BACS HW2

Student ID: 110065508

Due: 2022/02/27

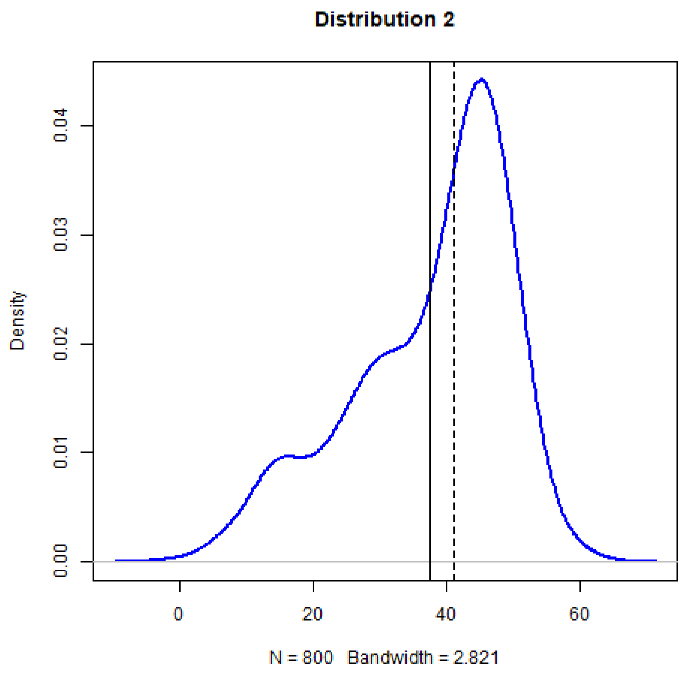
1. (a) The required distribution is generated by the following codes:

```r
# Question 1 (a)

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

d123 <- c(d1, d2, d3) # Distribution 2

png(filename = "1a.png")
plot(density(d123), col="blue", lwd=2, main = "Distribution 2") # plot pdf
abline(v=mean(d123)) # add vertical lines
abline(v=median(d123), lty="dashed")
dev.off()
```

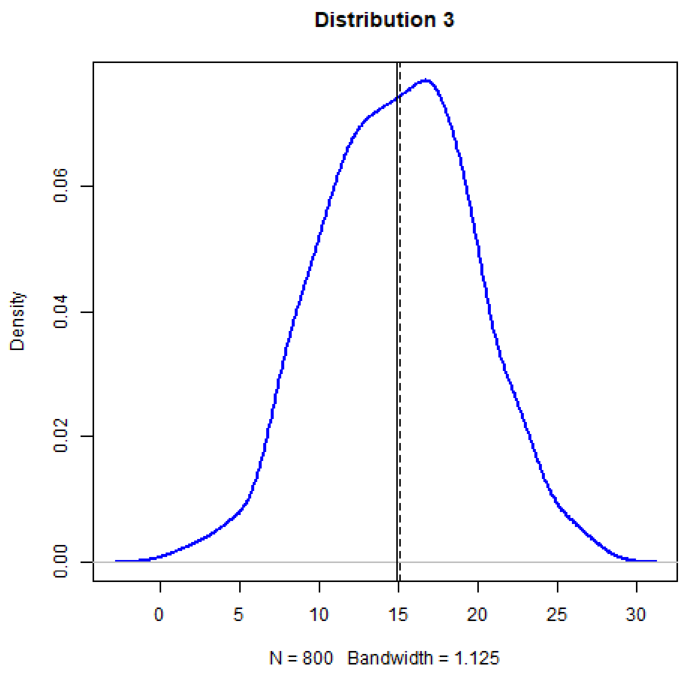


The result shows that mean=37.57582 and median=41.17198.

(b) The required distribution is generated by the following codes:

```r
# Question 1 (b)

d4 <- rnorm(n=800, mean=15, sd=5)
png(filename = "1b.png")
plot(density(d4), col="blue", lwd=2, main = "Distribution 3") # plot pdf
abline(v=mean(d4)) # add vertical lines
abline(v=median(d4), lty="dashed")
dev.off()
```



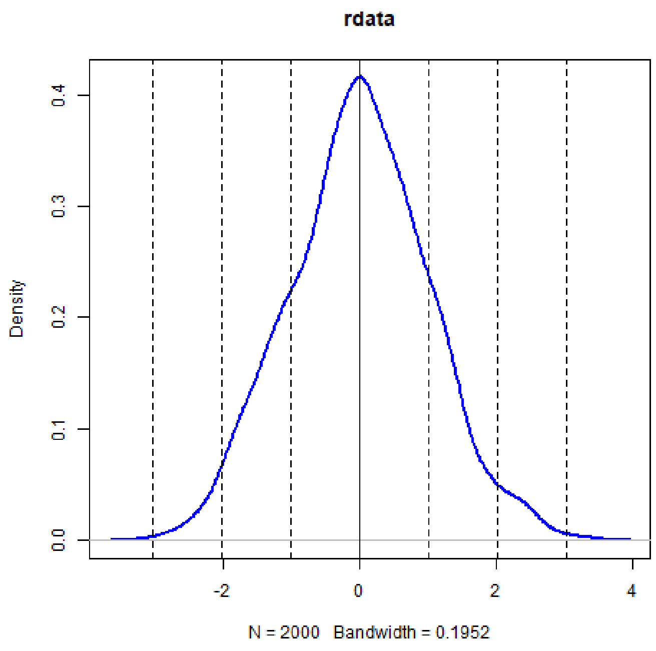
The result shows that mean= 14.93277 and median= 15.05891

(c) The mean does. Recall from probability theory that the median is an **order statistic**, which pdf does not depend on the value of the samples. However, in a computational aspect, the mean depends on the value of the samples. This makes the mean more sensitive.

2. (a) The required distribution is generated by the following codes:

```r
# Question 2 (a)

rdata <- rnorm(n=2000, mean=0, sd=1)
png(filename = "2a.png")
plot(density(rdata), col="blue", lwd=2, main = "rdata") # plot pdf
abline(v=mean(rdata)) # add vertical lines
for(i in c(-3:3))
{
  abline(v=i*sd(rdata), lty="dashed")
}
dev.off()
```



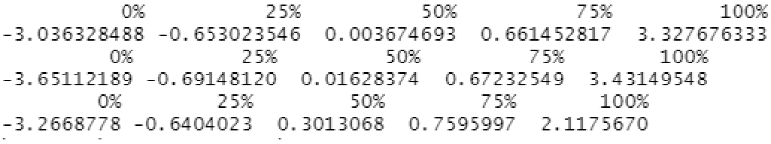
(b)(c)(d) It might be better if I answer these sub-questions together. First, by the following codes,

```r
# Question 2 (b)
q <- quantile(rdata)
q <- (q - mean(rdata))/sd(rdata)
print(q)

# Question 2 (c)
rdata <- rnorm(n=2000, mean=35, sd=3.5)
q <- quantile(rdata)
q <- (q - mean(rdata))/sd(rdata)
print(q)

# Question 2 (d)
q <- quantile(d123)
q <- (q - mean(d123))/sd(d123)
print(q)
```

I have the results:



```
           0%          25%          50%          75%         100% 
-3.036328488 -0.653023546  0.003674693  0.661452817  3.327676333 
         0%         25%         50%         75%        100% 
-3.65112189 -0.69148120  0.01628374  0.67232549  3.43149548 
        0%        25%        50%        75%       100% 
-3.2668778 -0.6404023  0.3013068  0.7595997  2.1175670 
```

That means, round up to 4 significant figures, the standard deviations that each quantile away from the mean is given by the following table.

quantiles	0%	25%	50%	75%	100%
(b)	-3.0363	-0.6530	0.0037	0.6615	3.3277
(c)	-3.6511	-0.6915	0.0163	0.6723	3.4315
(d)	-3.2669	-0.6404	0.3013	0.7596	2.1176

It is unlikely that the outcomes of (b) and (c) to be identical since we computes these number by sampling. However we may observed that the there's not much different, which meets our expects. However, the samples of (d) is not from a pure normal distribution, so we may expect that the outcomes to be somehow larger than (b) and (c) are, according to the plot of 1(a).

3. (a) He suggests the Freedman-Diaconis‘ choice, given by

$$h = 2 \times \frac{\text{IQR}}{\sqrt[3]{n}}$$

According to Wikipedia, (the Freedman-Diaconis‘ choice) **“is less sensitive than the standard deviation to outliers in data.”**

(b) Let the number of bins given by Sturges‘ formula be k, the bin widths given by Scott‘s normal reference rule be h_1, and Freedman-Diaconis‘ choice be h_2.

```r
# Question 3 (b)
rand_data <- rnorm(800, mean=20, sd = 5)

# Sturges' formula
k <- ceiling(log2(length(rand_data)))+1
# Scott's normal reference rule
h1 <- 3.49 * sd(rand_data) / length(rand_data)^(1/3)
# Freedman-Diaconis' choice
h2 <- 2 * IQR(rand_data) / length(rand_data)^(1/3)

cat("k=", k, ", h1=", h1, ", h2=", h2, "\n")
```

It turns out that $k = 11$, $h_1 \approx 1.8184$, $h_2 \approx 1.3980$.

(c) By the following codes,

```r
# Question 3 (c)
out_data <- c(rand_data, runif(10, min=40, max=60))

# Sturges' formula
k <- ceiling(log2(length(out_data)))+1
# Scott's normal reference rule
h1 <- 3.49 * sd(out_data) / length(out_data)^(1/3)
# Freedman-Diaconis' choice
h2 <- 2 * IQR(out_data) / length(out_data)^(1/3)

cat("k=", k, ", h1=", h1, ", h2=", h2, "\n")
```

It turns out that $k = 11$, $h_1 \approx 2.1804$, $h_2 \approx 1.4177$.

Generally speaking, Freedman-Diaconis‘ choice is least likely when outliers are added. That is because IQR is derived from quantiles, which is order statistics. In contrast, Scott‘s normal reference rule uses standard deviation form samples, which is likely to be effected by the outliers.