BACS HW2

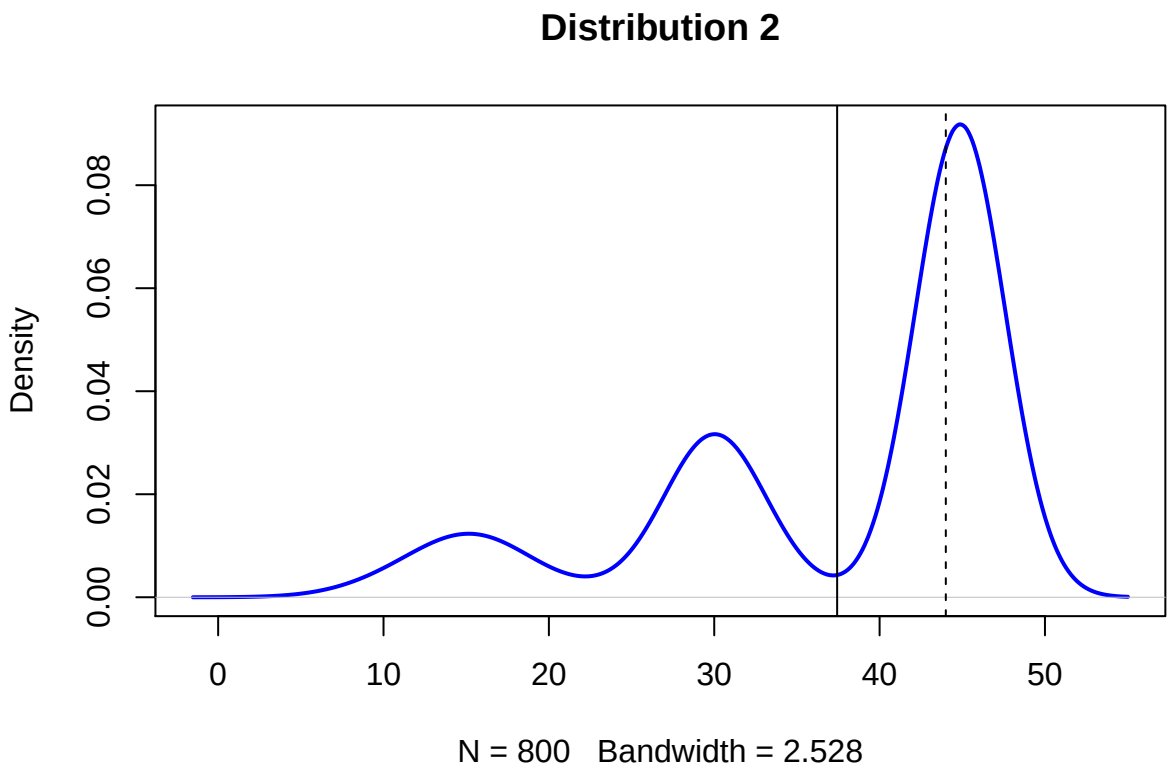
109062710

March 14, 2021

Question 1

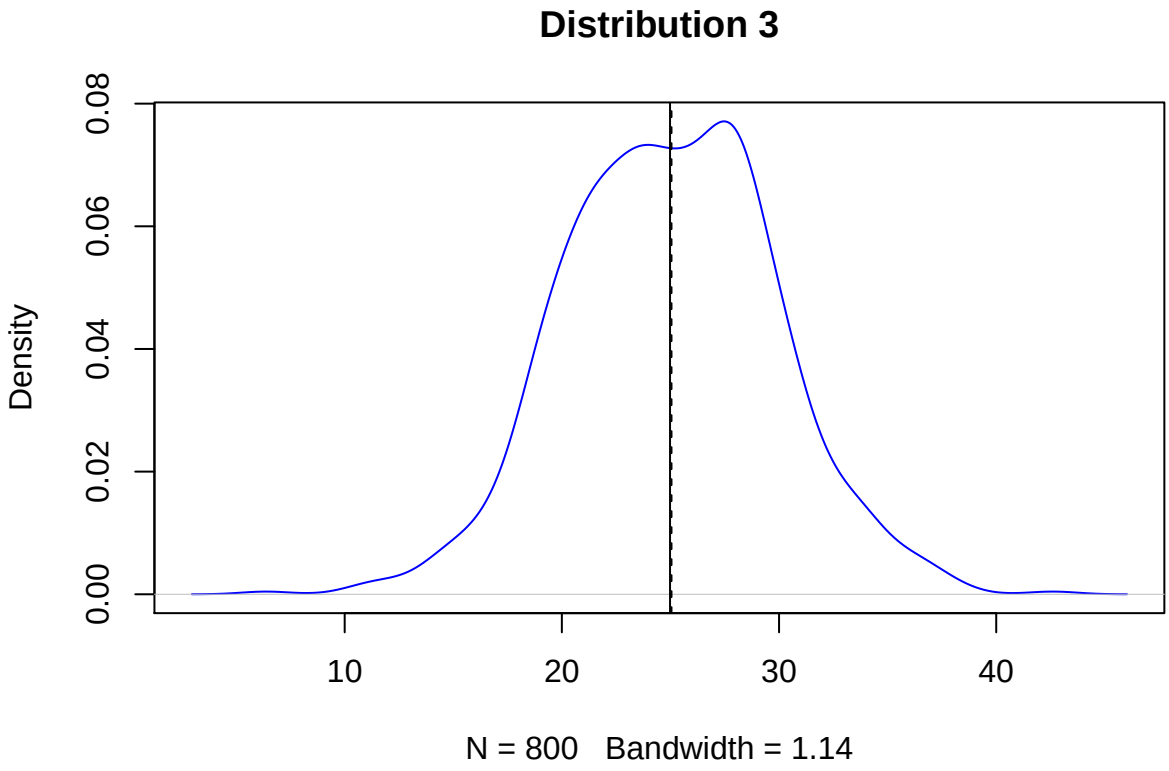
(a) A graph with negatively skewed tail

```
d1 <- rnorm(n=500, mean=45, sd=1)
d2 <- rnorm(n=200, mean=30, sd=2)
d3 <- rnorm(n=100, mean=15, sd=3)

dataset <- c(d1, d2, d3)
plot(density(dataset), col="blue", lwd=2, main="Distribution 2")

abline(v=mean(dataset))
abline(v=median(dataset), lty="dashed")
```



(b) A graph with bell shaped curve

```
dataset <- c(rnorm(n=800, mean=25,sd=5))

plot(density(dataset), col="blue", main="Distribution 3")

abline(v=mean(dataset))
abline(v=median(dataset), lty="dashed")
```



(c) Which of the central measurements is more likely to be affeccted by outliers in the data? Is it Mean or Median?

Mean will more likely to be affected by outliers in the data. Mean will increase because of the existence of higher outliers. Conversely, Mean will decrease because of the existence of lower outliers.

Question 2

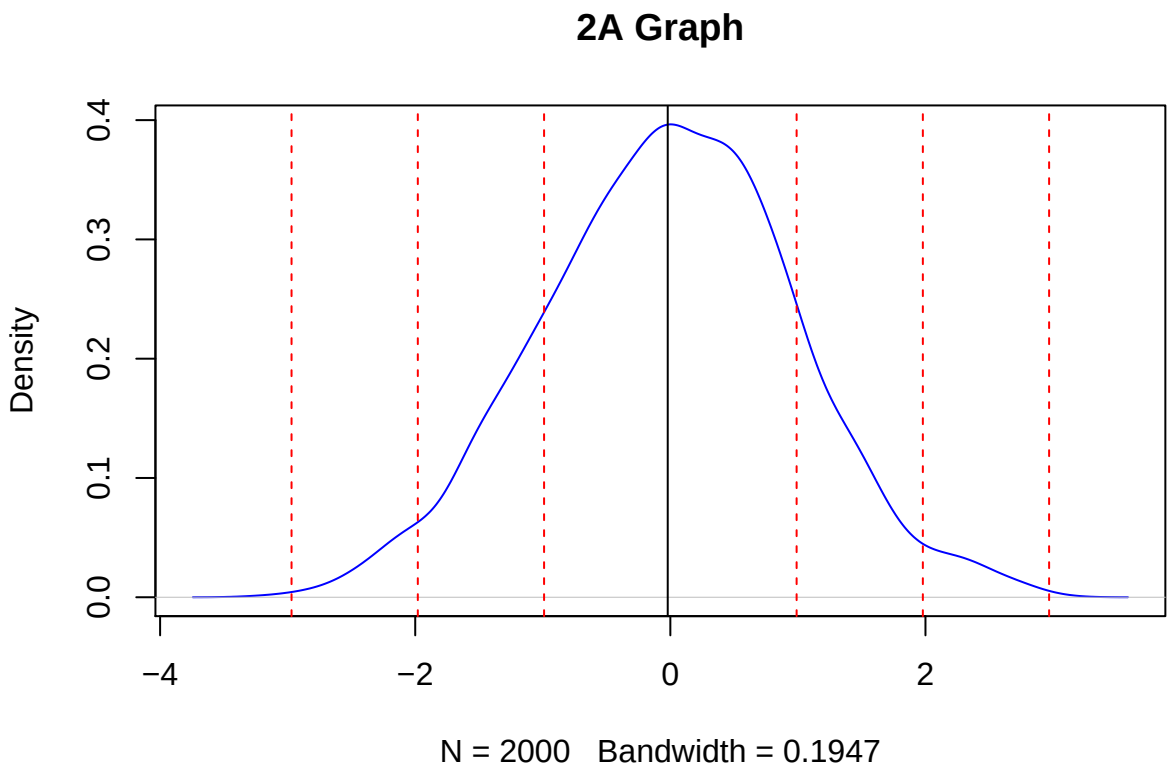
Below is the helper functions that are used in Q2

```
mean_std_lines <- function (dataset) {
  abline(v=mean(dataset))

  # negative Standard Deviation lines
  abline(v=sd(dataset) * -1, col="red", lty="dashed")
  abline(v=sd(dataset) * -2, col="red", lty="dashed")
  abline(v=sd(dataset) * -3, col="red", lty="dashed")

  # positive Standard Deviation lines
  abline(v=sd(dataset), col="red", lty="dashed")
  abline(v=sd(dataset) * 2, col="red", lty="dashed")
  abline(v=sd(dataset) * 3, col="red", lty="dashed")
}
```

```r
quartiles_lines <- function (dataset) {
  abline(v=quantile(dataset, 0.25), col="green", lty="dashed")
  abline(v=quantile(dataset, 0.50), col="green", lty="dashed")
  abline(v=quantile(dataset, 0.75), col="green", lty="dashed")
}

quartile_distance <- function(dataset, nth) {
  return(quantile(dataset, nth) - mean(dataset) / sd(dataset))
}

quartiles_distance <- function(dataset) {
  first_quantile_distance = quartile_distance(dataset, 0.25)
  second_quantile_distance = quartile_distance(dataset, 0.50)
  third_quantile_distance = quartile_distance(dataset, 0.75)

  print(paste("The distance of 1st quartile is", first_quantile_distance, "STD from the mean"))
  print(paste("The distance of 2nd quartile is", second_quantile_distance, "STD from the mean"))
  print(paste("The distance of 3rd quartile is", third_quantile_distance, "STD from the mean"))
}
```

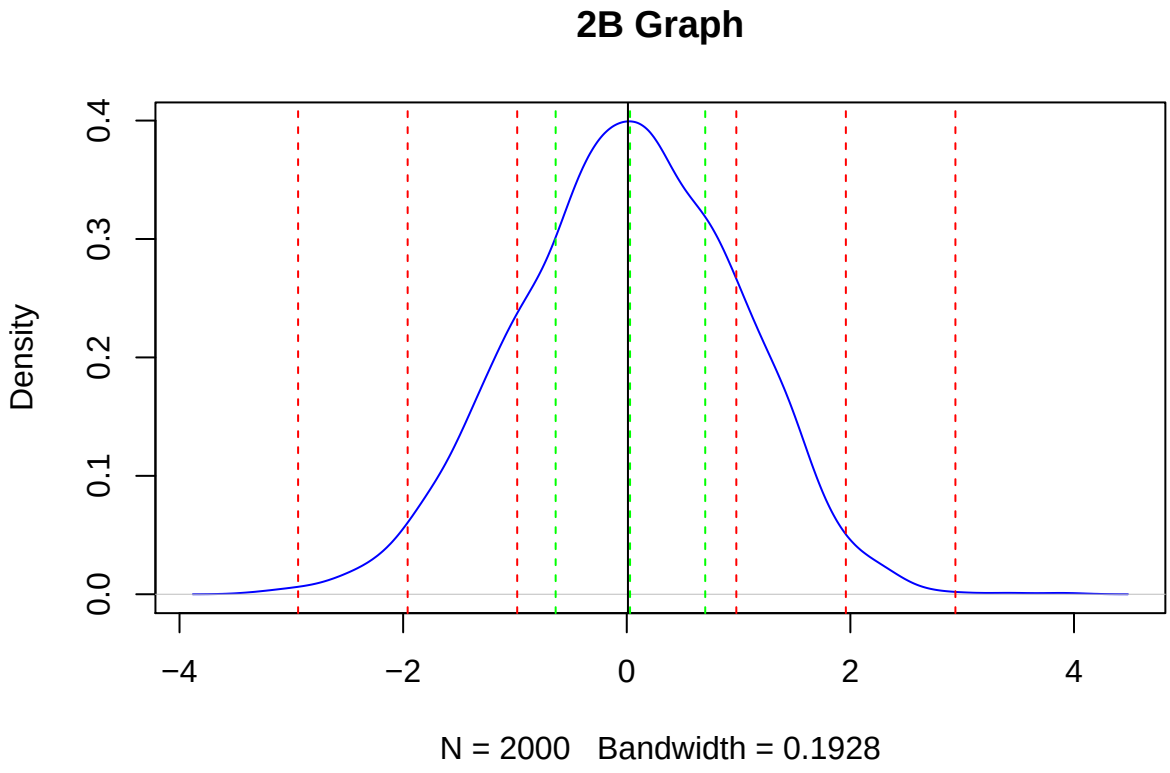
(a) `n = 2000, mean = 0, sd = 1` with Mean Line and STD lines

```r
first_dataset <- c(rnorm(n=2000, mean=0, sd=1))

plot(density(first_dataset), col="blue", main="2A Graph")

mean_std_lines(first_dataset)
```



(b) `n = 2000, mean = 0, sd = 1` with Mean Line, STD lines, and Quartile Lines

```
second_dataset <- c(rnorm(n=2000, mean=0, sd=1))

plot(density(second_dataset), col="blue", main="2B Graph")

mean_std_lines(second_dataset)
quartiles_lines(second_dataset)
```

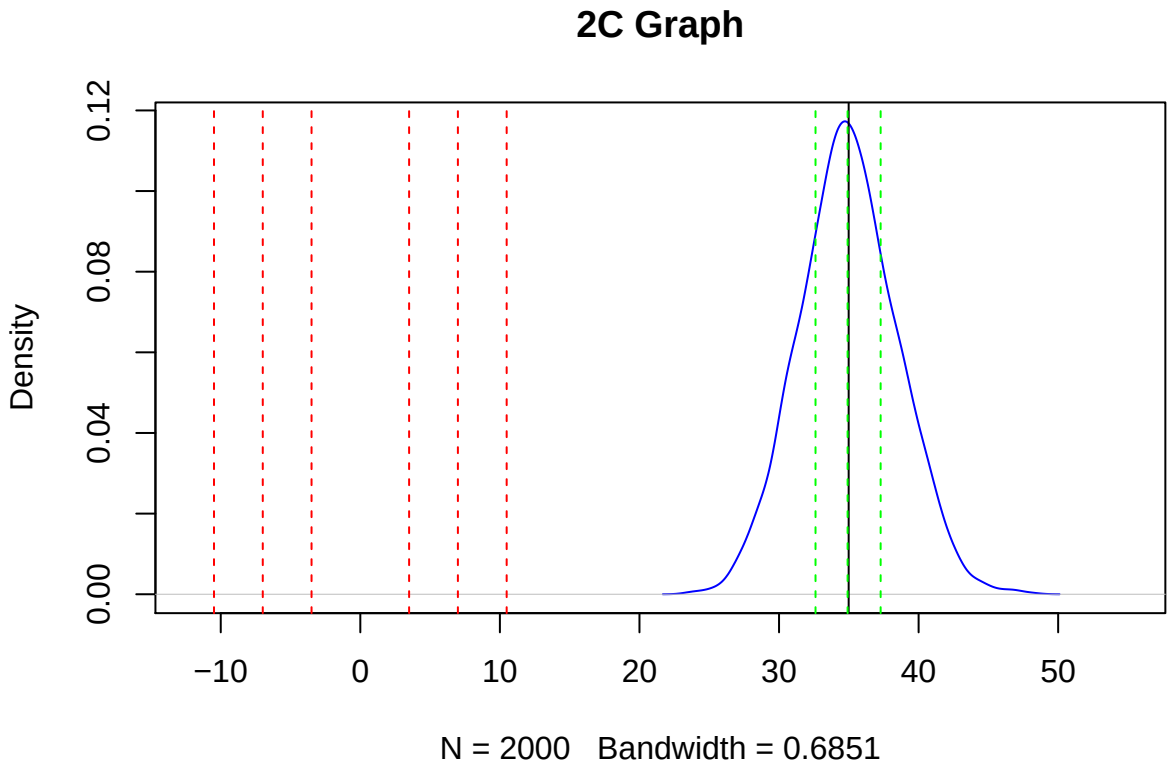



```
quartiles_distance(second_dataset)
```

```
## [1] "The distance of 1st quartile is -0.646676079271311 STD from the mean"
## [1] "The distance of 2nd quartile is 0.0187350101444562 STD from the mean"
## [1] "The distance of 3rd quartile is 0.691161393952457 STD from the mean"
```

(c) `n = 200, mean = 35, sd = 3.5` with Mean Line, STD lines, and Quartile Lines

```
third_dataset <- c(rnorm(n=2000, mean=35, sd=3.5))

plot(density(third_dataset), col="blue", main="2C Graph", xlim=c(-12,55))

mean_std_lines(third_dataset)
quartiles_lines(third_dataset)
```



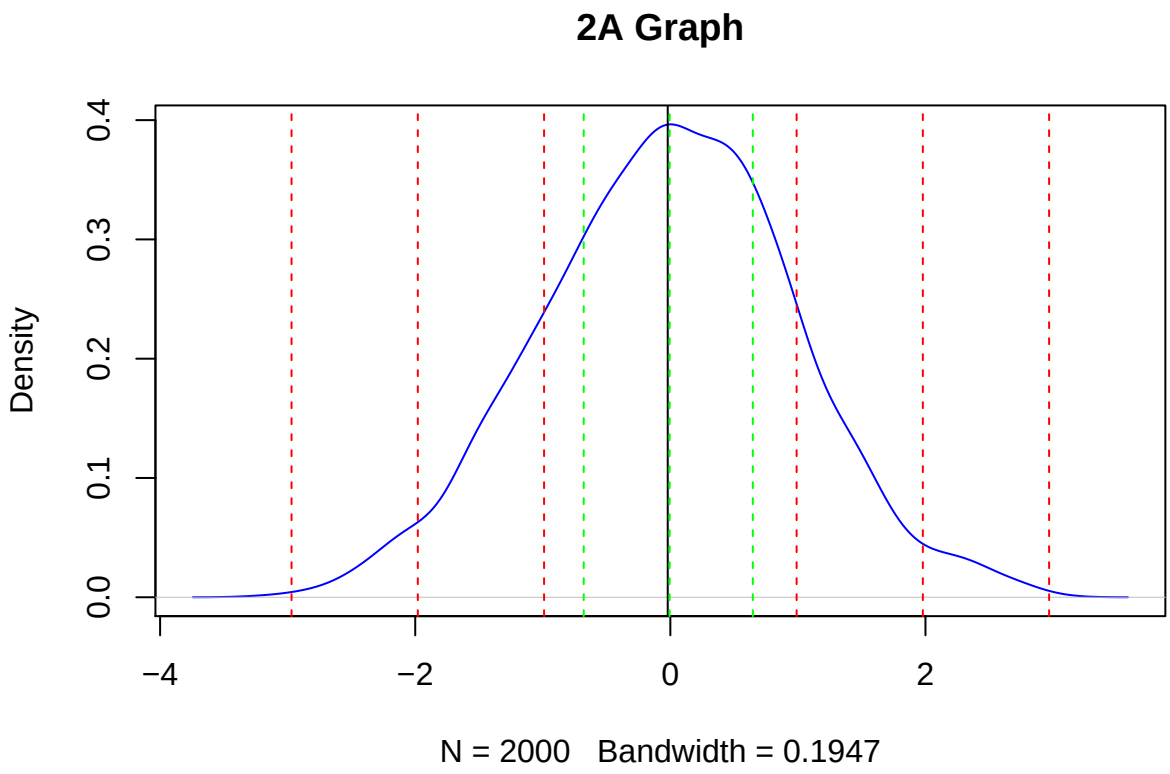
In the graph above, we can see that the 1st and the 3rd quartile is much further at the right hand side of the mean.

```
quartiles_distance(third_dataset)
```

```
## [1] "The distance of 1st quartile is 22.6035856912835 STD from the mean"
## [1] "The distance of 2nd quartile is 24.9019226172997 STD from the mean"
## [1] "The distance of 3rd quartile is 27.2685178440621 STD from the mean"
```

(d)

```
plot(density(first_dataset), col="blue", main="2A Graph")
mean_std_lines(first_dataset)
quartiles_lines(first_dataset)
```




```
quartiles_distance(first_dataset)
```

```
## [1] "The distance of 1st quartile is -0.658580821673416 STD from the mean"
## [1] "The distance of 2nd quartile is 0.0147199725278091 STD from the mean"
## [1] "The distance of 3rd quartile is 0.66729264114852 STD from the mean"
```

Question 3

(a) Calculate Bin Width/Number & Its Benefit

In his comment, the number of bins should be

$$k = (max - min)/h$$

where,

1. k is the number of bins
2. max is the maximum value in the observation group
3. min is the minimum value in the observation group
4. h is the size of each bin

In this case h is

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

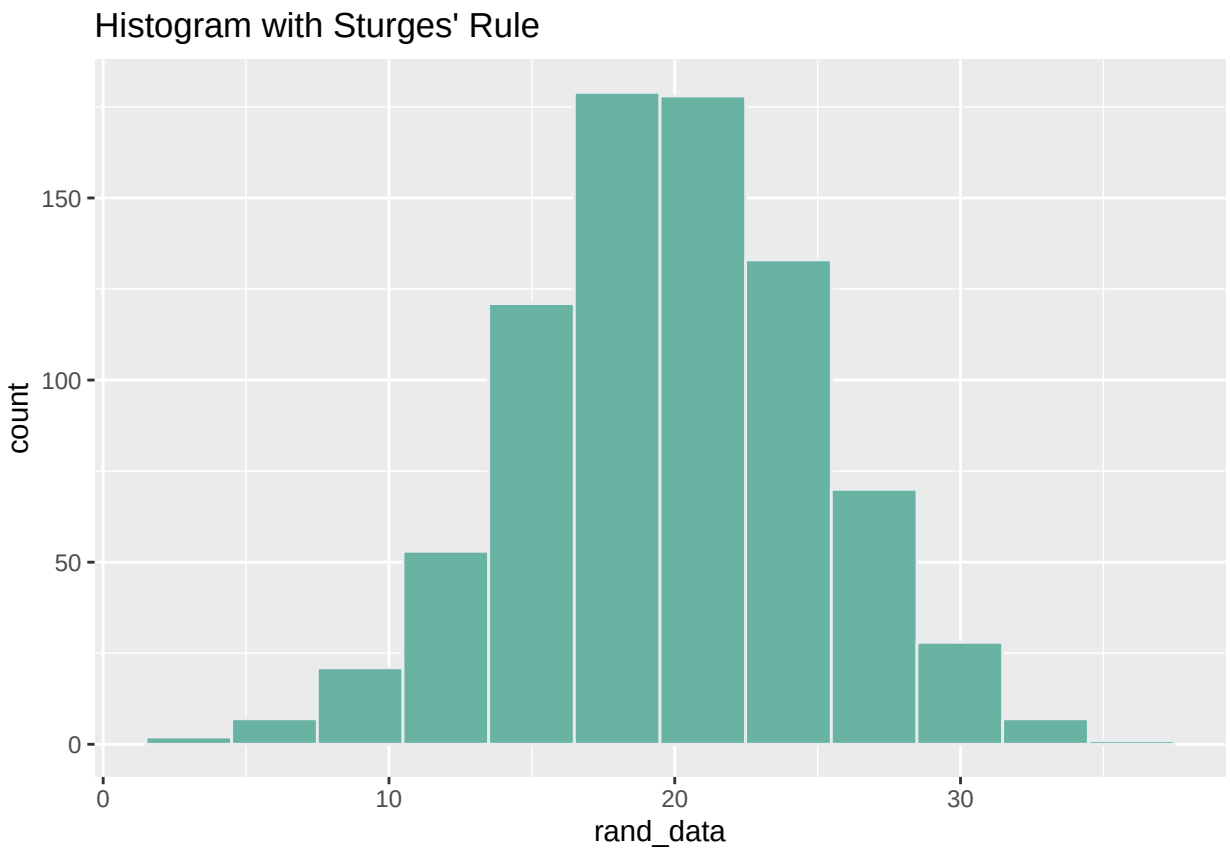
The benefit of Friedman-Diaconis' rule is

1. To minimize the difference between the area under the empirical probability distribution and the area under the theoretical probability distribution
2. The existence of outliers won't affect the bin width at ll with Friedman-Diaconis' Rule

(b)

1. Sturges' Formula

```
rand_data <- rnorm(800, mean=20, sd=5)
n <- length(rand_data)
sturges_bin_number = ceiling(log2(n)) + 1
sturges_bin_width = (max(rand_data) - min(rand_data)) / sturges_bin_number

ggplot(mapping = aes(rand_data)) +
  geom_histogram(
    bins=sturges_bin_number,
    binwidth=sturges_bin_width,
    fill="#69b3a2",
    color="#e9ecef"
  ) +
  ggtitle("Histogram with Sturges' Rule")
```




```
sturges_bin_number
```

```
## [1] 11
```

```
sturges_bin_width
```

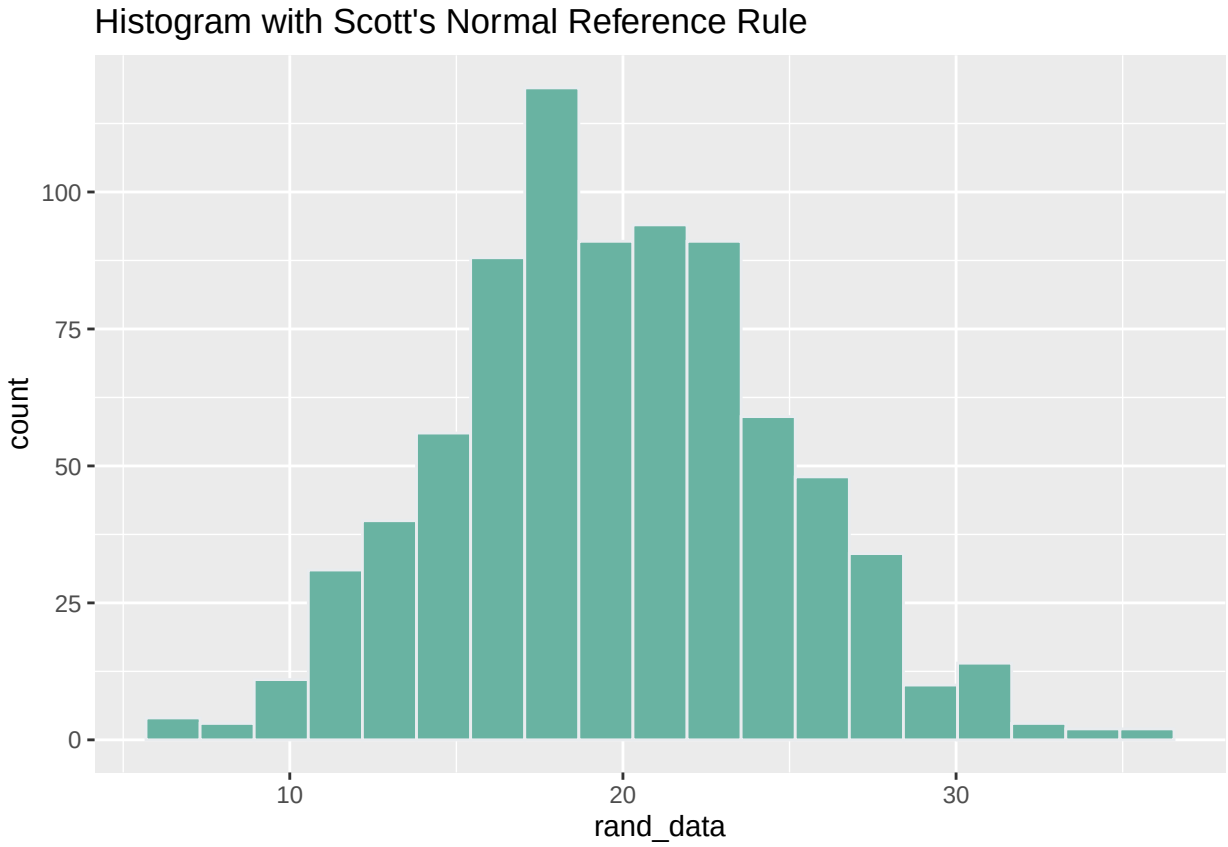
```
## [1] 2.997544
```

In Sturges' Rule, the number of bins is `11` and the bin width is `2.83`.

2. Scott's Normal Reference Rule

```
rand_data <- rnorm(800, mean=20, sd=5)
n <- length(rand_data)
scott_bin_width = (3.49 * sd(rand_data[c(1:80)])) / n^(1/3)
scott_bin_number = ceiling(max(rand_data) - min(rand_data)) + scott_bin_width

ggplot(mapping = aes(rand_data)) +
  geom_histogram(
    bins=scott_bin_number,
    binwidth=scott_bin_width,
    fill="#69b3a2",
```

```
    color="#e9ecef"
  ) +
  ggtitle("Histogram with Scott's Normal Reference Rule")
```




```
scott_bin_number
```

```
## [1] 31.62402
```

```
scott_bin_width
```

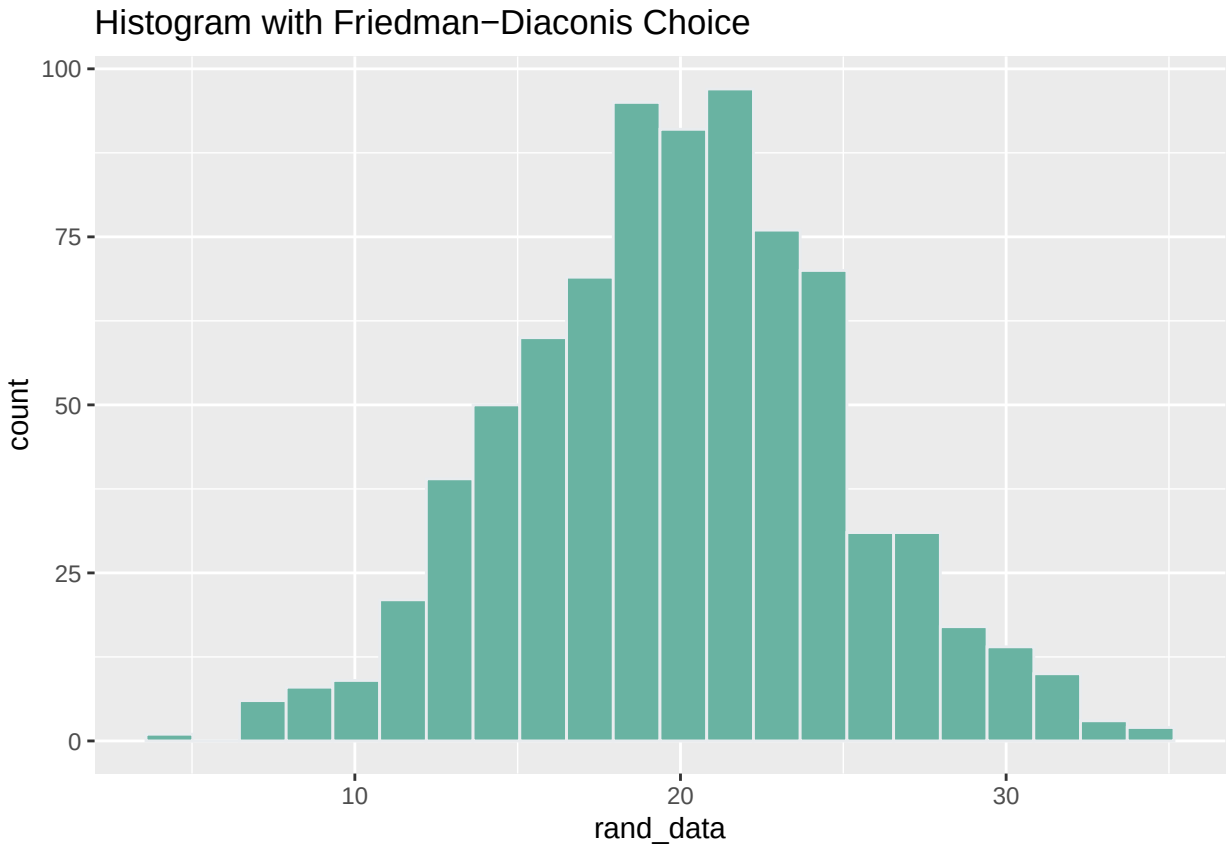
```
## [1] 1.624024
```

In Scott's Normal Reference Rule, the number of bins is `38` and the bin width is `1.89`.

3. Freedman-Diaconis' Choice (Using IQR)

```
rand_data <- rnorm(800, mean=20, sd=5)
n <- length(rand_data)
fd_bin_width = (2*IQR(rand_data)) / n^(1/3)
fd_bin_number = ceiling(max(rand_data) - min(rand_data)) + fd_bin_width

ggplot(mapping = aes(rand_data)) +
```

```
  geom_histogram(
    bins=fd_bin_number,
    binwidth=fd_bin_width,
    fill="#69b3a2",
    color="#e9ecef"
  ) +
  ggtitle("Histogram with Friedman-Diaconis Choice")
```




```
fd_bin_number
```

```
## [1] 32.43488
```

```
fd_bin_width
```

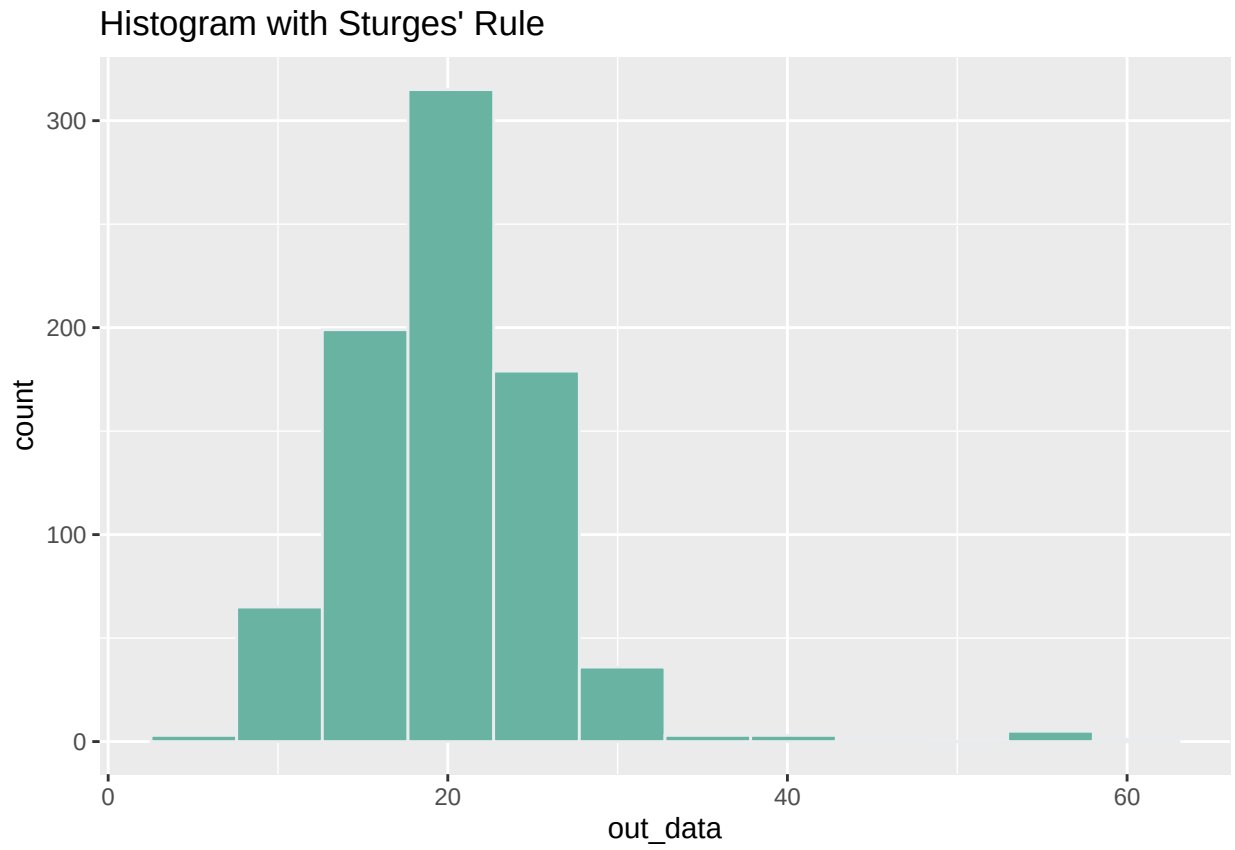
```
## [1] 1.434884
```

In Friedman-Diaconis Choice, the number of bins is `32.47621` and the bin width is `1.434905`.

(c)

1. Sturges' Formula

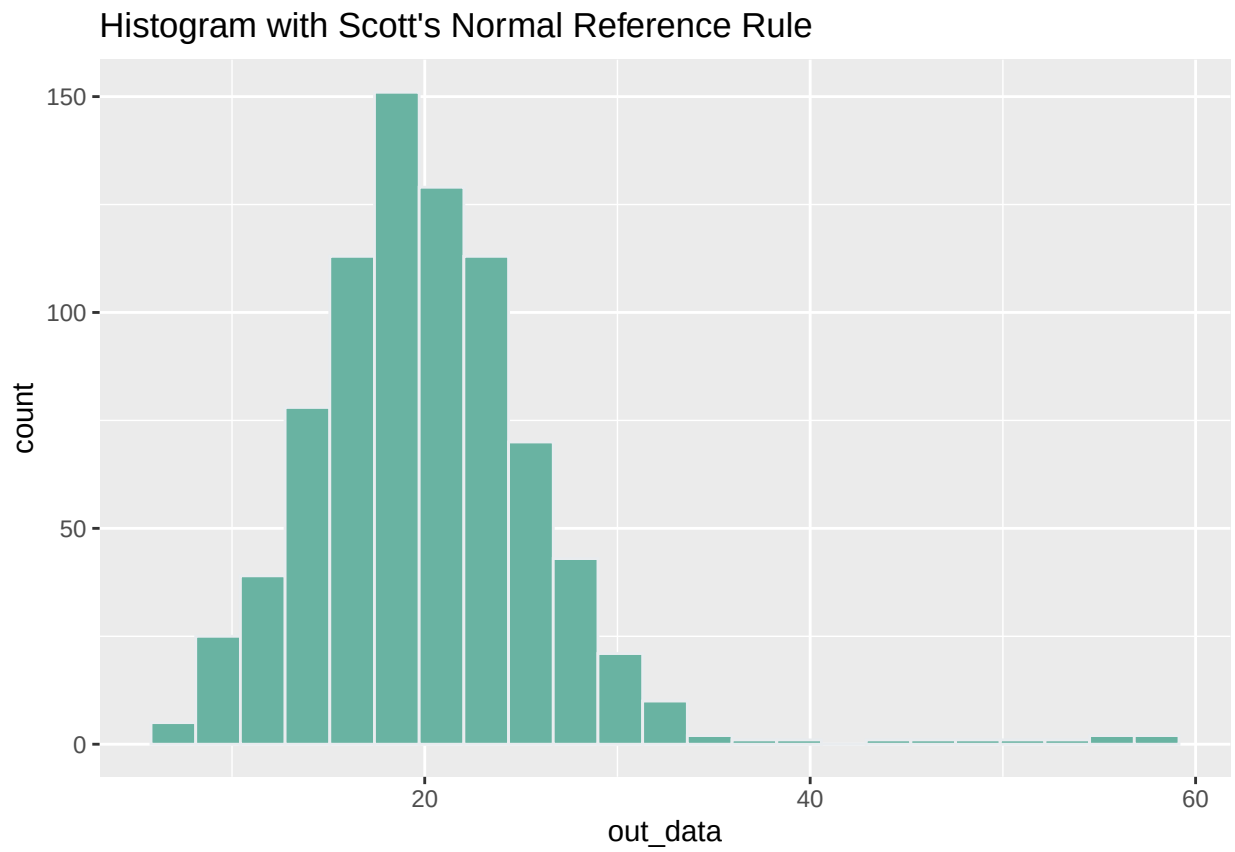
```r
rand_data <- rnorm(800, mean=20, sd = 5)
out_data <- c(rand_data, runif(10, min=40, max=60))
n <- length(out_data)
sturges_bin_number = ceiling(log2(n)) + 1
sturges_bin_width = (max(out_data) - min(out_data)) / sturges_bin_number

ggplot(mapping = aes(out_data)) +
  geom_histogram(
    bins=sturges_bin_number,
    binwidth=sturges_bin_width,
    fill="#69b3a2",
    color="#e9ecef"
  ) +
  ggtitle("Histogram with Sturges' Rule")
```



2. Scott's Normal Reference Rule

```r
rand_data <- rnorm(800, mean=20, sd = 5)
out_data <- c(rand_data, runif(10, min=40, max=60))
n <- length(out_data)
scott_bin_width = (3.49 * sd(out_data)) / n^(1/3)
scott_bin_number = ceiling(max(out_data) - min(out_data)) + scott_bin_width
```

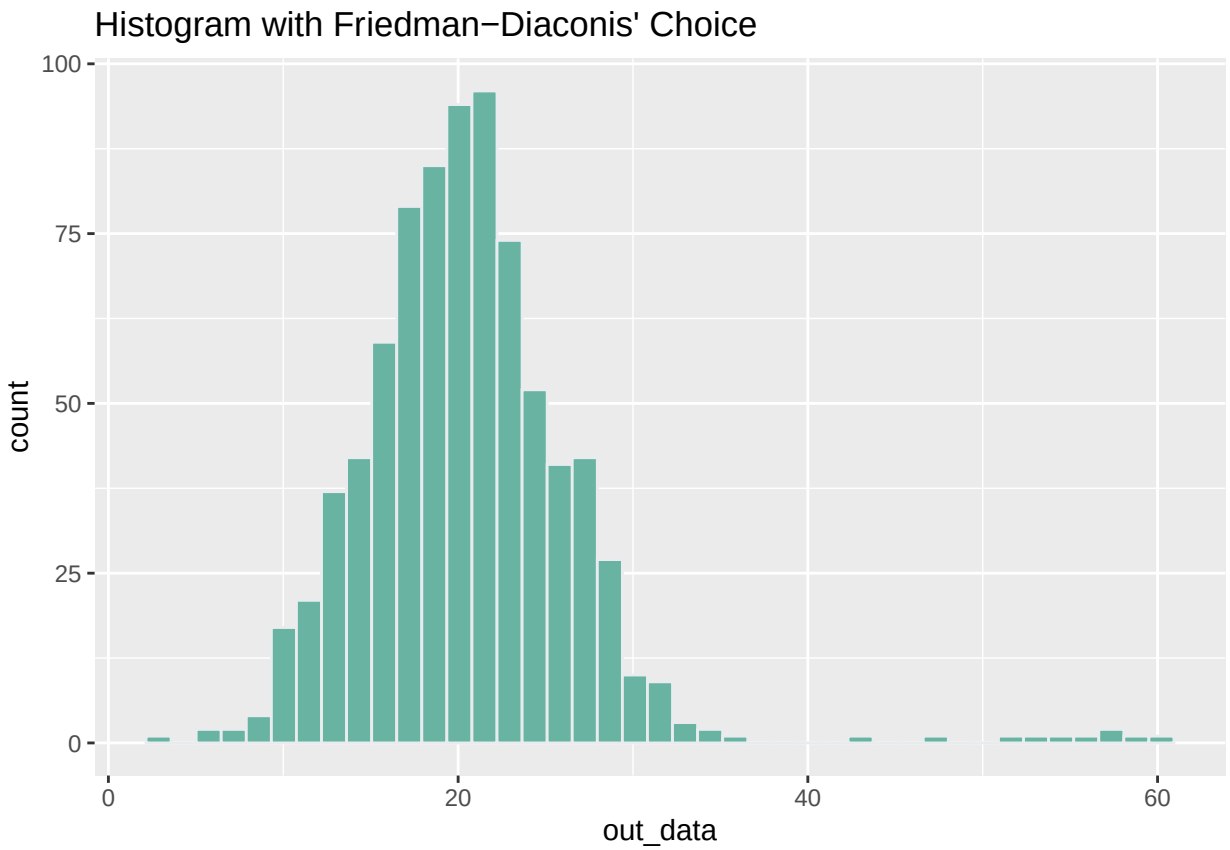
```
ggplot(mapping = aes(out_data)) +
  geom_histogram(
    bins=scott_bin_number,
    binwidth=scott_bin_width,
    fill="#69b3a2",
    color="#e9ecef"
  ) +
  ggtitle("Histogram with Scott's Normal Reference Rule")
```



3. Freedman-Diaconis' Choice (Using IQR)

```
rand_data <- rnorm(800, mean=20, sd = 5)
out_data <- c(rand_data, runif(10, min=40, max=60))
n <- length(out_data)
fd_bin_width = (2*IQR(out_data)) / n^(1/3)
fd_bin_number = ceiling(max(out_data) - min(out_data)) + fd_bin_width

ggplot(mapping = aes(out_data)) +
  geom_histogram(
    bins=fd_bin_number,
    binwidth=fd_bin_width,
    fill="#69b3a2",
    color="#e9ecef"
```

```
) +
ggtitle("Histogram with Friedman-Diaconis' Choice")
```



(d)

It has been said above that outliers are able to increase or decrease Mean value dramatically. Thus, the bin width in Scott's Normal Reference Rule will change as soon as outliers are included in the system. Unlike Friedman-Diaconis' Choice which uses Inter Quartile Range (IQR). IQR can be acquired with

$$IQR = Q(3/4) - Q(1/4)$$

Clearly, there is no mean in the equation, and outliers won't affect the bin width at all in this scenario.