HW3 (week3)

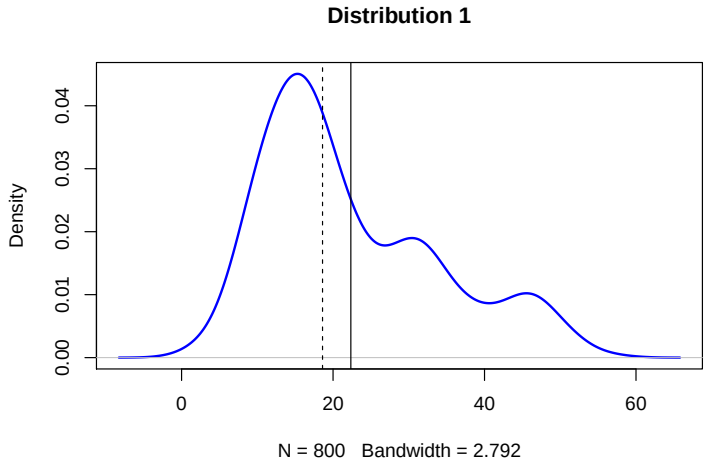
StudentID:106022103

2021/3/14

Question 1

Example

```
# 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)

# Let's combine them into a single dataset
d123 <- c(d1, d2, d3)

# Let's plot the density function of abc
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")
```



useful function

To simplify the code to be writing, we can create a function to show the plot and

```
dist_show <- function(dist,number=1,plot_mean=TRUE,plot_median =TRUE){
  plot(density(dist), col="blue", lwd=2,
```

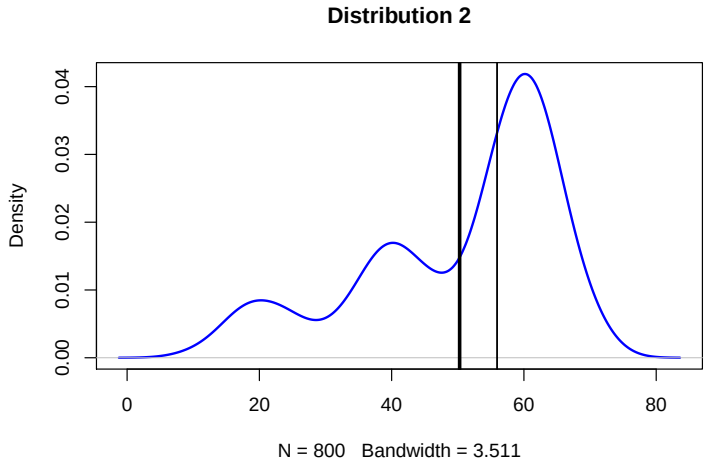
```
       main = paste("Distribution",number))
  if(plot_mean){
    abline(v=mean(dist),lwd=3)
  }
  if(plot_median){
    abline(v=median(dist),lwd=1.5)
  }
}
```

```
dist_print <- function(dist){
  print(sprintf("Mean: %.3f",mean(dist)))
  print(sprintf("Median: %.3f",median(dist)))
}
```

(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).

```
# Change the mean and standard deviation
d1 <- rnorm(n=500, mean=60, sd=5)
d2 <- rnorm(n=200, mean=40, sd=5)
d3 <- rnorm(n=100, mean=20, sd=5)

# Let's combine them into a single dataset
dist2 <- c(d1, d2, d3)

# Show the results
dist_show(dist2,2)
```

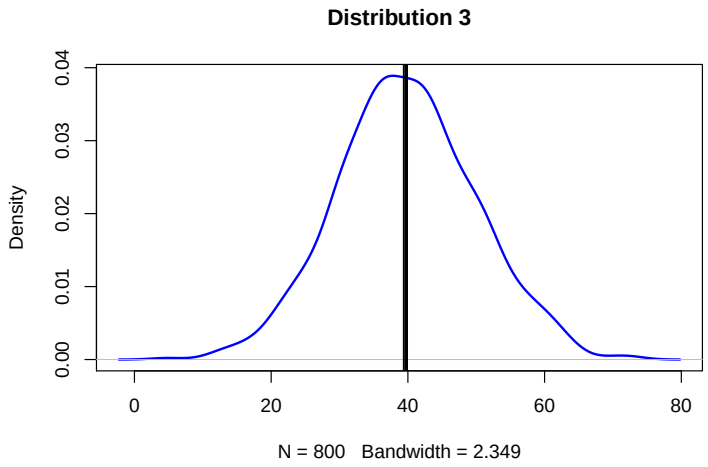



```
dist_print(dist2)
```

```
## [1] "Mean: 50.279"
## [1] "Median: 55.939"
```

(b) Create a "Distribution 3":

```
dist3 <- rnorm(800,mean=40,sd=10)
dist_show(dist3,3)
```




```
dist_print(dist3)
```

```
## [1] "Mean: 39.755"
## [1] "Median: 39.396"
```

(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?

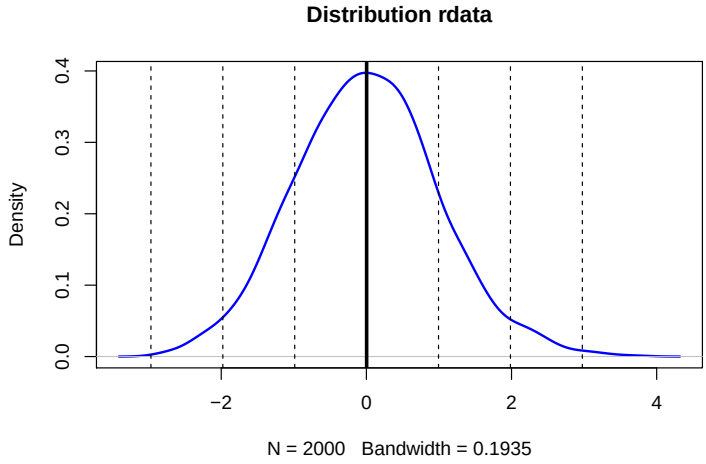
ANSWER: I think the **mean** of data is more sensitive in general.
Most of time, the outliers are few and far from center, so the mean is more sensitive.
In a few cases, when the outliers are many but not very far from the center (near the threshold of the outliers), then the median may be more sensitive.

Question 2

(a) plot 7 lines

> 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)
dist_show(rdata,"rdata",plot_mean=FALSE,plot_median=FALSE)

# create a list with the dash line
sd_list <- c(-3:3) * sd(rdata) + mean(rdata)
# use `lapply` just learned this week to plot  it
lapply(sd_list, function(x){abline(v=mean(x),lty="dashed")})

# plot the mean line
abline(v=mean(rdata),lwd=3)
```



(b) How far quantiles from mean?

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

```
distance_quantile <- function(data,th){
  distance <- (quantile(data,th/100) - mean(data)) / sd(data)
  print(sprintf("The %dth quentile is %.4f (sd) away from the mean.",th,distance))
  distance
}

lapply(c(25,50,75),function(x){distance_quantile(rdata,th=x)})
```

(c) How far quantiles from mean?

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?

```
rdata2 <- rnorm(n=2000, mean=35, sd=3.5)
lapply(c(25,75),function(x){distance_quantile(rdata2,th=x)})
```

Compare your answer to (b)

ANSWER: The 1st and 3rd quartiles of (c) are similar to the results of (b) because the distances calculated here are divided by sd.

(d) How far quantiles from mean?

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?

```
lapply(c(25,75),function(x){distance_quantile(d123,th=x)})
```

```
## [1] "The 25th quentile is -0.7383 (sd) away from the mean."
## [1] "The 75th quentile is 0.6631 (sd) away from the mean."
```

```
## [[1]]
##        25%
## -0.7382531
##
## [[2]]
##       75%
## 0.6631089
```

Compare your answer to (b)

ANSWER: The first quartile of (d) is farther away and the third quartile is closer than in (b). The reason is that `d123` is a left-biased distribution causing more data to be concentrated on the right side.

Question 3

We mentioned in class that there might be some objective ways of determining the bin size of histograms. Take a quick look at the Wikipedia article on Histograms ("Number of bins and width") to see the different ways to calculate bin width (h) and number of bins (k).
Note that, for any dataset d, we can calculate number of bins (k) from the bin width (h):
$k = ceiling((max(d) - min(d))/h)$
and bin width from number of bins:
$h = (max(d) - min(d))/k$
Now, read the following discussion on the Q&A forum called "Cross Validated" about choosing the number of bins

(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: For the bin-width is $h = 2 \times IQR \times n^{-1/3}$, so there are $\frac{max-min}{h}$ bins. (n is the number of obervations,max is the maximum value and min is the minimum value.)
The benifit is less sensitive than the standard deviation to outliers in data, because it replaces 3.5σ of Scott's rule with $2IQR$,

(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)

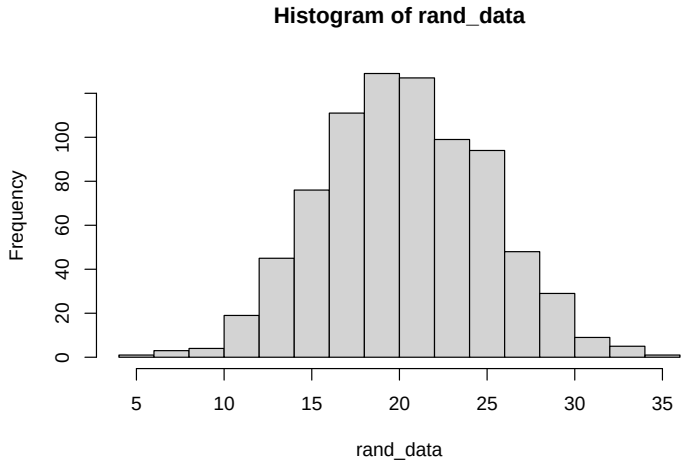
```r
rand_data <- rnorm(800, mean=20, sd = 5)

band <- function(data,method){
  nb <- method(data) # number of bands
  bw <- (max(data)-min(data)) / nb # band-width
  print(sprintf("There are %d of bands",nb))
  print(sprintf("The band-width is %.3f",bw))

  if (identical(method,nclass.Sturges)){
    hist(rand_data,breaks="Sturges")
  }

  if (identical(method,nclass.scott)){
    hist(rand_data,breaks="Scott")
```

```
  }

  if (identical(method,nclass.FD)){
    hist(rand_data,breaks="FD")
  }

}
```

```
band(rand_data,nclass.Sturges)
```

i. Sturges formula

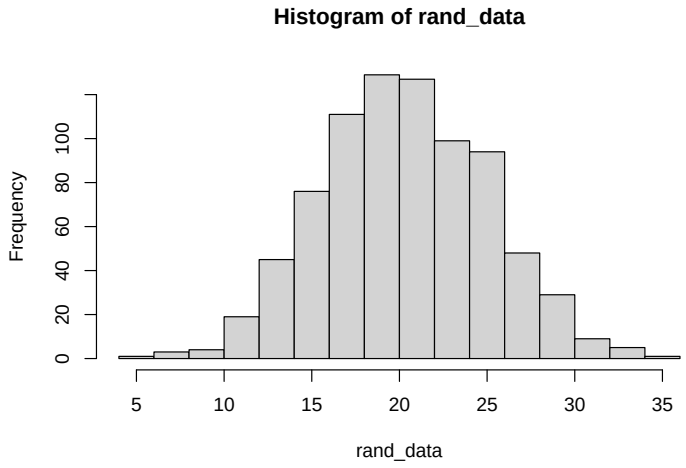
```
## [1] "There are 11 of bands"
## [1] "The band-width is 2.636"
```




```
band(rand_data,nclass.scott)
```

ii. Scotts normal reference rule (uses standard deviation)

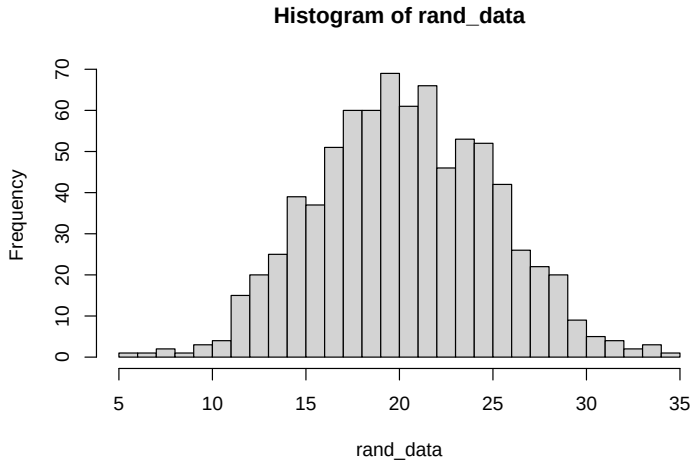
```
## [1] "There are 17 of bands"
## [1] "The band-width is 1.705"
```




```
band(rand_data,nclass.FD)
```

iii. Freedman-Diaconis choice (uses IQR)

```
## [1] "There are 21 of bands"
## [1] "The band-width is 1.381"
```



(c)

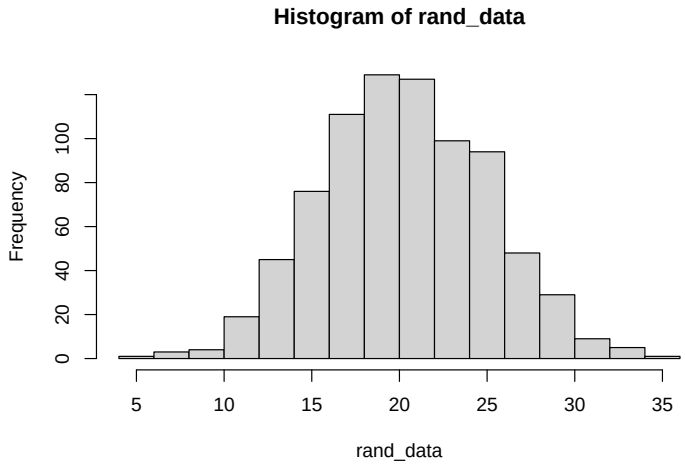
Repeat part (b) but extend the rand_data dataset with some outliers (use a new dataset out_data): `out_data <- c(rand_data, runif(10, min=40, max=60))`

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

```
band(out_data,nclass.Sturges)
```

i. Sturges formula

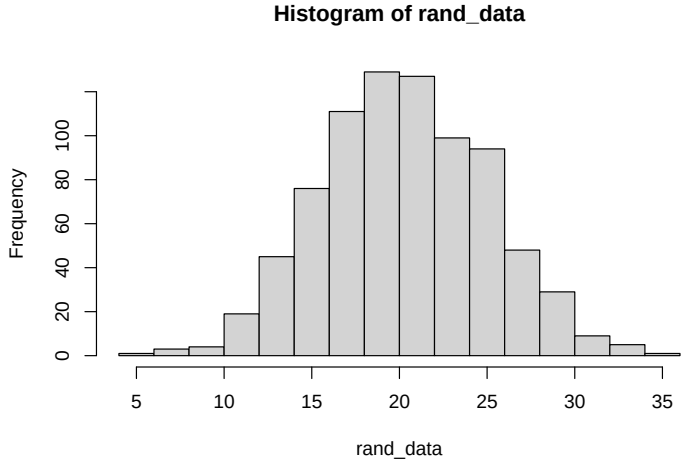
```
## [1] "There are 11 of bands"
## [1] "The band-width is 4.910"
```




```
band(out_data,nclass.scott)
```

ii. Scotts normal reference rule (uses standard deviation)

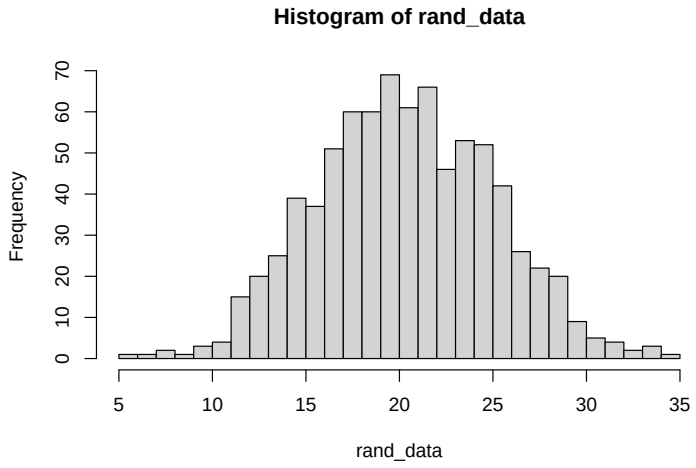
```
## [1] "There are 26 of bands"
## [1] "The band-width is 2.077"
```




```
band(out_data,nclass.FD)
```

iii. Freedman-Diaconis choice (uses IQR)

```
## [1] "There are 38 of bands"
## [1] "The band-width is 1.421"
```



(d)

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?

ANSWER: Freedman-Diaconis choice(FD) because it is much less affected by outliers.

References

- nclass
- sprintf not ouput?
- Skewness
- 106070038
  - part of Question 3
- R Markdown