hw3

109078501

2021/3/9

Help by 109078507, 109078513, 109078519

Question 1) Let's have a look at how the mean and median behave. 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!

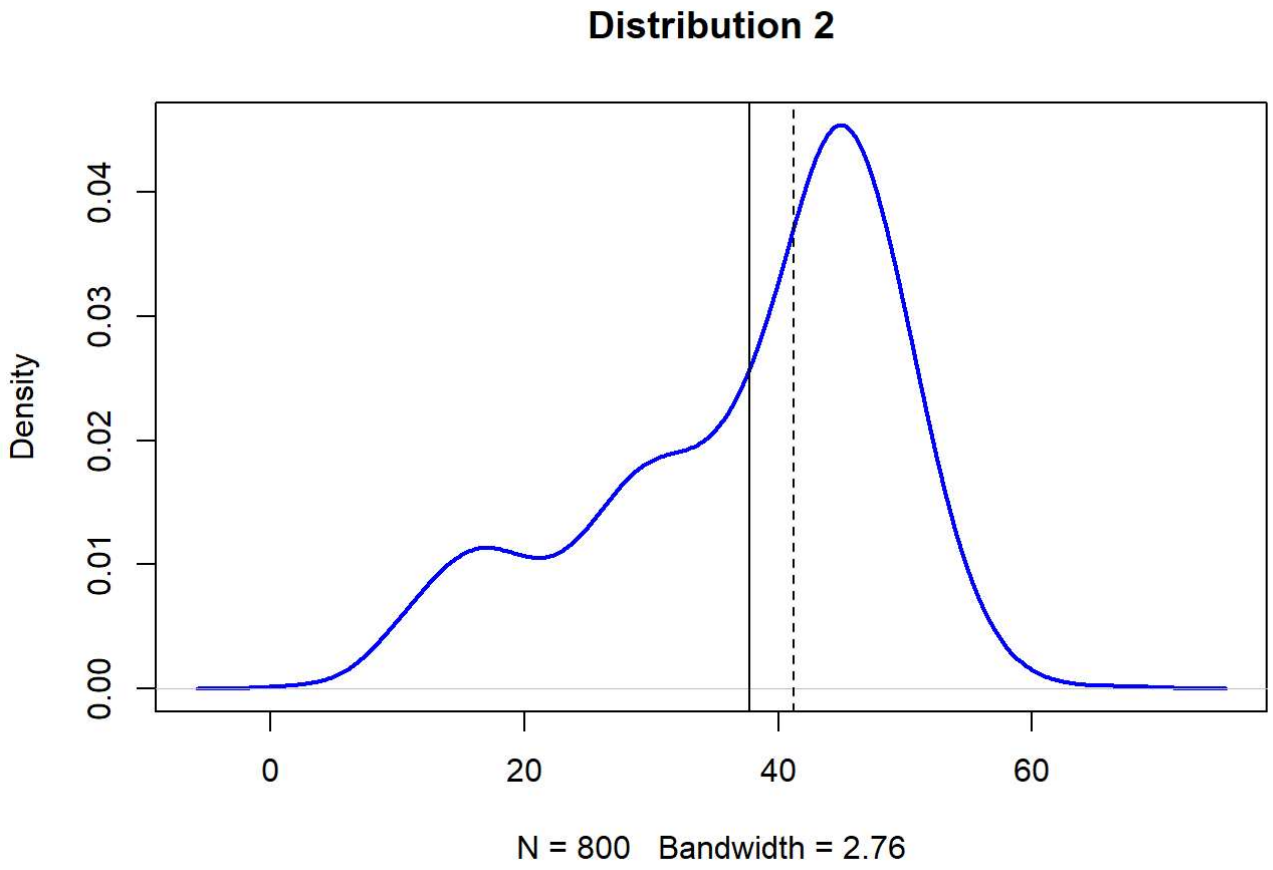
Now, try to to match the following distributions. You can reuse and modify the code above.

(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).

```
#construct 3 normal distribution data set
d1<-rnorm(n = 500,mean = 45,sd = 5)
d2<-rnorm(n = 200,mean = 30,sd = 5)
d3<-rnorm(n = 100,mean = 15,sd = 5)
#combine them into single data set
d123<-c(d1,d2,d3)
#compute the mean and median
my_mean <- mean(d123)
my_median <- median(d123)
paste("The mean is", my_mean,"The median is", my_median)
```

```
## [1] "The mean is 37.7203312701946 The median is 41.2327590282397"
```

```
#plot
plot(density(d123), col="blue", lwd=2, main = "Distribution 2")
abline(v=mean(d123))
abline(v=median(d123), lty="dashed")
```

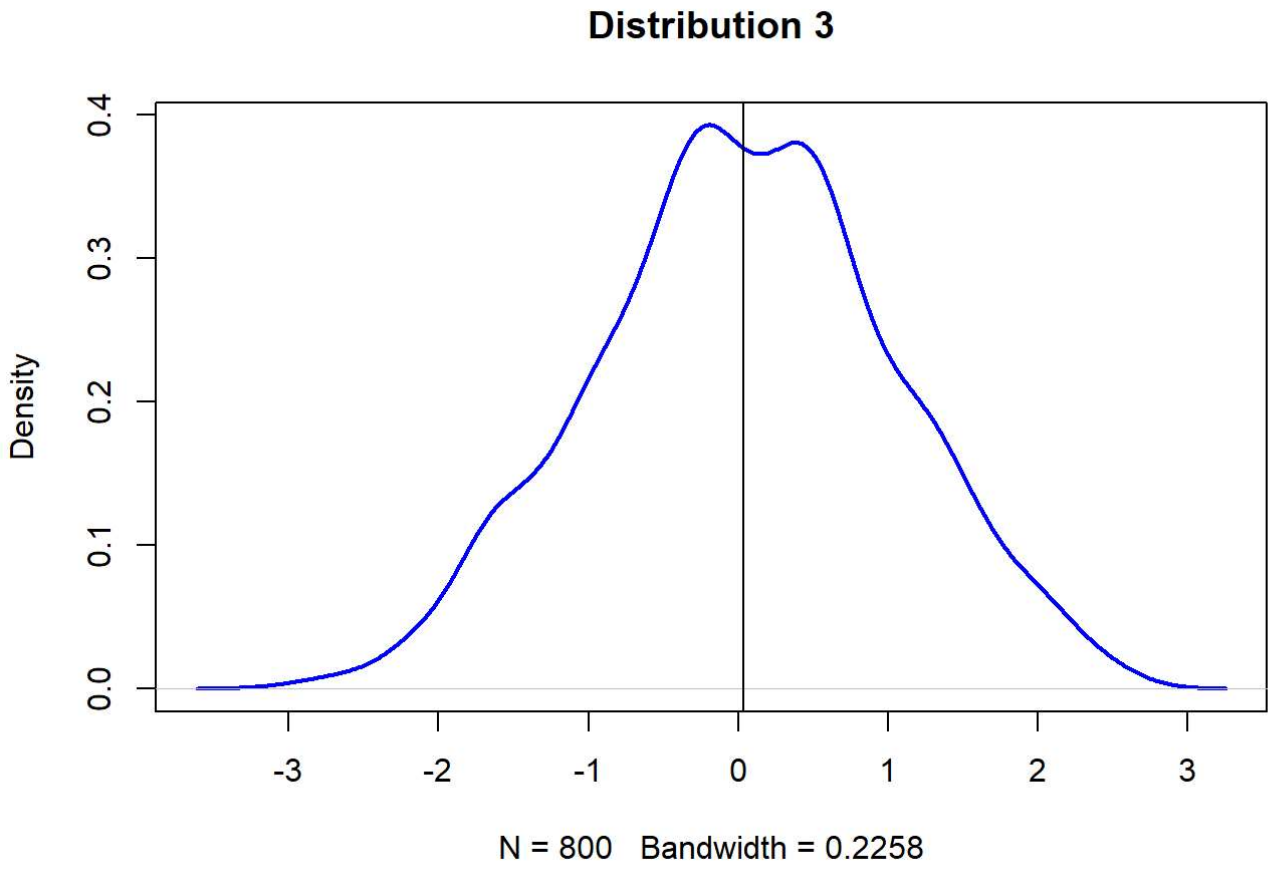


(b) Create a "Distribution 3": a single dataset that is normally distributed (bell-shaped, symmetric) – you do not need to combine datasets, just use the rnorm function to create a single large dataset (n=800). Show your code, compute the mean and median, and draw lines showing the mean (thick line) and median (thin line).

```
#construct a normal distribution
d<-rnorm(n = 800)
#compute the mean and median
my_mean <- mean(d)
my_median <- median(d)
paste("The mean is", my_mean,"The median is", my_median)
```

```
## [1] "The mean is 0.0374133693315439 The median is 0.0332393562379957"
```

```
#plot
plot(density(d), main="Distribution 3", col="blue", lwd=2)
#mean and median is overlapped.
abline(v=mean(d))
abline(v=median(d),lty="dashed")
```



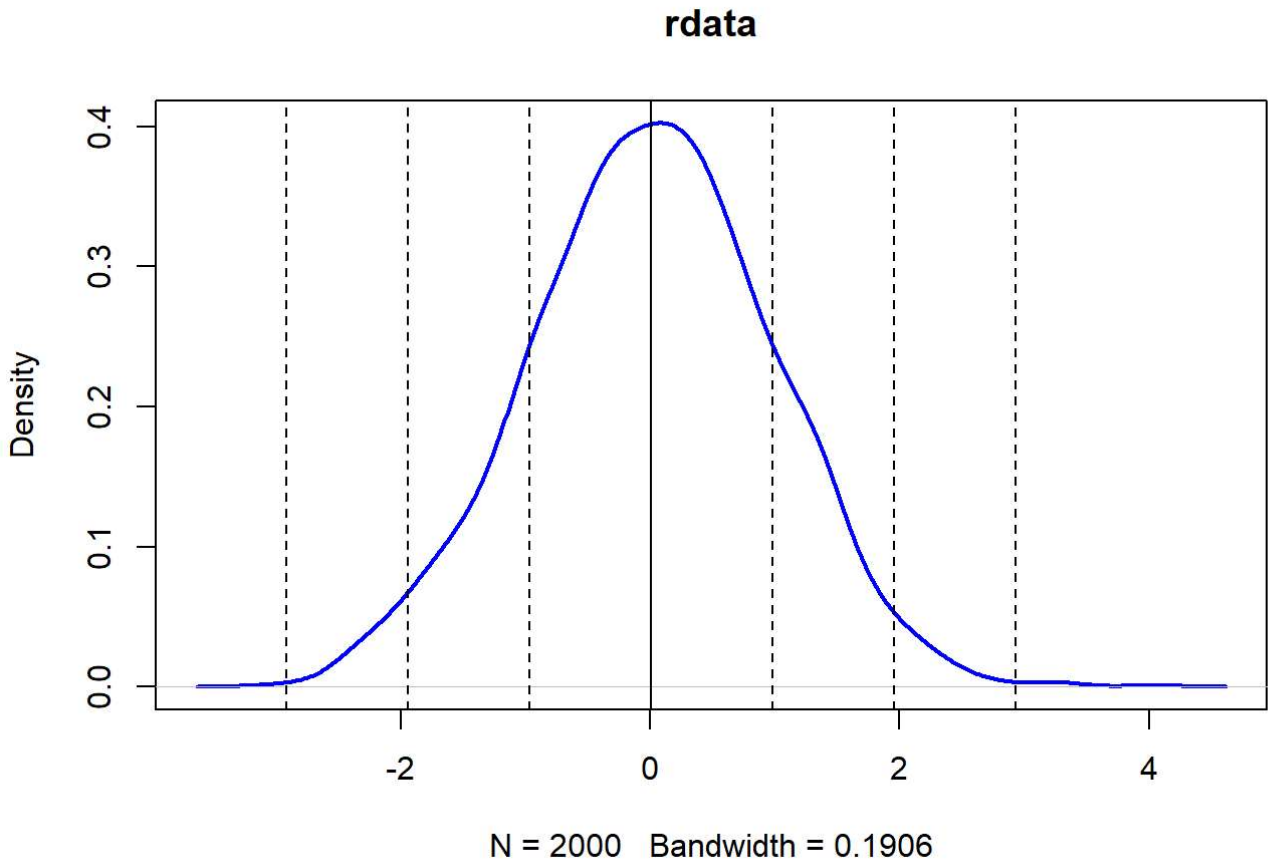
(c) In general, which measure of central tendency (mean or median) do you think will be more sensitive (will change more) to outliers being added to your data?

I think mean is more sensitive to outliers. Outliers are numbers in a data set that are vastly larger or smaller than the other values in the set. Mean, median and mode are measures of central tendency. When we calculate mean, we have to sum up all the values and divide by length of dataset. Mean is the only measure of central tendency that is always affected by an outlier. If we add outliers to the dataset, the data will change the order in the dataset. Due to median standing for middle point of the data, if the order of the data changed, the median will change a little.

Question 2) Let's try to get some more insight about what standard deviations are.

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).

```
#construct a noraml distribution with mean=0 sd=1
rdata<-rnorm(n = 2000,mean = 0,sd = 1)
#create a sequence from -3 to 3
grid<-seq(from=-3,to = 3)
#the cutting line which contains the 1st,2nd,3rd standard deviations and the mean(from left t
o right ,7 points totally.)
lines<-mean(rdata)+grid*sd(rdata)
#plot
plot(density(rdata),main="rdata", col="blue", lwd = 2)
#lty=2 means dashed line,1 means solid line
abline(v=lines,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)? 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?

```
#calculate quantile (25th,50th,75th percentiles)
rdata.quantile<-quantile(x = rdata, probs = c(.25,.5,.75))
paste("The data points of 1st, 2nd and 3rd quartiles are:", rdata.quantile[1], rdata.quantile
[2], "and", rdata.quantile[3])
```

```
## [1] "The data points of 1st, 2nd and 3rd quartiles are: -0.640206575427653 0.0104187710376
267 and 0.657349098739422"
```

```
#calculate how many sd away from the mean
dist<-unname((rdata.quantile-mean(rdata))/sd(rdata))
paste("The distance of 1st, 2nd, and 3rd quartiles are", dist[1], dist[2], dist[3], "standard
deviations away from the mean.")
```

```
## [1] "The distance of 1st, 2nd, and 3rd quartiles are -0.661412779658191 0.0058464538708641
2 0.669316201789818 standard deviations away from the mean."
```

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)
#calculate quantile (25th,50th,75th percentiles)
rdata.new.quantile<-quantile(x = rdata.new,probs = c(.25,.5,.75))
#calculate how many sd away from the mean (called dist)
dist<-unname((rdata.new.quantile-mean(rdata.new))/sd(rdata.new))

paste("The distance of 1st, and 3rd quartiles are", dist[1], dist[3], "standard deviations aw
ay from the mean.")
```

```
## [1] "The distance of 1st, and 3rd quartiles are -0.695311524202633 0.658528701481384 stand
ard deviations away from the mean."
```

```
paste("Due to normal distributions ,there are similar between (b) and (c).")
```

```
## [1] "Due to normal distributions ,there are similar between (b) and (c)."
```

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)

```
#Distribution 1
#construct 3 normal distribution data set
d1<-rnorm(n = 500,mean = 15,sd = 5)
d2<-rnorm(n = 200,mean = 30,sd = 5)
d3<-rnorm(n = 100,mean = 45,sd = 5)
#combine them into single data set
d123<-c(d1,d2,d3)
#calculate the quantile
d123.quantile<-quantile(x = d123,probs = c(.25,.75))
#substract the mean and see how many sd away from the mean
dist_d123<-unname((d123.quantile-mean(d123))/sd(d123))

paste("The distance of 1st, and 3rd quartiles are", dist[1], dist[3], "standard deviations aw
ay from the mean.")
```

```
## [1] "The distance of 1st, and 3rd quartiles are -0.695311524202633 0.658528701481384 stand
ard deviations away from the mean."
```

```
paste("Because (d) is not a normal distribution, there are more differences between (b) and
 (d).")
```

```
## [1] "Because (d) is not a normal distribution, there are more differences between (b) and
(d)."
```

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?

Rob Hyndman suggested us to use Freedman-Diaconis' choice, and the benefit of this formula 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

```
rand_data <- rnorm(800, mean=20, sd = 5)
k = ceiling(log2(length(rand_data)))+1
h = (max(rand_data) - min(rand_data)) / k
paste("The number of bins are", k, "and the bin widths are", h)
```

```
## [1] "The number of bins are 11 and the bin widths are 2.81636853924947"
```

ii. Scott's normal reference rule (uses standard deviation)

```
# sd
scott_sd = sd(rand_data)
h = 3.5*scott_sd / (length(rand_data)^(1/3))
k = ceiling((max(rand_data) - min(rand_data)) / h)
paste("The number of bins are", k, "and the bin widths are", h)
```

```
## [1] "The number of bins are 17 and the bin widths are 1.88574349874969"
```

iii. Freedman-Diaconis' choice (uses IQR)

```
h = 2 * IQR(rand_data) / (length(rand_data)^(1/3))
k = ceiling((max(rand_data) - min(rand_data)) / h)
paste("The number of bins are", k, "and the bin widths are", h)
```

```
## [1] "The number of bins are 23 and the bin widths are 1.40378441277714"
```

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))

i. Sturges' formula

```
out_data <- c(rand_data, runif(10, min=40, max=60))
k = ceiling(log(length(out_data), 2)) + 1
h = (max(out_data) - min(out_data)) / k
paste("The number of bins are", k, "and the bin widths are", h)
```

```
## [1] "The number of bins are 11 and the bin widths are 5.05827726616003"
```

ii. Scott's normal reference rule (uses standard deviation)

```
# sd
scott_sd = sd(out_data)
h = 3.5*scott_sd / (length(out_data)^(1/3))
k = ceiling((max(out_data) - min(out_data)) / h)
paste("The number of bins are", k, "and the bin widths are", h)
```

```
## [1] "The number of bins are 25 and the bin widths are 2.30466441530265"
```

iii. Freedman-Diaconis' choice (uses IQR)

```
h = 2 * IQR(out_data) / (length(out_data)^(1/3))
k = ceiling((max(out_data) - min(out_data)) / h)
paste("The number of bins are", k, "and the bin widths are", h)
```

```
## [1] "The number of bins are 39 and the bin widths are 1.43084016359895"
```

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?

Sturges' formula and Scott's normal reference rule don't exclude outliers, so they are sensitive to outliers. Freedman-Diaconis's choice is least sensitive to outliers because it use the interquartile range (IQR), which don't take very much care of the outliers.