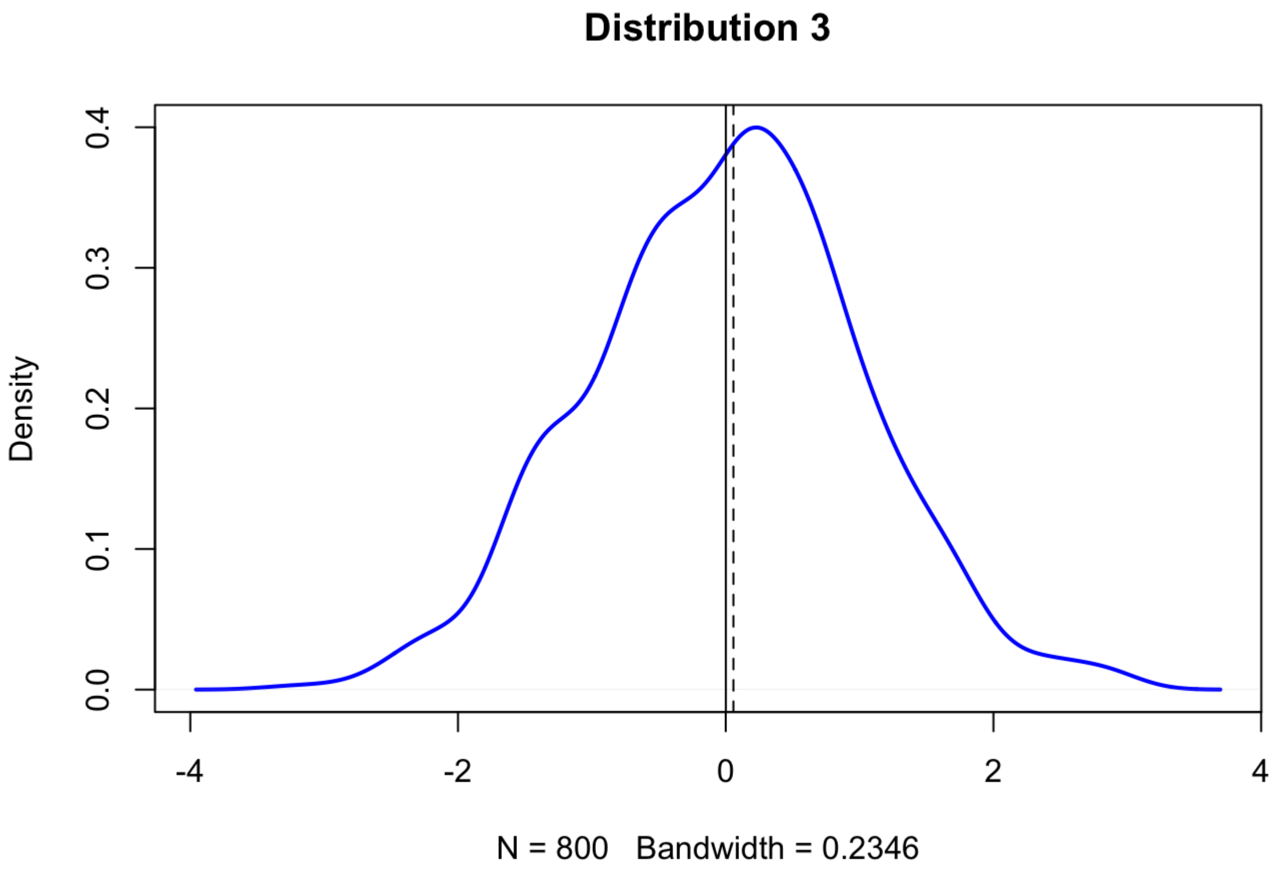
```
## [1] 0.001098512
```

```
sd(d1)
```

```
## [1] 1.012586
```

```
# Let’s plot the density function of d1
plot(density(d1), col="blue", lwd=2,
     main = "Distribution 3")

# Add vertical lines showing mean and median
abline(v=mean(d1))
abline(v=median(d1), lty="dashed")
```



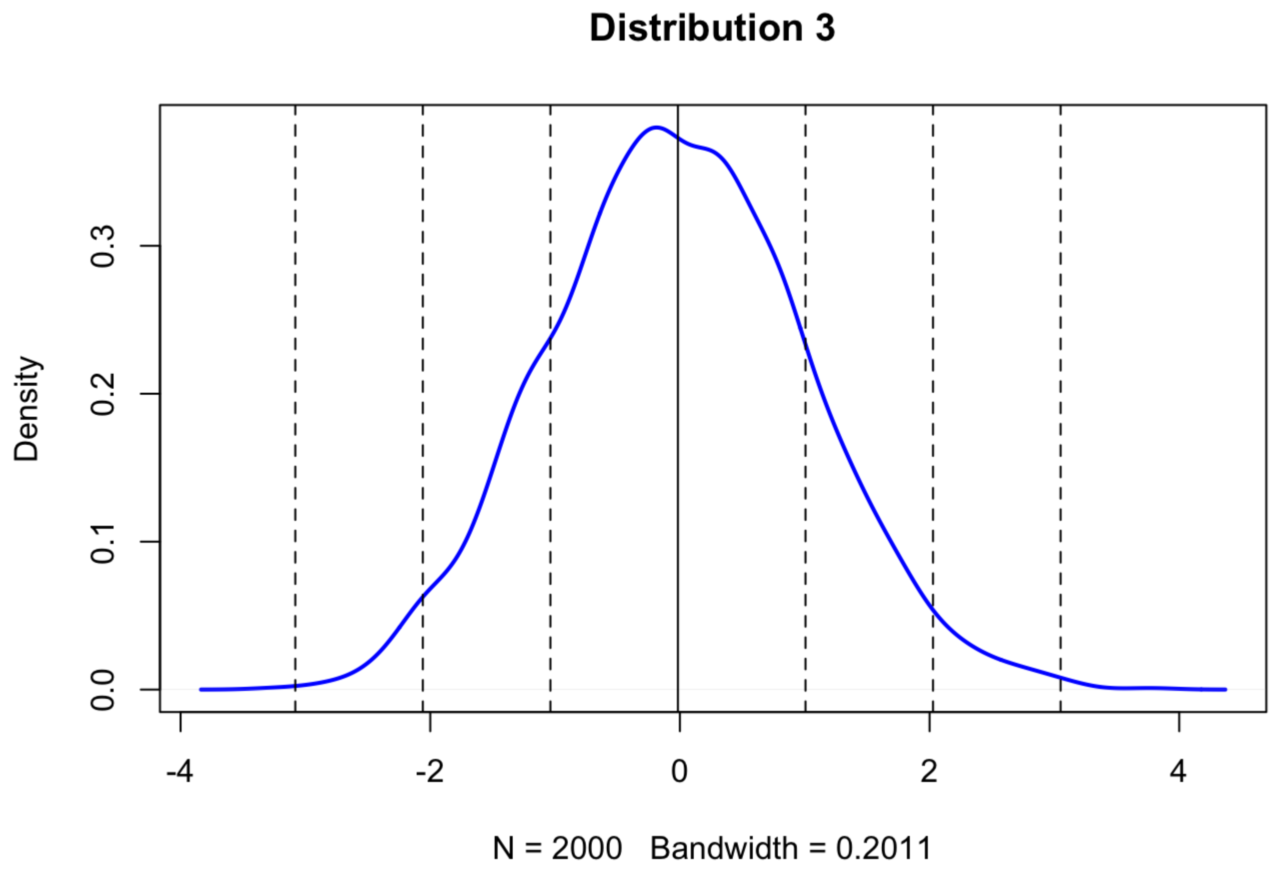
## (c) In general, which measure of central tendency (mean or median) do you think will be more sensitive (will change more) to outliers being added to your data?

Answer: Extreme values do not influence the center portion of a distribution. This means that the median of a sample taken from a distribution is not influenced so much,hence the mean is more sensitive to the existence of outliers than the median.

Q2

(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).

```
# Three normally distributed data sets
rdata <- rnorm(n=2000, mean=0, sd=1)
# Let’s plot the density function of d1
plot(density(rdata), col="blue", lwd=2,
     main = "Distribution 3")
# Add vertical lines showing mean and median
abline(v=mean(rdata))
abline(v=mean(rdata) + sd(rdata), lty="dashed")
abline(v=mean(rdata) + 2*sd(rdata), lty="dashed")
abline(v=mean(rdata) + 3*sd(rdata), lty="dashed")
abline(v=mean(rdata) - sd(rdata), lty="dashed")
abline(v=mean(rdata) - 2*sd(rdata), lty="dashed")
abline(v=mean(rdata) - 3*sd(rdata), 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?

```
#The  1st, 2nd, and 3rd quartiles are correspond to these point
quantile(rdata, probs = c(0.25, 0.5, 0.75))
```

```
##         25%         50%         75% 
## -0.70381756 -0.04016711  0.68695962
```

How many standard deviations are 1st, 2nd, and 3rd quartiles away from the mean

```
quantile(rdata, probs = 0.25)  - mean(rdata)/ sd(rdata)
```

```
##        25%
## -0.6883343
```

```
quantile(rdata, probs = 0.5) - mean(rdata) / sd(rdata)
```

```
##         50%
## -0.02468386
```

```
quantile(rdata, probs = 0.75) - mean(rdata) / sd(rdata)
```

```
##       75%
## 0.7024429
```

(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)

```
rdata2 <- rnorm(n=2000, mean=35, sd=3.5)
#how many standard deviations are they away from the mean
(quantile(rdata2, probs = 0.25) - mean(rdata2)) / sd(rdata2)
```

```
##        25%
## -0.6782319
```

```
(quantile(rdata2, probs = 0.5) - mean(rdata2))/ sd(rdata2)
```

```
##          50%
## -0.003597197
```

```
(quantile(rdata2, probs = 0.75) - mean(rdata2))/ sd(rdata2)
```

```
##       75%
## 0.6882169
```

```
#the 1st, 2nd, and 3rd quartile
quantile(rdata2, probs = 0.25)
```

```
##      25%
## 32.71492
```

```
quantile(rdata2, probs = 0.5)
```

```
##     50%
## 35.0834
```

```
quantile(rdata2, probs = 0.75)
```

```
##     75%
## 37.5122
```

Answer : We can see if we have a normal distribution with mean = 0 and standard devaition = 1, the 1st, 2nd, and 3rd quartiles position are correspond to there standard deviation distance away from their mean. But if we have a normal distribution with mean = 35 and standard devaition = 3.5,the 1st, 2nd, and 3rd quartiles position are not correspond to there standard deviation distance away from their mean but they are very close.

(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)

```
#how many standard deviations are they away from the mean
(quantile(d123, probs = 0.25) - mean(d123))/ sd(d123)
```

```
##        25%
## -0.5869297
```

```
(quantile(d123, probs = 0.5) - mean(d123))/ sd(d123)
```

```
##       50%
## 0.2633816
```

```
(quantile(d123, probs = 0.75) - mean(d123))/ sd(d123)
```

```
##       75%
## 0.7417857
```

```
#the 1st, 2nd, and 3rd quartile
quantile(d123, probs = 0.25)
```

```
##      25%
## 30.81035
```

```
quantile(d123, probs = 0.5)
```

```
##      50%
## 40.72216
```

```
quantile(d123, probs = 0.75)
```

```
##      75%
## 46.29877
```

Answer : We can see if we have a normal distribution with mean = 0 and standard devaition = 1, the 1st, 2nd, and 3rd quartiles position are correspond to there standard deviation distance away from their mean. But if we have a distribution d123 that its tail skew to the right with mean = 37.25264 and standard devaition = 11.5354 ,the 1st, 2nd, and 3rd quartiles position are not correspond to there standard deviation distance away from their mean, but there posotion are very far compared to quesion(b).

Q3

(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: Rob Hyndman suggest using “The Freedman-Diaconis rule”because it work very robust and works well in practice. The bin-width is set to h=2×IQR×n−1/3. So the number of bins is (max−min)/h, where n is the number of observations, max is the maximum value and min is the minimum value.

The benefit is that The Freedman-Diaconis rule is based on the interquartile range, denoted by IQR. It replaces 3.5σ of Scott’s rule with 2 IQR, so it make the method which is less sensitive than the standard deviation to outliers in data.

Note:refer to the above scott’s rule, Scott’s normal reference rule is optimal for random samples of normally distributed data, in the sense that it minimizes the integrated mean squared error of the density estimate.

(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

```
library(agricolae)
rand_data <- rnorm(800, mean=20, sd = 5)
#Sturges’ formula
Sturges <- sturges.freq(rand_data,k=0)
k <- Sturges$classes
h1 <- (Sturges$maximum - Sturges$minimum)/k
paste('the number of  bins is ', k )
```

```
## [1] "the number of  bins is  11"
```

```
paste('the bin widths is', h1 )
```

```
## [1] "the bin widths is 2.78109421942241"
```

ii. Scott’s normal reference rule (uses standard deviation) Given bin width : h=3.5σ^ / n^1/3

```
h2 <- (3.5*sd(rand_data))/(length(rand_data)^1/3)
k <- ceiling((max(rand_data) - min(rand_data))/h2)

paste('the number of  bins is ', k )
```

```
## [1] "the number of  bins is  452"
```

```
paste('the bin widths is', h2 )
```

```
## [1] "the bin widths is 0.0677850790776222"
```

iii. Freedman-Diaconis’ choice (uses IQR) Given bin width : h=2*IQR(x) / n ^1/3

```
h3 <- (2*IQR(rand_data))/(length(rand_data)^1/3)
k <- ceiling((max(rand_data) - min(rand_data))/h3)
paste("the number of bins is ",k)
```

```
## [1] "the number of bins is  591"
```

```
paste('the bin widths is', h3 )
```

```
## [1] "the bin widths is 0.0517870566334696"
```

(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?

i. Sturges’ formula

```
library(agricolae)
out_data <- c(rand_data, runif(10, min=40, max=60))
#Sturges’ formula
Sturges <- sturges.freq(out_data,k=0)
k <- Sturges$classes
h4 <- (Sturges$maximum - Sturges$minimum)/k
paste('the number of  bins is ', k )
```

```
## [1] "the number of  bins is  11"
```

```
paste('the bin widths is', h4 )
```

```
## [1] "the bin widths is 5.03867233169183"
```

ii. Scott’s normal reference rule (uses standard deviation) Given bin width : h=3.5σ^ / n^1/3

```
h5 <- (3.5*sd(out_data))/(length(out_data)^1/3)
k <- ceiling((max(out_data) - min(out_data))/h5)

paste('the number of  bins is ', k )
```

```
## [1] "the number of  bins is  691"
```

```
paste('the bin widths is', h5 )
```

```
## [1] "the bin widths is 0.0802365936085188"
```

iii. Freedman-Diaconis’ choice (uses IQR) Given bin width : h=2*IQR(x) / n ^1/3

```
h6 <- (2*IQR(out_data))/(length(out_data)^1/3)
k <- ceiling((max(out_data) - min(out_data))/h6)
paste("the number of bins is ",k)
```

```
## [1] "the number of bins is  1068"
```

```
paste('the bin widths is', h6 )
```

```
## [1] "the bin widths is 0.0519328308132877"
```

Calculate the bin width change when adding outliers of each method use

```
paste('the change bin width using Scott method ', (h1-h4)/h1 )
```

```
## [1] "the change bin width using Scott method  -0.811758945994388"
```

```
paste('the change bin width using Sturge method ', (h2-h5)/h2 )
```

```
## [1] "the change bin width using Sturge method  -0.1836910821722"
```

```
paste('the change bin width using Freedman-Diaconis method ', (h3-h6)/h3 )
```

```
## [1] "the change bin width using Freedman-Diaconis method  -0.00281487671427065"
```

From the above result we can conclude that using Freedman–Diaconis method is the least sensitive to change binwidth when there are outliers.

It is cause by Freedman–Diaconis method use IQR to calculate bin width which is less affect by outlier. Sturge method, Scott method will be affect by sd(sigma hat (σ)) and the number of data(n) when dealing with ouliers.