BACS_HW2

109090046

2023-03-04

Question 1)

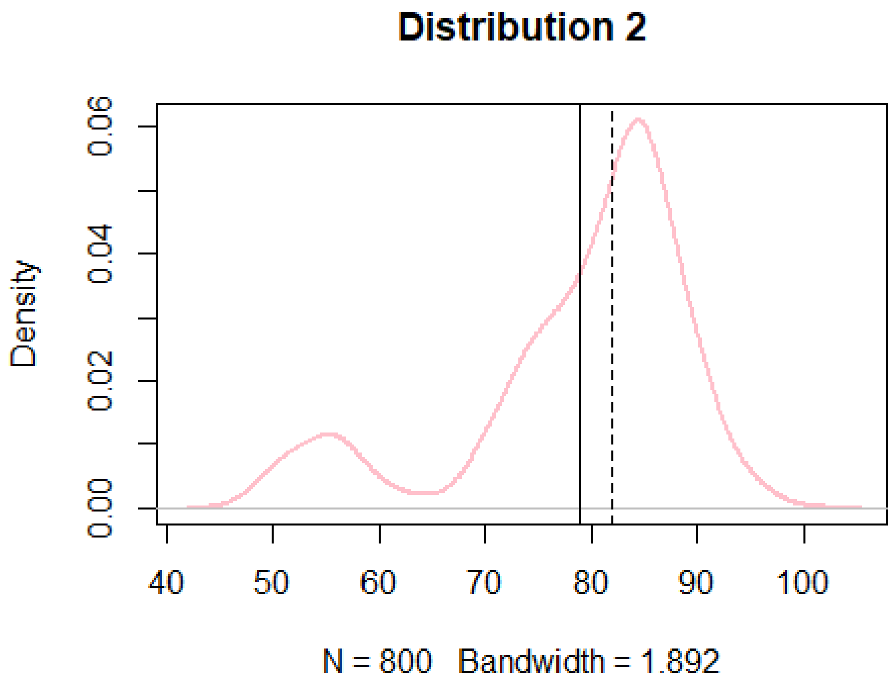
(a) Create and visualize a new "**Distribution 2**": a combined dataset (n=800) that is negatively skewed (tail stretches to the left). Change the mean and standard deviation of d1, d2, and d3 to achieve this new distribution. Compute the mean and median, and draw lines showing the mean (thick line) and median (thin line).

```
# Three normally distributed data sets
d1 <- rnorm(n=500, mean=85, sd=4)
d2 <- rnorm(n=200, mean=75, sd=4)
d3 <- rnorm(n=100, mean=55, sd=4)

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

# Plot the density function of d123
plot(density(d123), col="pink", lwd=2,
     main = "Distribution 2")

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

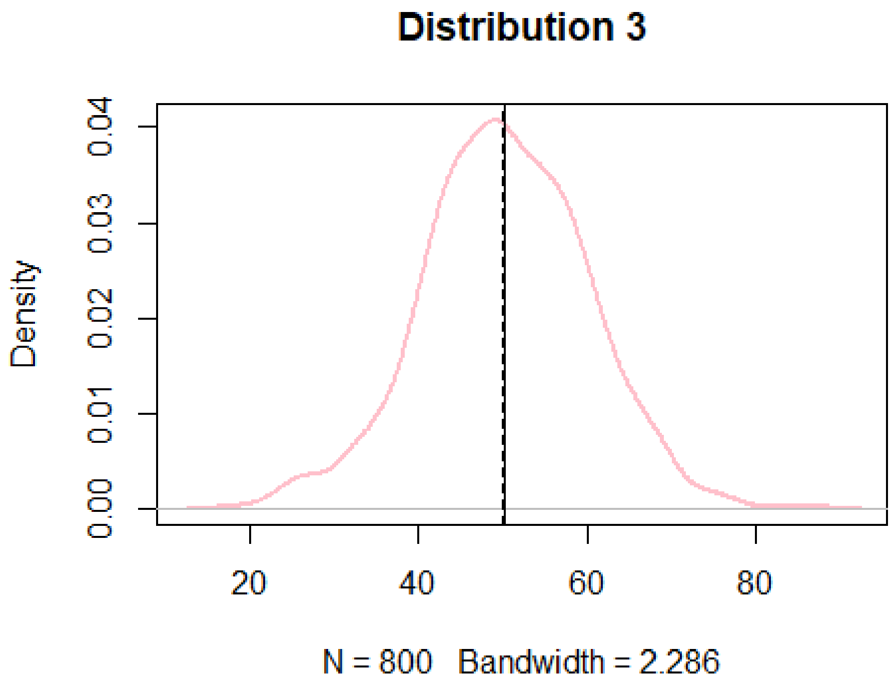


(b) Create a "**Distribution 3**": a single dataset that is normally distributed (bell-shaped, symmetric) – you do not need to combine datasets, just use the rnorm() function to create a single large dataset (n=800). Show your code, compute the mean and median, and draw lines showing the mean (thick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

```
bell_shaped <- rnorm(n=800, mean=50, sd=10)

# Plot the density function
plot(density(bell_shaped), col="pink", lwd=2,
     main = "Distribution 3")

# Add vertical lines showing mean and median
abline(v=mean(bell_shaped))
abline(v=median(bell_shaped), 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?

```
# Mean will be more sensitive to outliers being added to data because i
f the outliers are increadibly big, they may change the mean but the me
dian.
```

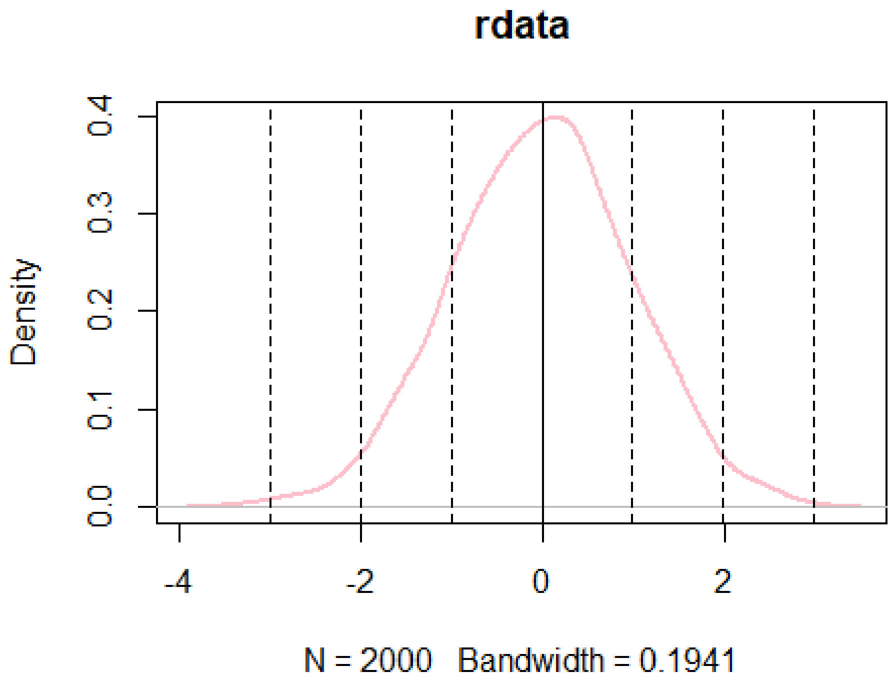
Question 2)

(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 the density function
plot(density(rdata), col="pink", lwd=2,
     main = "rdata")

# Add vertical lines showing mean and the 1st, 2nd, and 3rd standard de
viations
abline(v=mean(rdata))
for (i in 3:-3){
  abline(v=mean(rdata+i), 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?

```
q1 <- quantile(rdata, 0.25)
q2 <- quantile(rdata, 0.5)
q3 <- quantile(rdata, 0.75)

z1 <- (q1 - mean(rdata)) / sd(rdata)
z2 <- (q2 - mean(rdata)) / sd(rdata)
z3 <- (q3 - mean(rdata)) / sd(rdata)

## 1st quartile is -0.67 which is -0.67 standard deviations away from t
he mean.

## 2nd quartile is 0 which is 0.01 standard deviations away from the me
an.

## 3rd quartile is 0.66 which is 0.68 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_2 <- rnorm(n=2000, mean=35, sd=3.5)

q1 <- quantile(rdata_2, 0.25)
q2 <- quantile(rdata_2, 0.5)
q3 <- quantile(rdata_2, 0.75)

z1 <- (q1 - mean(rdata_2)) / sd(rdata_2)
z2 <- (q2 - mean(rdata_2)) / sd(rdata_2)
z3 <- (q3 - mean(rdata_2)) / sd(rdata_2)
```

```
## 1st quartile is 32.7 which is -0.65 standard deviations away from th
e mean.

## 2nd quartile is 34.97 which is 0 standard deviations away from the m
ean.

## 3rd quartile is 37.37 which is 0.69 standard deviations away from th
e mean.
```

The answer of standard deviations away from the mean are very close but not the same.

(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)

```
d1 <- rnorm(n=500, mean=15, sd=5)
d2 <- rnorm(n=200, mean=30, sd=5)
d3 <- rnorm(n=100, mean=45, sd=5)
d123 <- c(d1, d2, d3)

q1 <- quantile(d123, 0.25)
q3 <- quantile(d123, 0.75)

z1 <- (q1 - mean(d123)) / sd(d123)
z3 <- (q3 - mean(d123)) / sd(d123)
```

```
## 1st quartile is 14.34 which is -0.72 standard deviations away from t
he mean.

## 3rd quartile is 30.61 which is 0.69 standard deviations away from th
e mean.
```

The answer are close, but not that close compare to (c). I think it's because of the large standard deviation of the d123 (which is 5).

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?

```
# Freedman-Diaconis' choice is that formula that the author suggests us
 to use.
# The benefit of the formula is that it is less sensitive to outliers i
n 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)

# Sturges' formula
k1 <- log(length(rand_data), 2) + 1
h1 <- (max(rand_data) - min(rand_data)) / k1

# Scott's normal reference rule
h2 <- 3.49 * sd(rand_data) / (length(rand_data) ^ (1/3))
k2 <- ceiling(max(rand_data) - min(rand_data) / h2)

# Freedman-Diaconis' choice
h3 <- 2 * IQR(rand_data) / (length(rand_data) ^ (1/3))
k3 <- ceiling(max(rand_data) - min(rand_data) / h3)

## Using Sturges' formula, the result bin widths (h) is 3.095042 and nu
mber of bins (k) is 10.64386 .

## Using Scott's normal reference rule, the result bin widths (h) is 1.
899665 and number of bins (k) is 34 .

## Using Freedman-Diaconis' choice, the result bin widths (h) is 1.4755
28 and number of bins (k) is 34 .
```

(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?

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

# Sturges' formula
k1 <- log(length(out_data), 2) + 1
h1 <- (max(out_data) - min(out_data)) / k1
```

```
# Scott's normal reference rule
h2 <- 3.49 * sd(out_data) / (length(out_data) ^ (1/3))
k2 <- ceiling(max(out_data) - min(out_data) / h2)

# Freedman-Diaconis’ choice
h3 <- 2 * IQR(out_data) / (length(out_data) ^ (1/3))
k3 <- ceiling(max(out_data) - min(out_data) / h3)
```

```
## Using Sturges’ formula, the result bin widths (h) is 5.54079 and num
ber of bins (k) is 10.66178 .
```

```
## Using Scott's normal reference rule, the result bin widths (h) is 2.
275914 and number of bins (k) is 60 .
```

```
## Using Freedman-Diaconis’ choice, the result bin widths (h) is 1.4903
94 and number of bins (k) is 60 .
```

Freedman-Diaconis’ choice method’s bin widths changes the least. By using the IQR instead of the full range or standard deviation, Freedman-Diaconis’ rule is able to avoid the influence from the outliers.