Mathematical Stats Homework #9

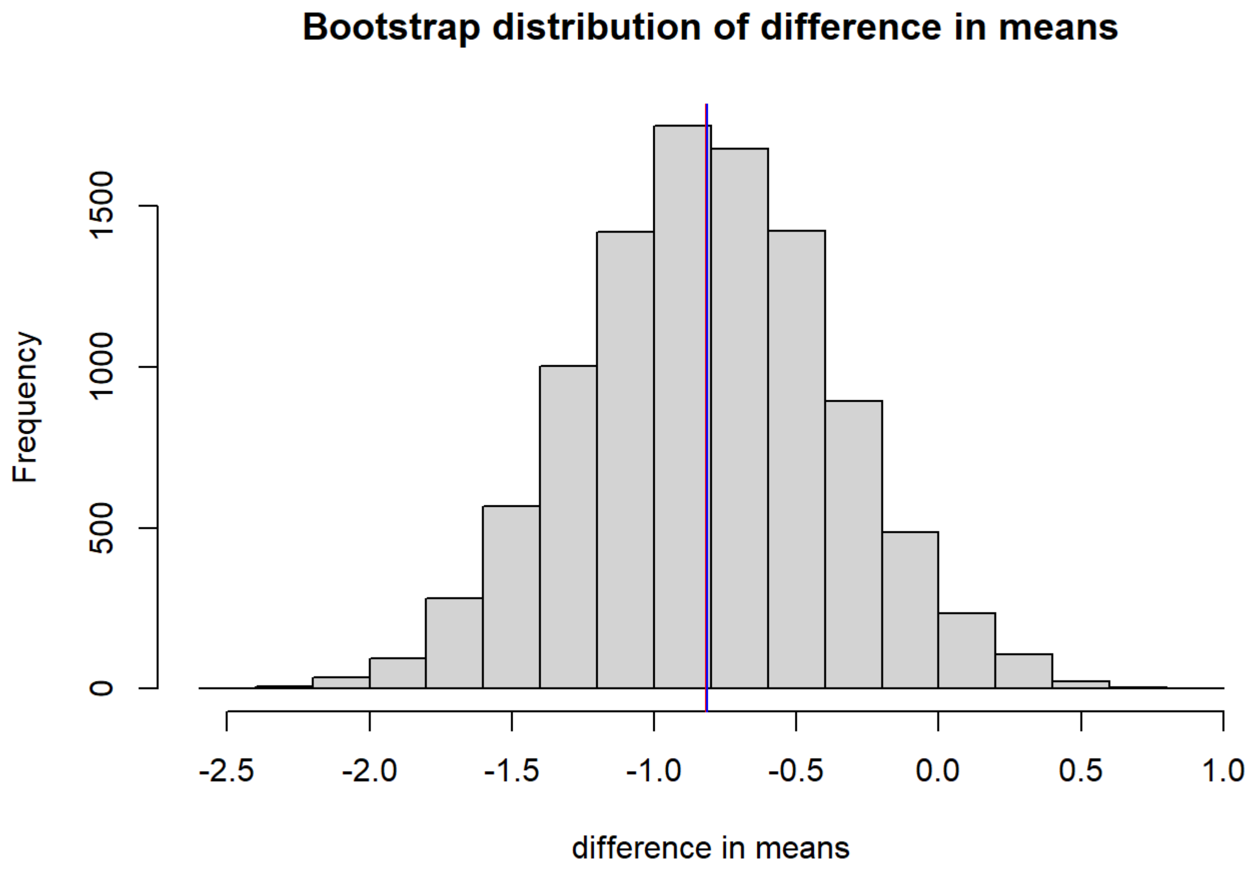
Vishesh Saharan

4/29/2019

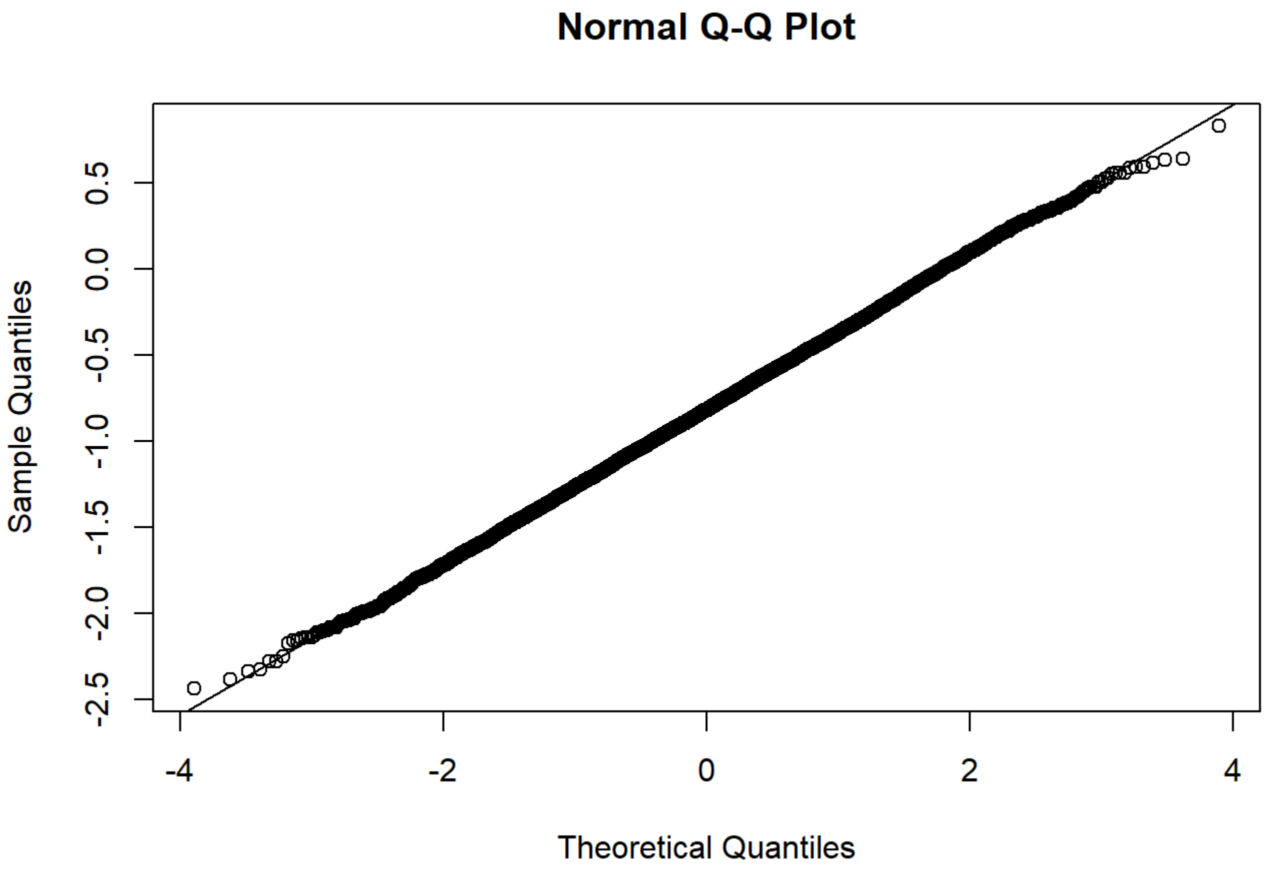
Problem 5.16c

Gender differences in math anxiety among Italian primary and secondary school children. AMAS is a unit of measurement that measures anxiety, with higher scores ndicating more anxiety. The dataset MathAnxiety contains the result for a subset if children in the study. What is the bootstrap estimate of the bias? What faction of the bootstrap standard error does this represent?

```
MathAnxiety<-read.csv("http://people.kzoo.edu/enordmoe/math365/data/MathAnxiety.csv")
MaleAMAS<-subset(MathAnxiety,select=AMAS,subset=Gender=="Boy",drop=T)
FemaleAMAS<-subset(MathAnxiety,select=AMAS,susbet=Gender=="Girl",drop=T)
n.MaleAMAS <- length(MaleAMAS)
n.FemaleAMAS <- length(FemaleAMAS)
N <- 10^4
diff <- numeric(N)
#set.seed(100)
for (i in 1:N)
{
  x.MaleAMAS <- sample(MaleAMAS, n.MaleAMAS, replace = TRUE)
  x.FemaleAMAS <- sample(FemaleAMAS, n.FemaleAMAS, replace = TRUE)
diff[i] <- mean(x.MaleAMAS) - mean(x.FemaleAMAS)
}
hist(diff, main = "Bootstrap distribution of difference in means",
    xlab = "difference in means")
abline(v = mean(diff),col = "red")
abline(v = mean(MaleAMAS) - mean(FemaleAMAS),
     col = "blue")
```



```
qqnorm(diff)
qqline(diff)
```



```
mean(diff)
```

```
## [1] -0.8161624
```

```
sd(diff)
```

```
## [1] 0.4499703
```

```
mean(diff)-(mean(MaleAMAS)-mean(FemaleAMAS))
```

```
## [1] -0.001708974
```

```
quantile(diff, c(0.025,0.975))
```

```
##        2.5%       97.5%
## -1.70265122  0.07662086
```

```
bias<-mean(diff)-(mean(MaleAMAS)-mean(FemaleAMAS))
bias/sd(diff)
```

```
## [1] -0.003797971
```

The bootstrap estimate of error is 0.021, which represents 2.1% of the bootstrap standard error.

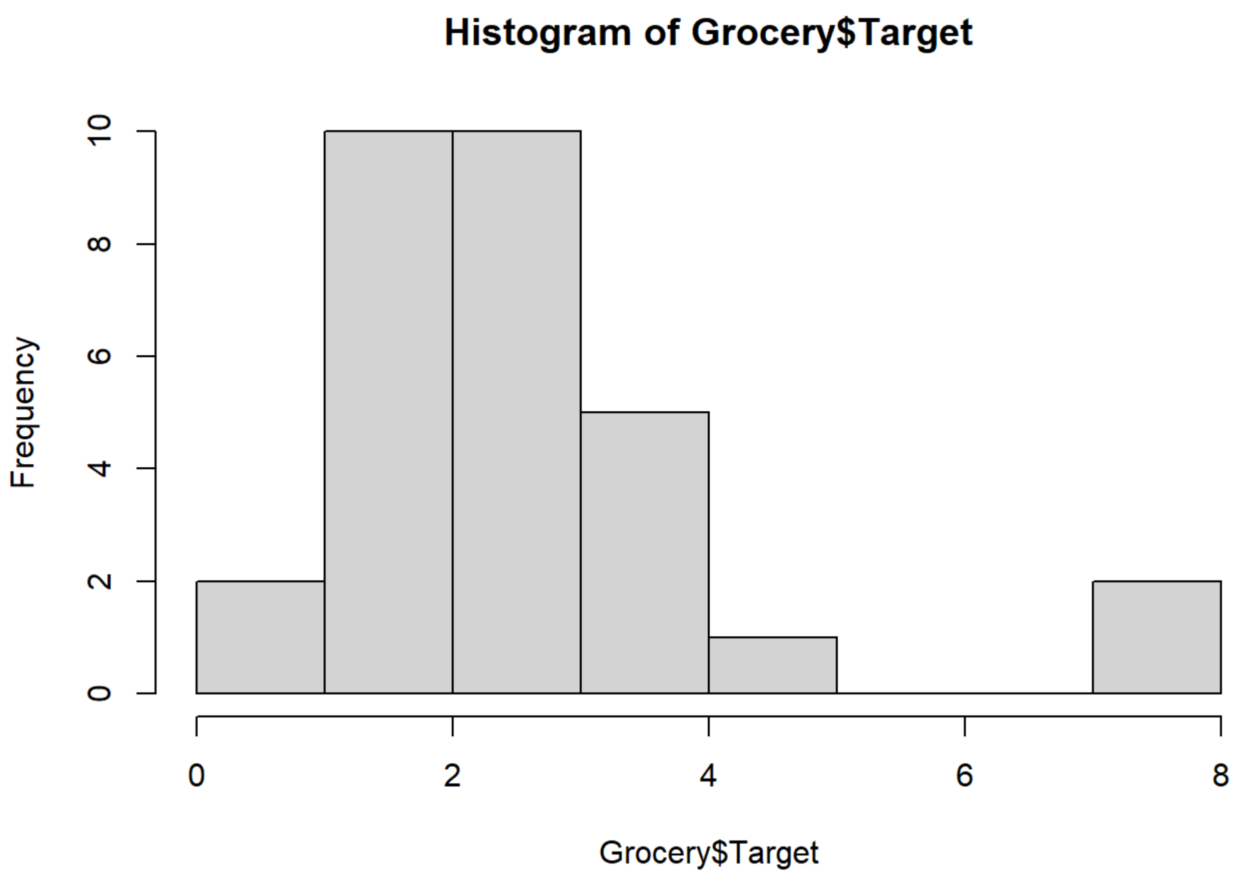
Problem 5.18

Is there a difference between the price of grocers sold in two retailers Target and Walmart? The data set Groceries contain, a sample of grocery items and their prices advertised on their respective websites on one specific day.

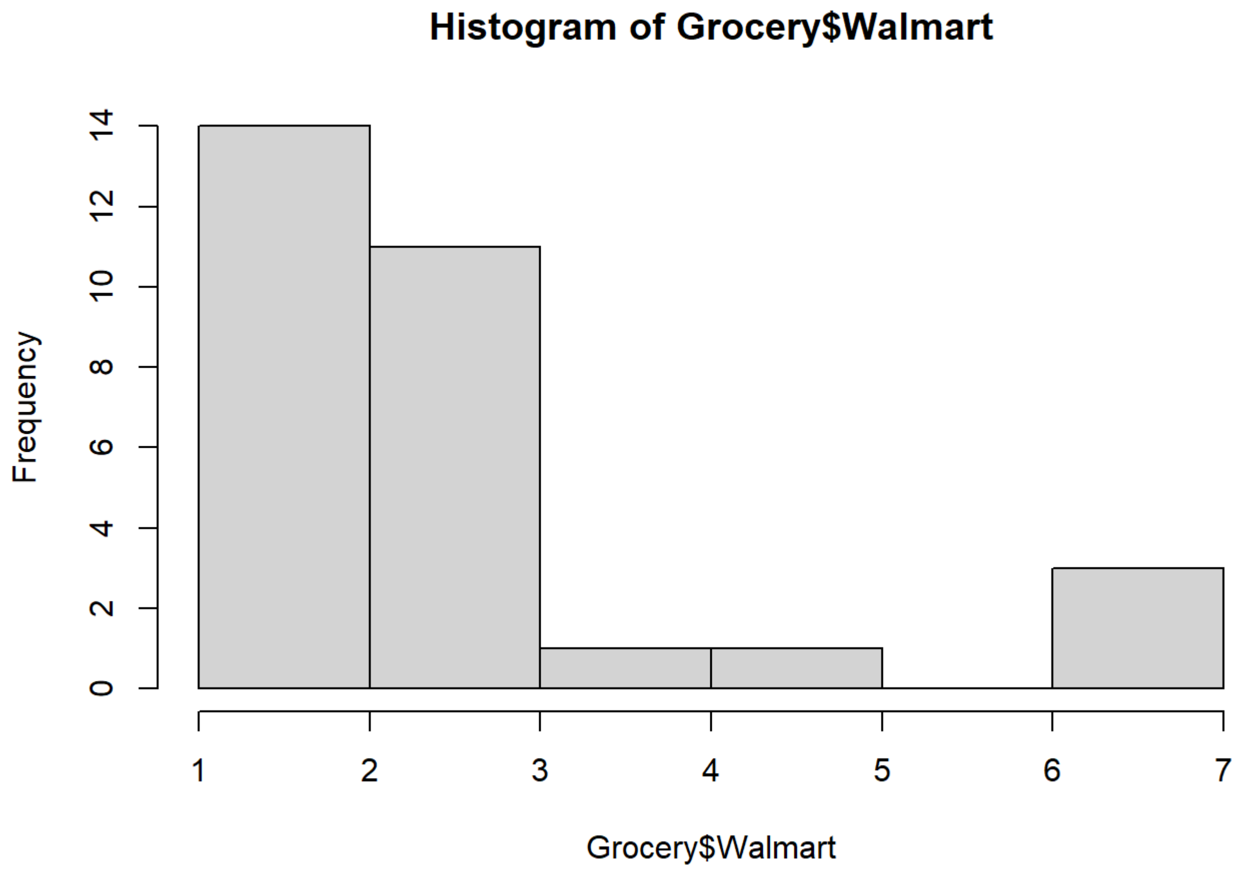
(a).Compute the summary statistics of the prices for each store. (b). Use bootstrap to determine whether or not there is a difference in the mean prices (c). Create a histrogram of the difference in prices. What is unusual about Quaker Oats Life Cereal? (d). Recompute the bootstrap percentile interval without this observation. What do you conclude?

(a).

```
Grocery<-read.csv("http://people.kzoo.edu/enordmoe/math365/data/Groceries.csv")
hist(Grocery$Target)
```



```
hist(Grocery$Walmart)
```



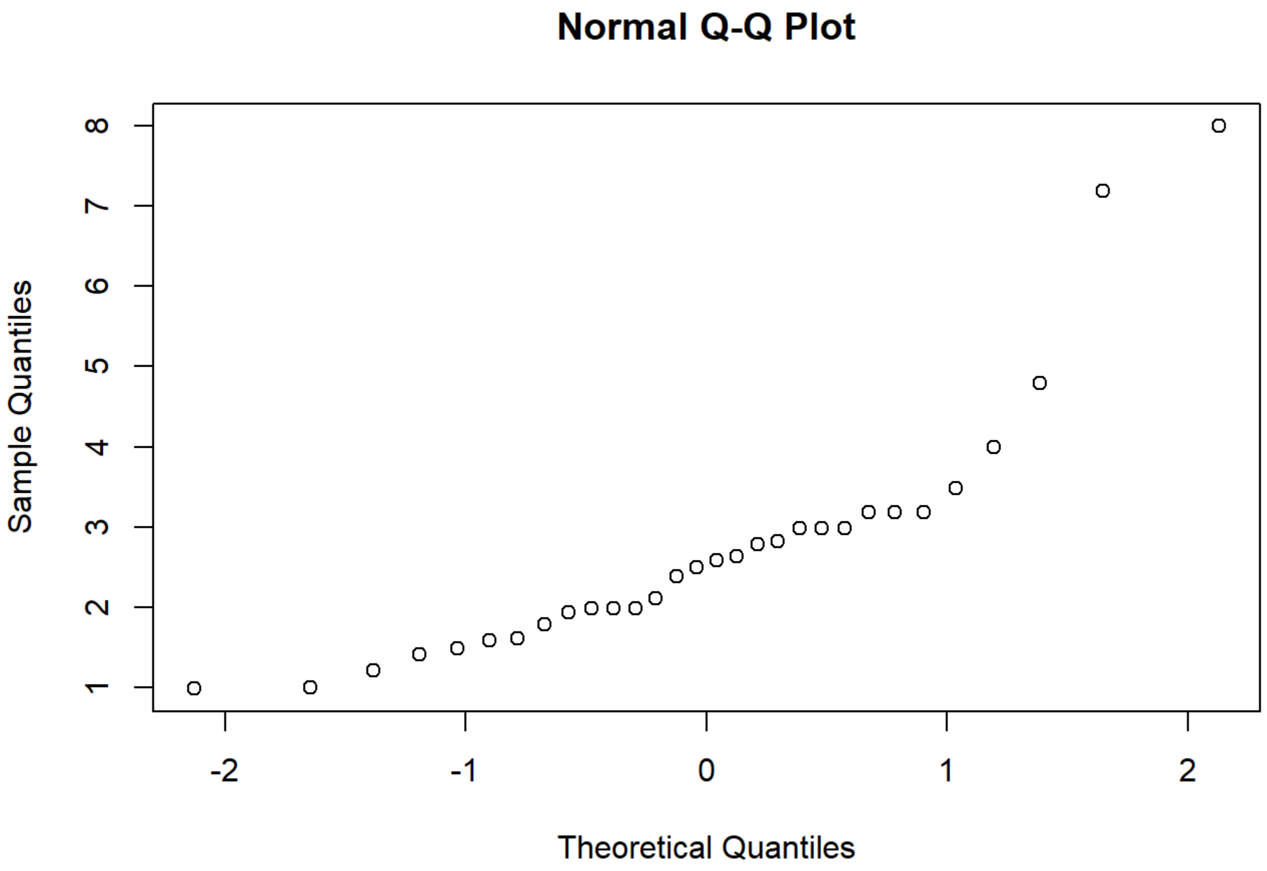
```
summary(Grocery$Target)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.990   1.827   2.545   2.762   3.140   7.990
```

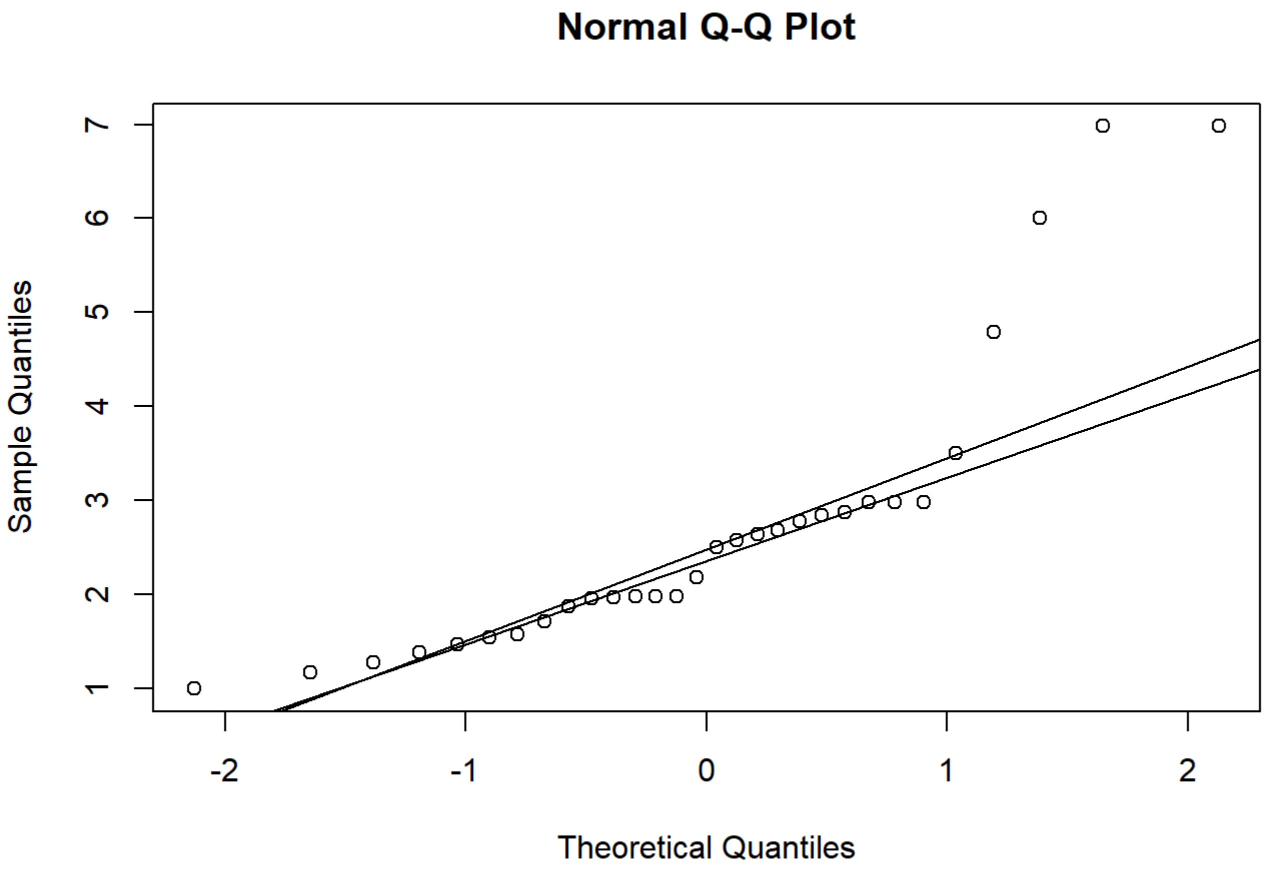
```
summary(Grocery$Walmart)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   1.760   2.340   2.706   2.955   6.980
```

```
qqnorm(Grocery$Target)
```

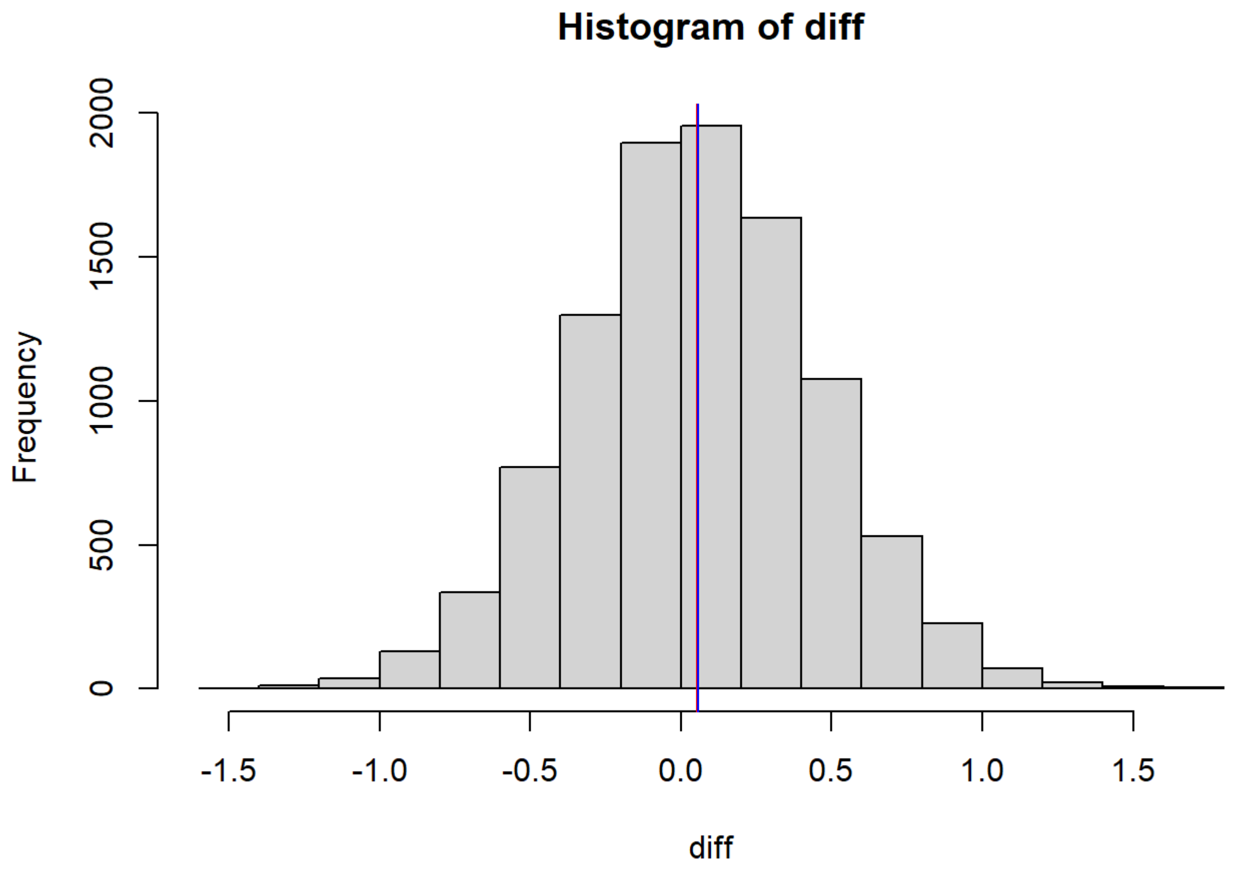


```
qqnorm(Grocery$Walmart)
qqline(Grocery$Target)
qqline(Grocery$Walmart)
```

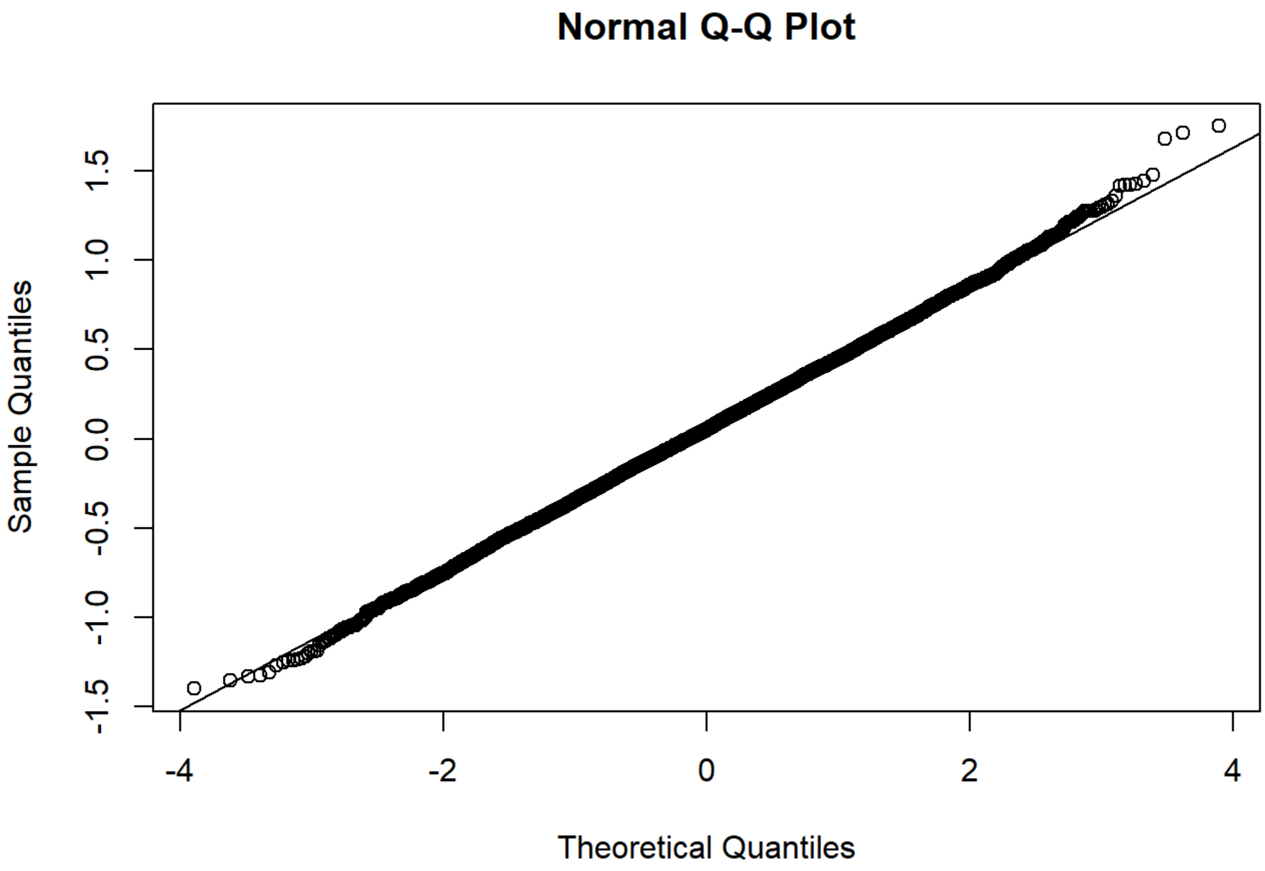


(b).

```
n.Target <- length(Grocery$Target)
n.Walmart <- length(Grocery$Walmart)
N <- 10^4
diff <- numeric(N)
#set.seed(100)
for (i in 1:N)
{
  x.Target <- sample(Grocery$Target, n.Target, replace = TRUE)
  x.Walmart <- sample(Grocery$Walmart, n.Walmart, replace = TRUE)
diff[i] <- mean(x.Target) - mean(x.Walmart)
}
hist(diff)
abline(v = mean(diff),col = "red")
abline(v = mean(Grocery$Target) - mean(Grocery$Walmart),
     col = "blue")
```



```
qqnorm(diff)
qqline(diff)
```



```
mean(diff)
```

```
## [1] 0.055032
```

```
sd(diff)
```

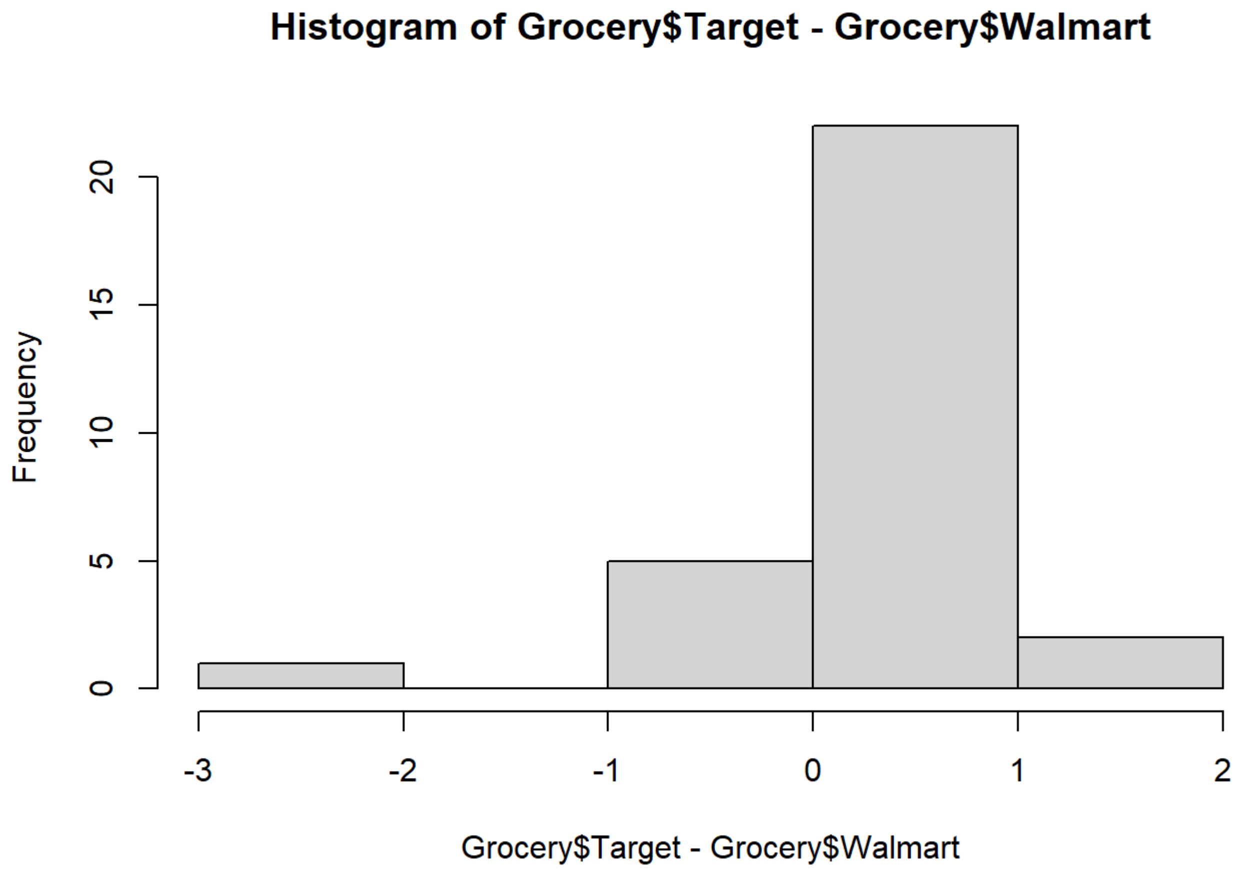
```
## [1] 0.4006754
```

```
quantile(diff, c(0.025,0.975))
```

```
##       2.5%      97.5% 
## -0.7393667  0.8397250
```

(c).

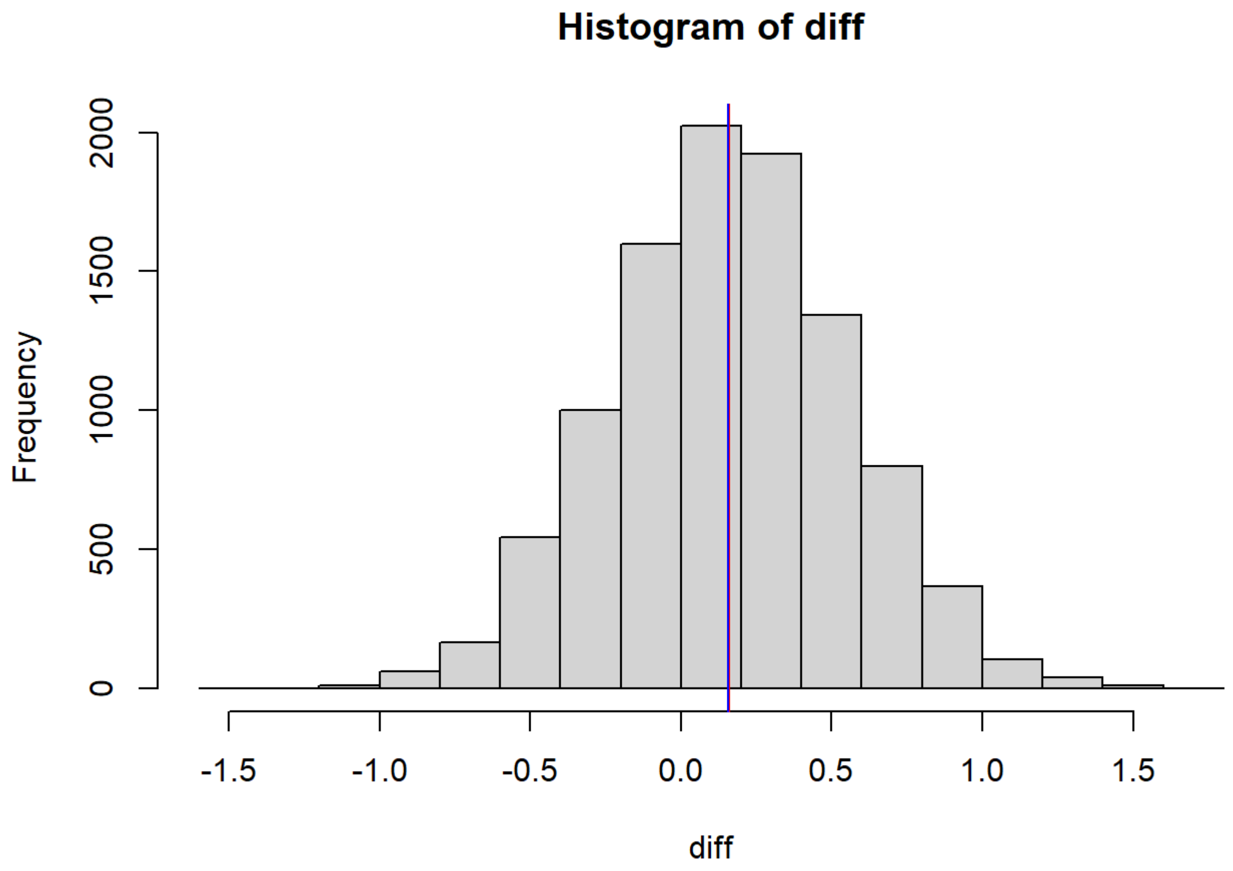
```
hist(Grocery$Target-Grocery$Walmart)
```



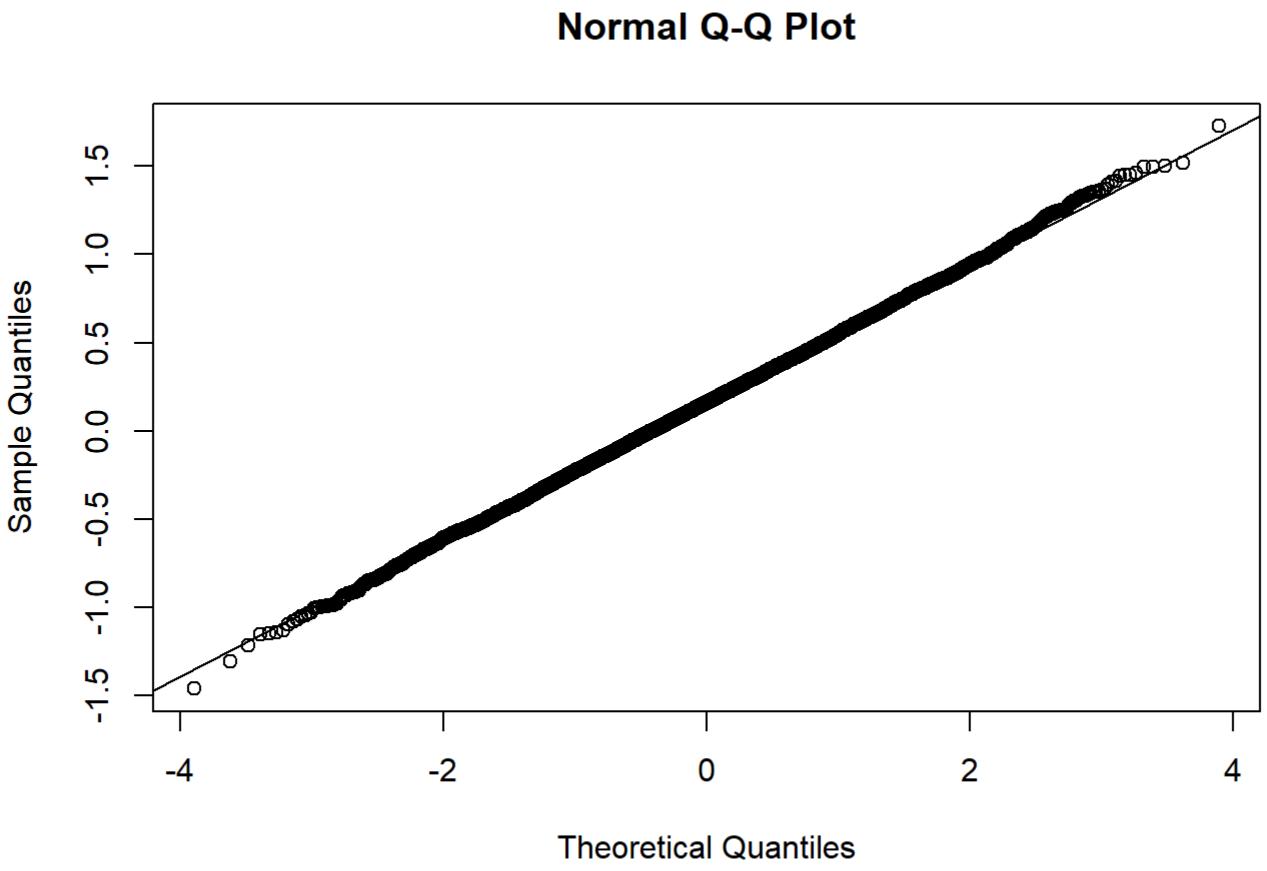
Difference between the two stores on the Quaker Oats Life Cereal is nearly $3, which makes it a heavy outlier compared to the other products.

(d).

```
NewGrocery<-Grocery[-2,c(1:4)]
n.Target <- length(NewGrocery$Target)
n.Walmart <- length(NewGrocery$Walmart)
N <- 10^4
diff <- numeric(N)
#set.seed(100)
for (i in 1:N)
{
  x.Target <- sample(NewGrocery$Target, n.Target, replace = TRUE)
  x.Walmart <- sample(NewGrocery$Walmart, n.Walmart, replace = TRUE)
diff[i] <- mean(x.Target) - mean(x.Walmart)
}
hist(diff)
abline(v = mean(diff),col = "red")
abline(v = mean(NewGrocery$Target) - mean(NewGrocery$Walmart),
     col = "blue")
```



```
qqnorm(diff)
qqline(diff)
```



```
mean(diff)
```

```
## [1] 0.1597298
```

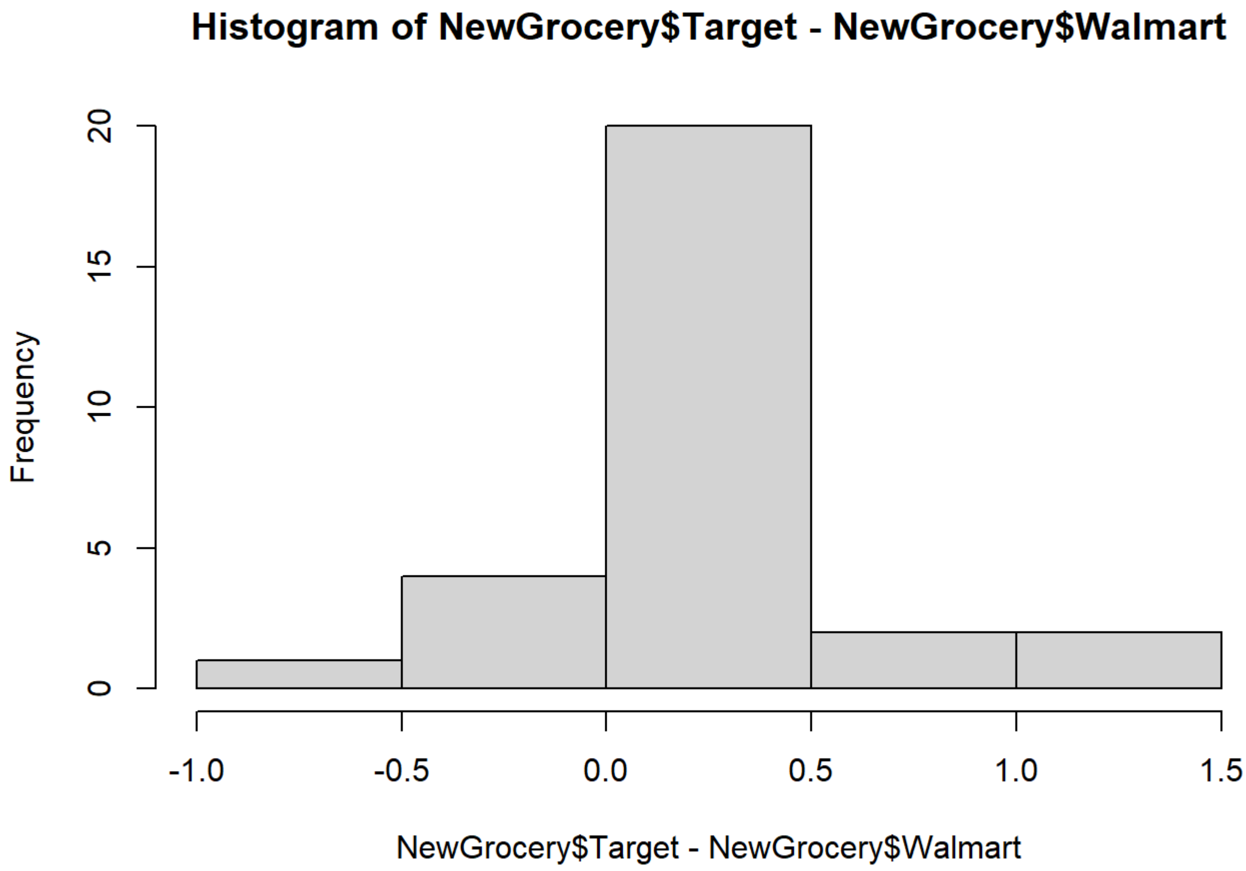
```
sd(diff)
```

```
## [1] 0.3923788
```

```
quantile(diff, c(0.025,0.975))
```

```
##       2.5%      97.5%
## -0.5965776  0.9242155
```

```
hist(NewGrocery$Target-NewGrocery$Walmart)
```



Without that outlier, it appears to have a more central and normal distribution.

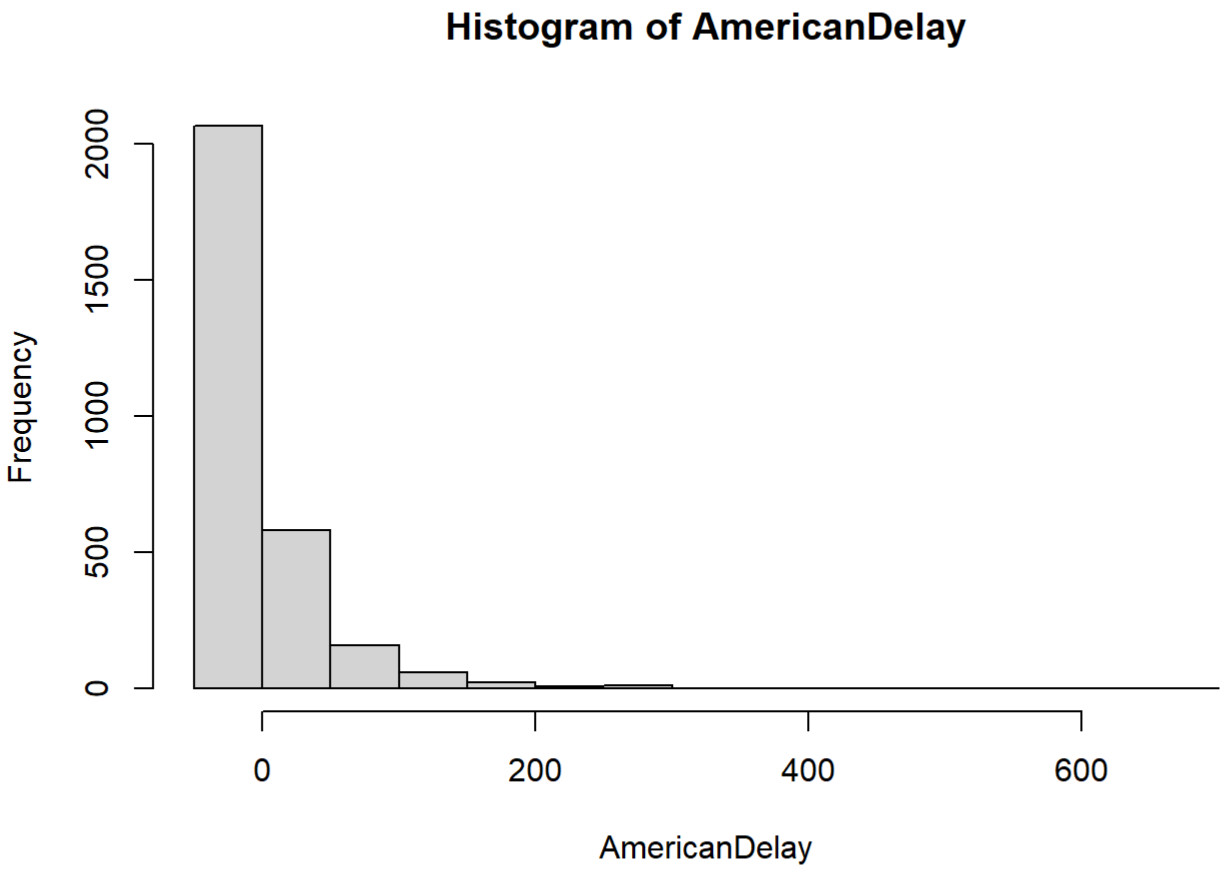
Problem 5.22

Import the data from flight delays case study. Although the data represent all UA and AA flights in May and June of 2009, we will assume they represent a sample from a larger population of UA and AA flights flown under similar circumstances. We will consider the ratio of means of the flight delay lengths, muUA/muAA

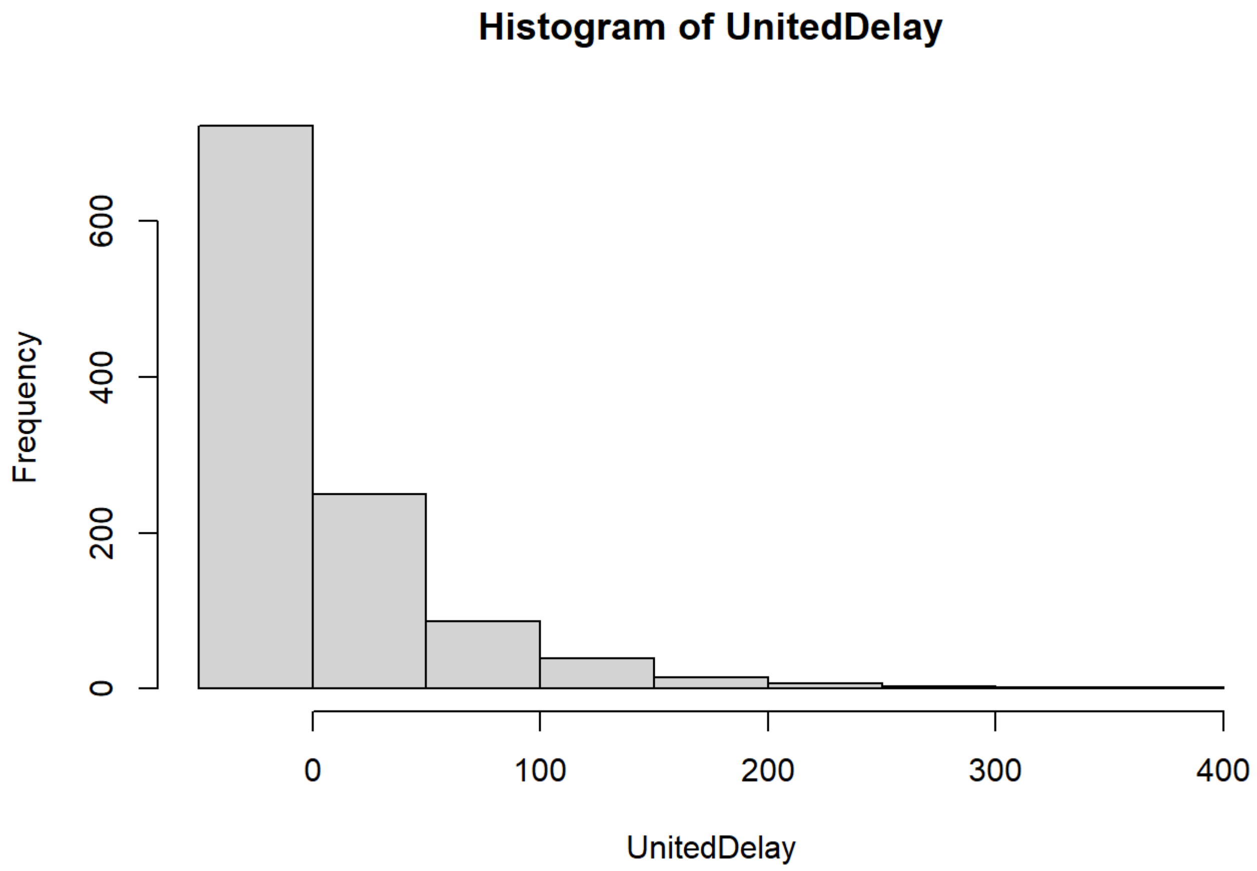
(a). Perform some exploratory data analysis on flight delay lengths for each of UA and AA flights. (b). Bootstrap the mean of flight delay lengths for each airline separately, and describe the distribution. (c). Bootstrap the ratio of means. Provide plots of the bootstrap distribution and describe the distribution. (d). Find the 95% bootstrap percentile interval for the ratio of means. Interpret this interval (e). What is the bootstrap estimate of the bias? What fraction of the bootstrap error does it represent? (f). For inference in this text, we assume that the observations are independent. Is that condition met here?

(a).

```
FlightDelay<-read.csv("http://people.kzoo.edu/enordmoe/math365/data/FlightDelays.csv")
AmericanDelay<-subset(FlightDelay,select=Delay,subset=Carrier=="AA",drop=T)
UnitedDelay<-subset(FlightDelay,select=Delay,subset=Carrier=="UA",drop=T)
Delay1<-subset(FlightDelay,select=Delay,drop=T)
hist(AmericanDelay)
```



```
hist(UnitedDelay)
```



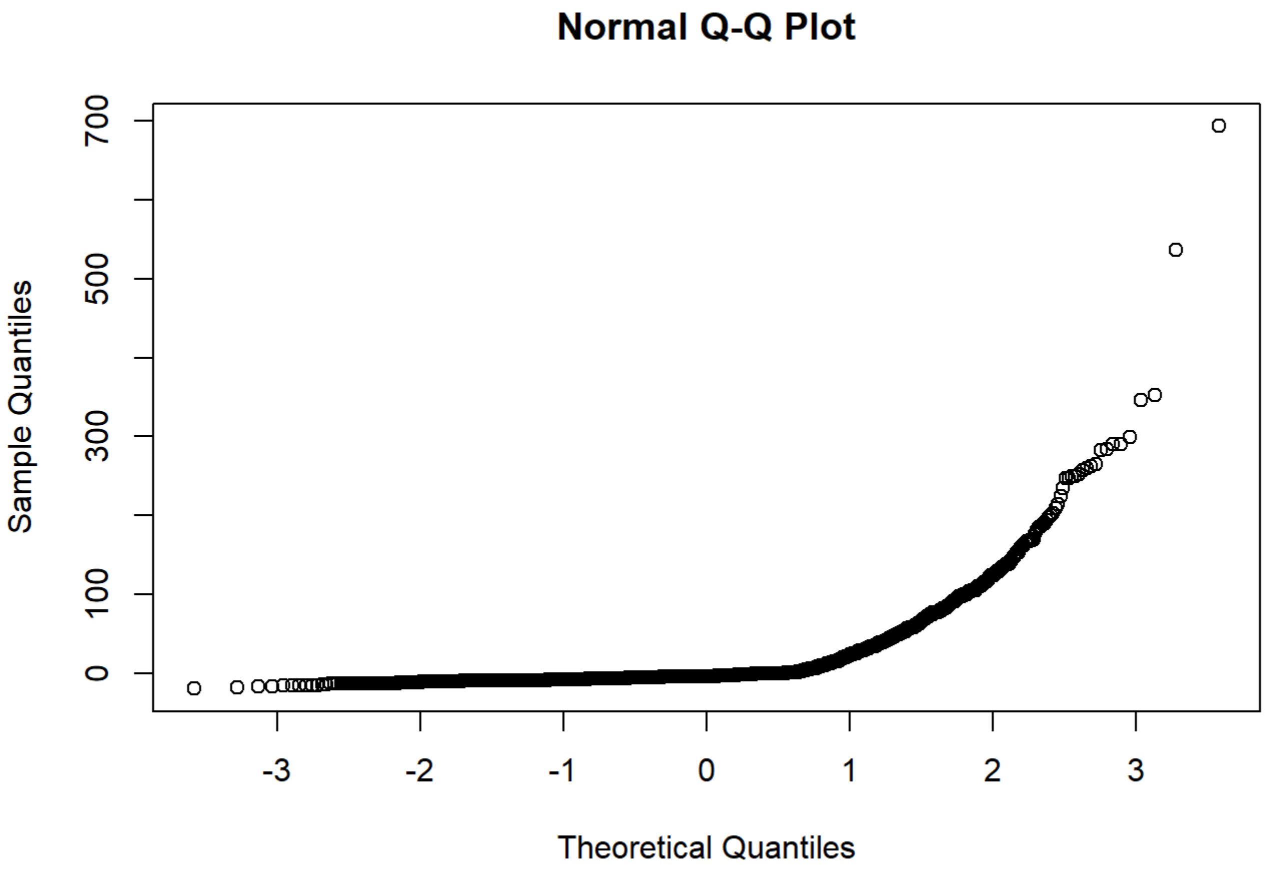
```
summary(AmericanDelay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   -19.0    -6.0    -3.0    10.1     4.0   693.0
```

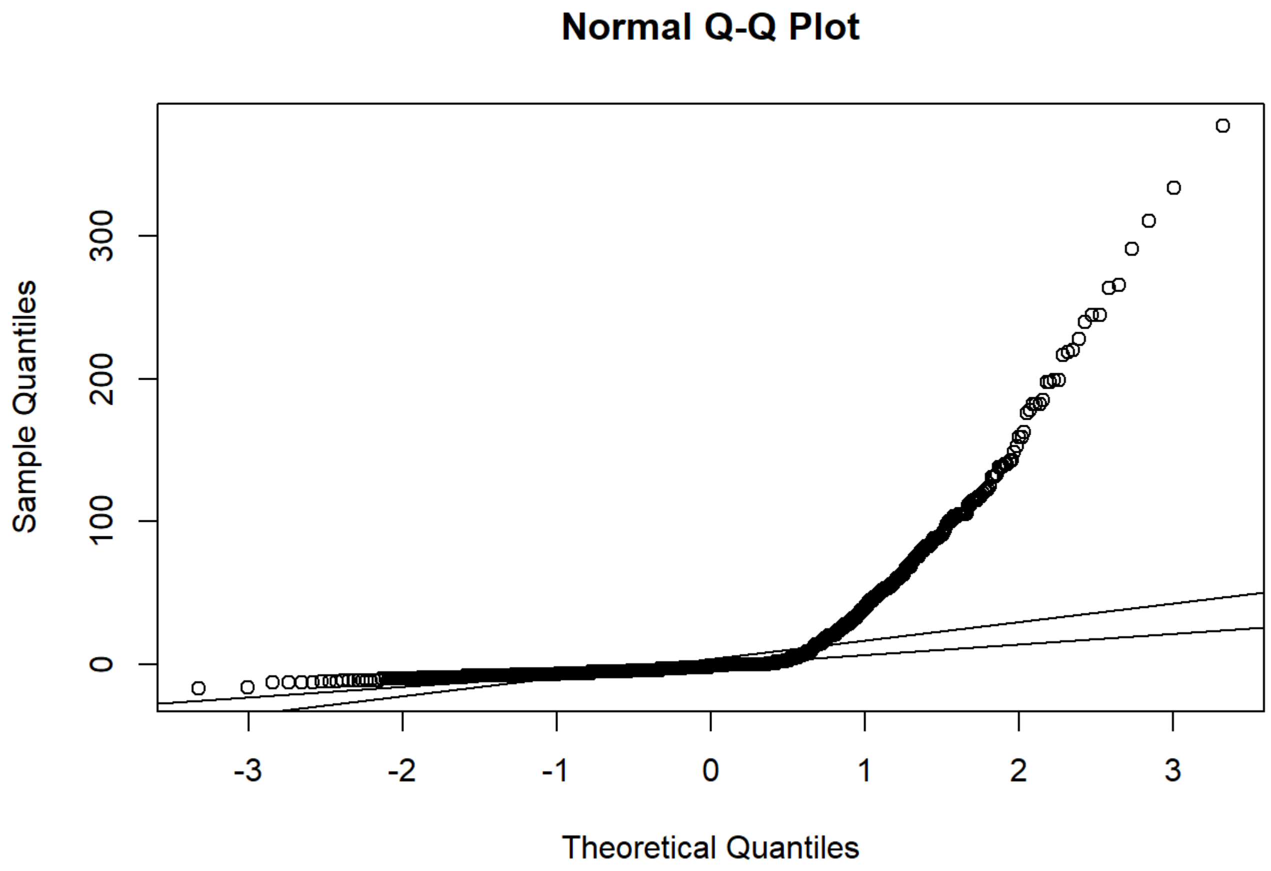
```
summary(UnitedDelay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  -17.00   -5.00   -1.00   15.98   12.50  377.00
```

```
qqnorm(AmericanDelay)
```



```
qqnorm(UnitedDelay)
qqline(AmericanDelay)
qqline(UnitedDelay)
```

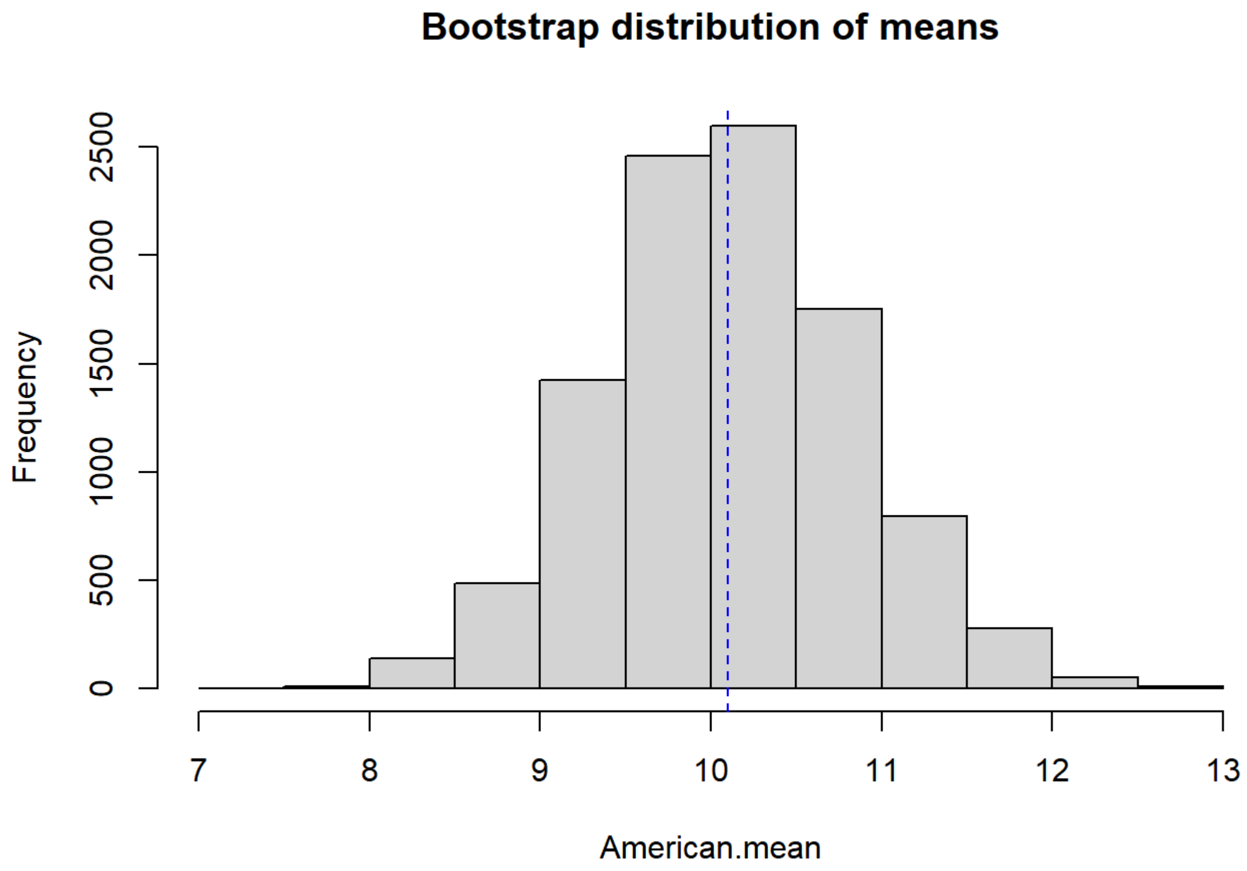


(b).

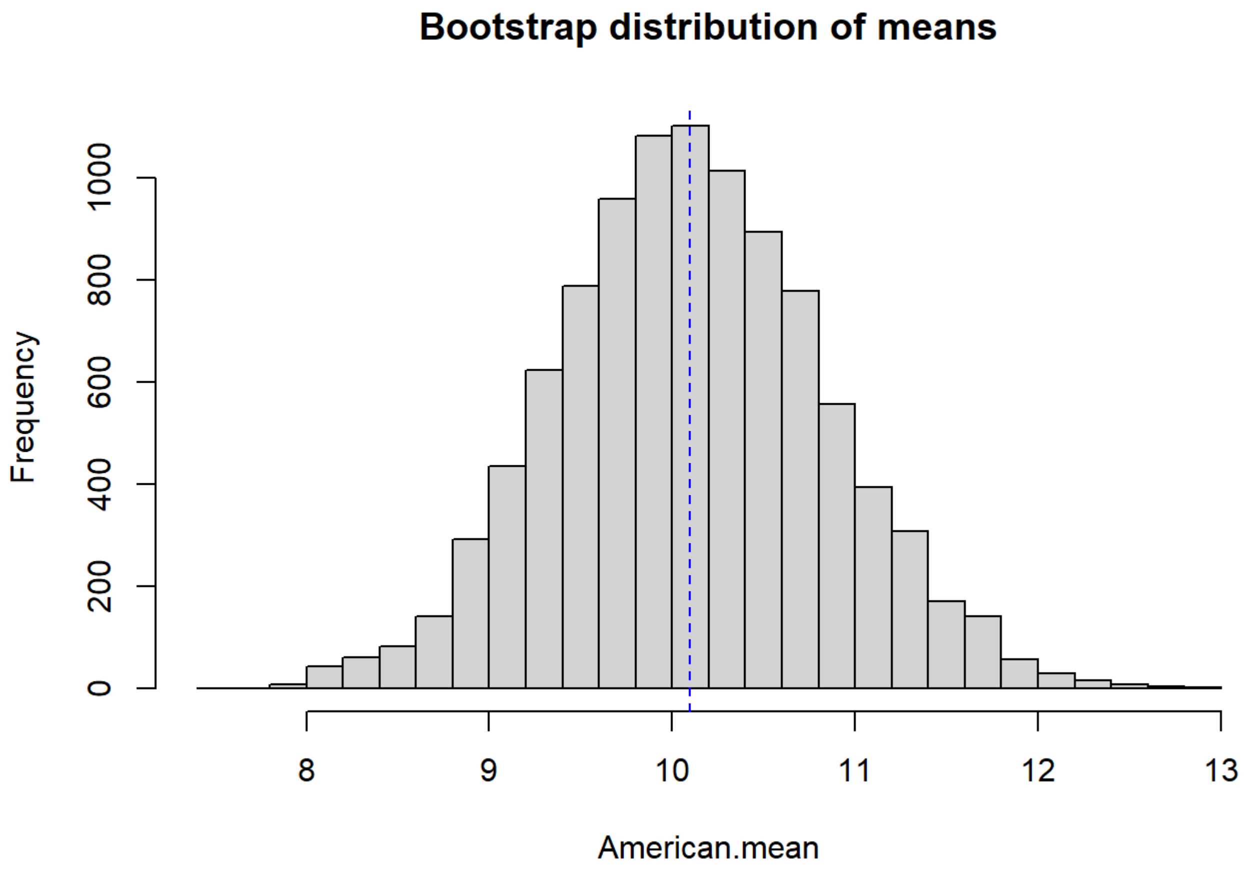
```
(n <- length(AmericanDelay))
```

```
## [1] 2906
```

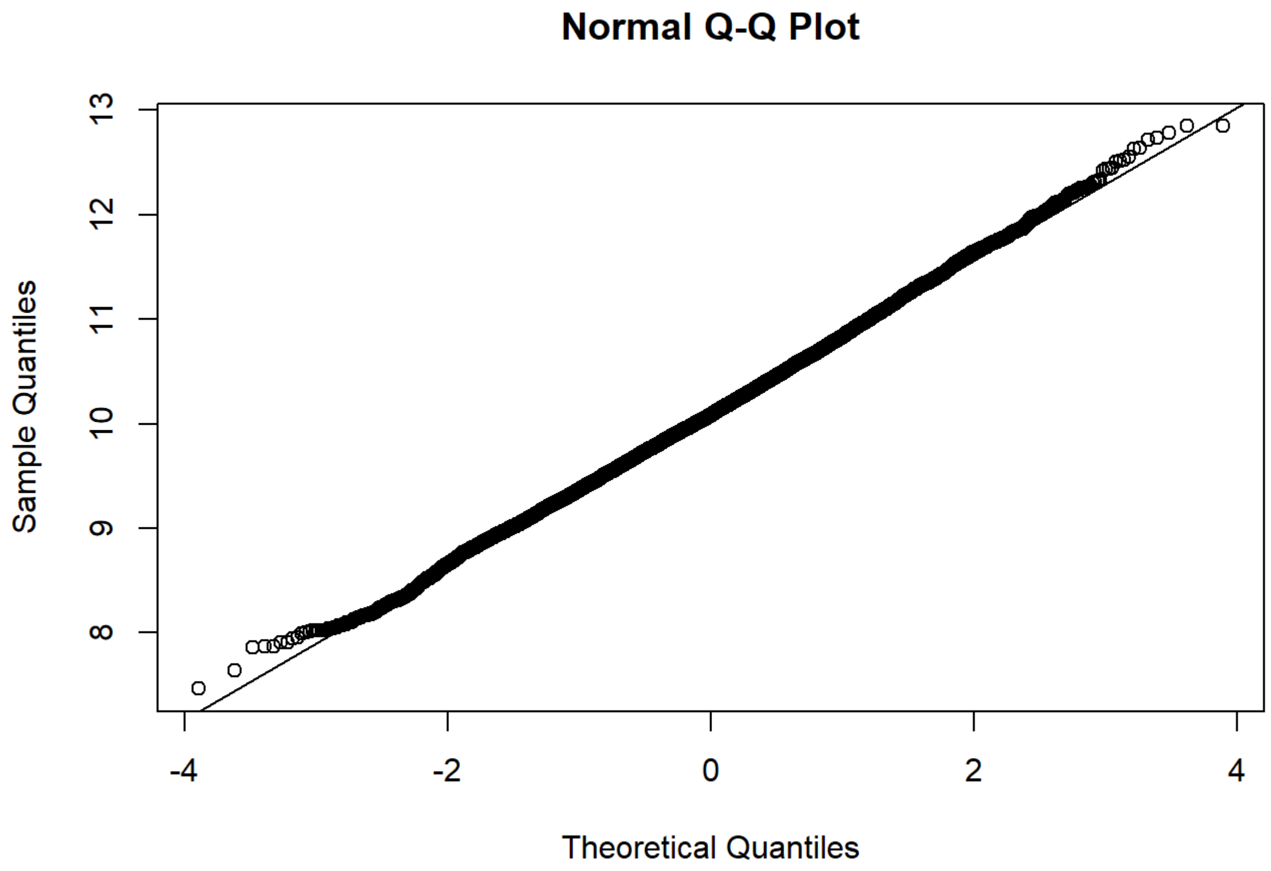
```
N <- 10^4
American.mean <- numeric(N)
for (i in 1:N)
 {
   x <- sample(AmericanDelay, n, replace = TRUE)
   American.mean[i] <- mean(x)
 }
hist(American.mean, main = "Bootstrap distribution of means")
abline(v = mean(AmericanDelay), col = "blue", lty = 2)
```



```
                            # vertical line at observed mean
hist(American.mean, main = "Bootstrap distribution of means", breaks = 30)
abline(v = mean(AmericanDelay), col = "blue", lty = 2)
```



```
                                  # vertical line at observed mean
qqnorm(American.mean)
qqline(American.mean)
```



```
summary(American.mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   7.467   9.604  10.082  10.102  10.593  12.844
```

```
(se <- sd(American.mean))
```

```
## [1] 0.7388852
```

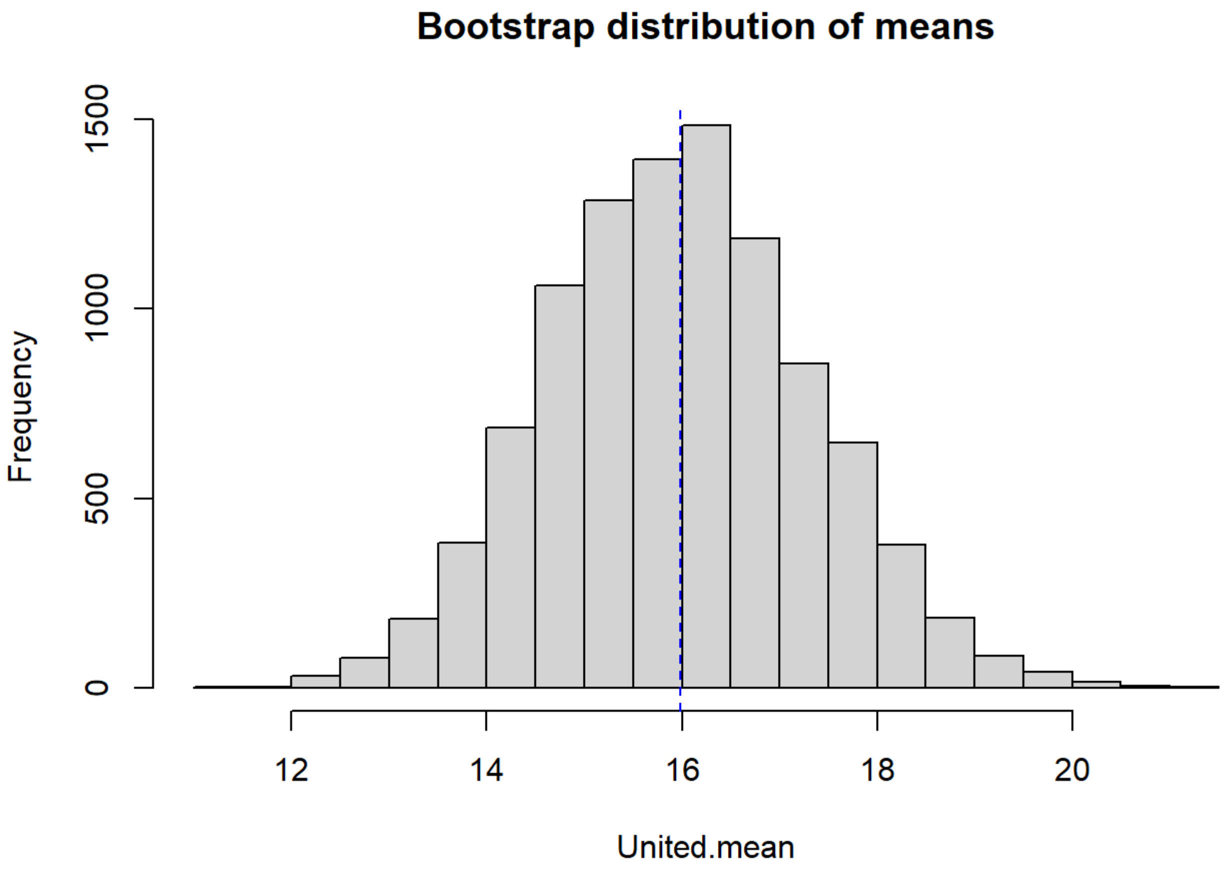
Normal distribution. Summary stats coded above

(c).

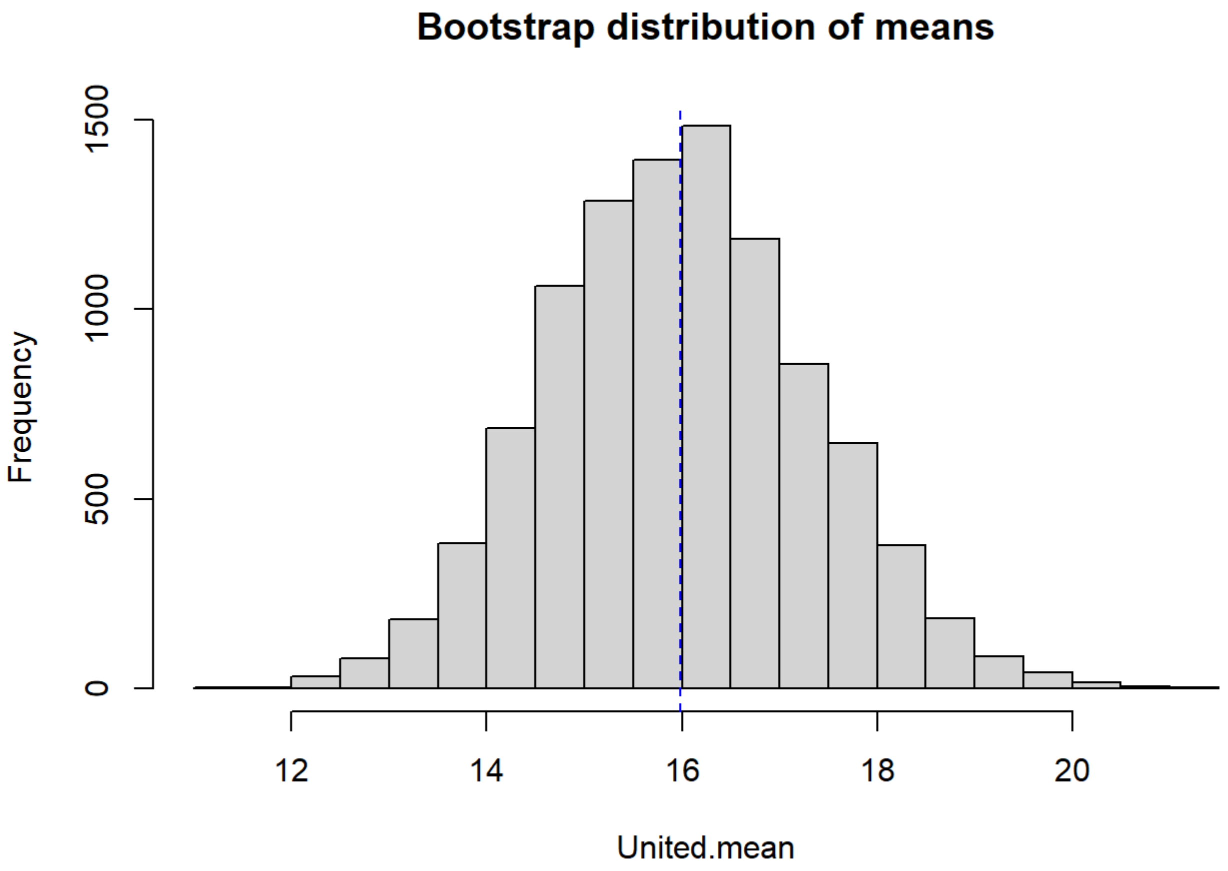
```
(m <- length(UnitedDelay))
```

```
## [1] 1123
```

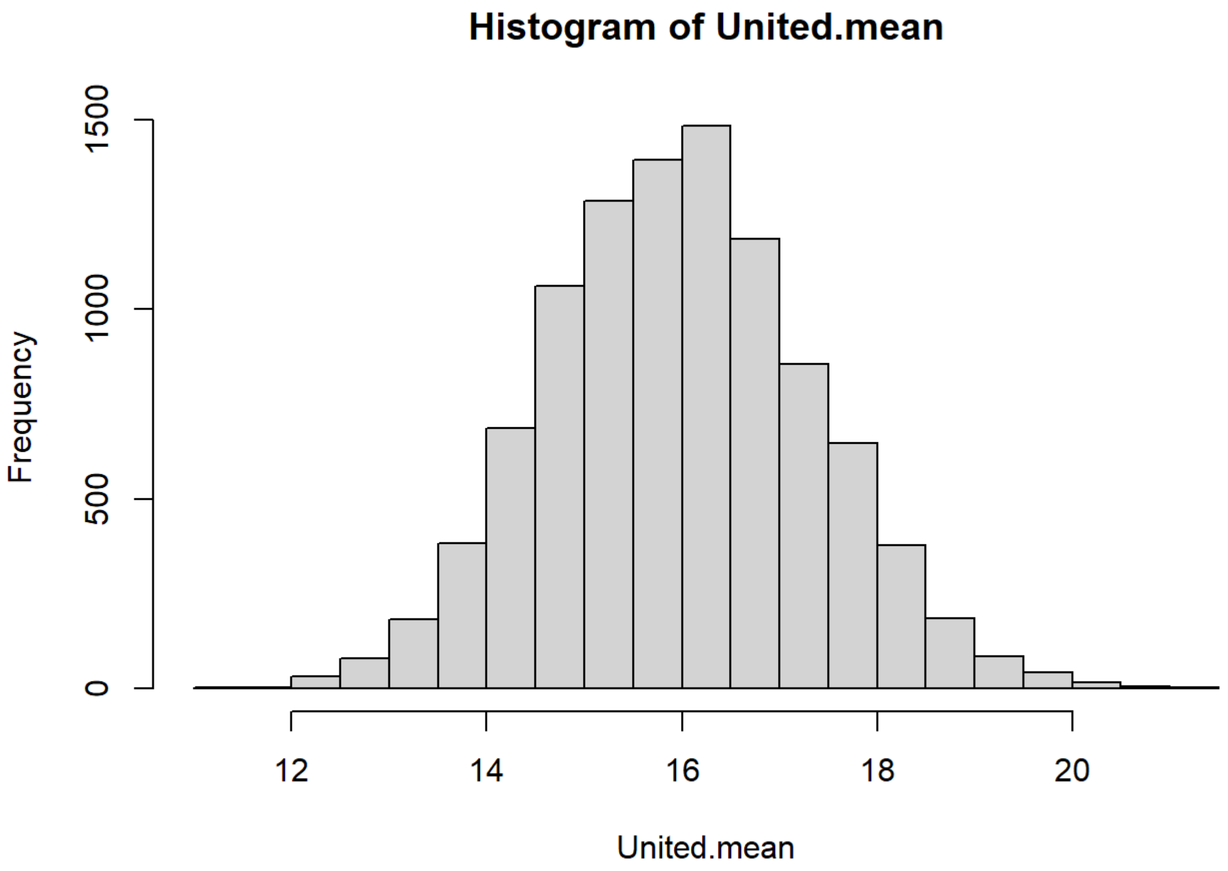
```
N <- 10^4
United.mean <- numeric(N)
for (i in 1:N)
 {
   x <- sample(UnitedDelay, m, replace = TRUE)
   United.mean[i] <- mean(x)
 }
hist(United.mean, main = "Bootstrap distribution of means")
abline(v = mean(UnitedDelay), col = "blue", lty = 2)
```



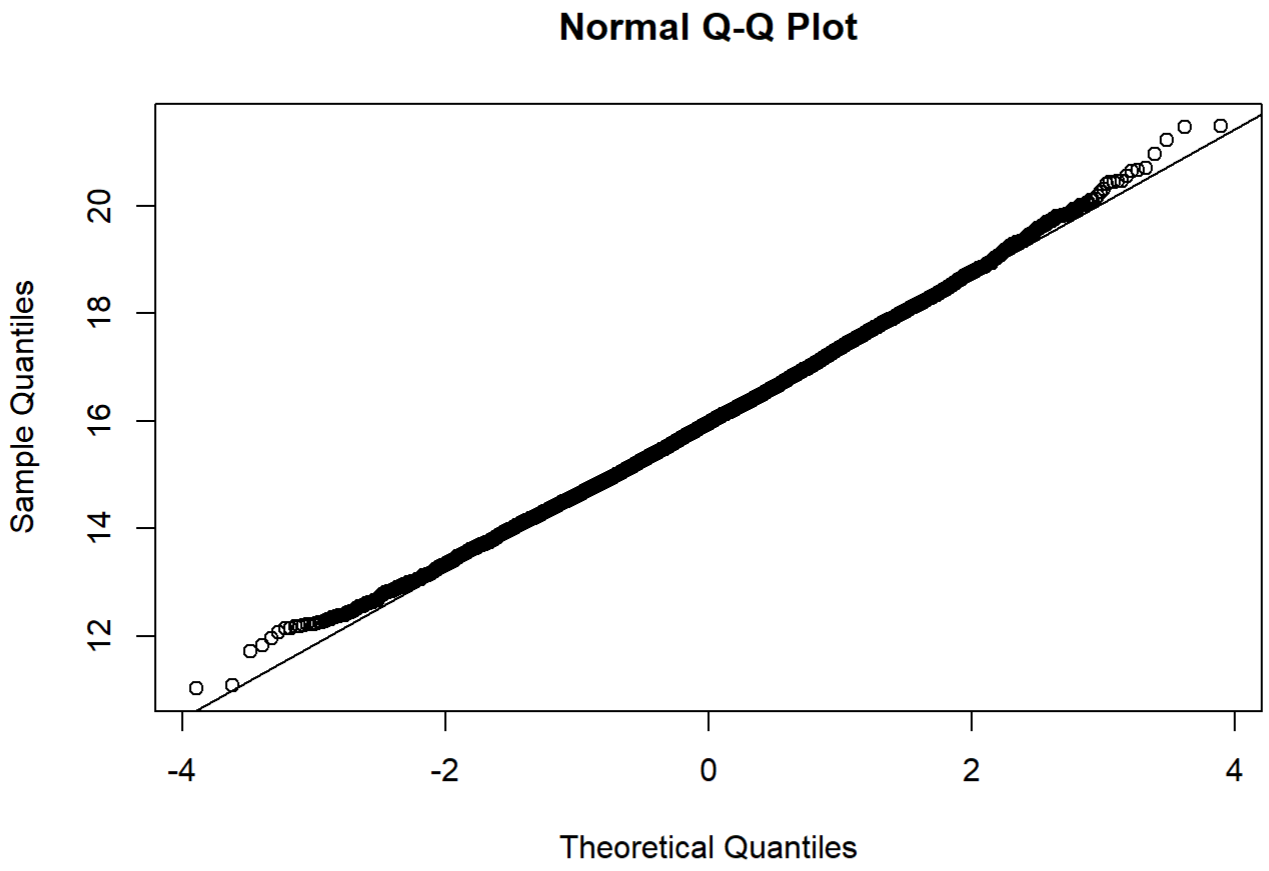
```
                            # vertical line at observed mean
hist(United.mean, main = "Bootstrap distribution of means", breaks = 30)
abline(v = mean(UnitedDelay), col = "blue", lty = 2)
```



```
                                # vertical line at observed mean
hist(United.mean)
```



```
qqnorm(United.mean)
qqline(United.mean)
```



```
summary(United.mean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   11.03   15.03   15.96   15.98   16.87   21.47
```

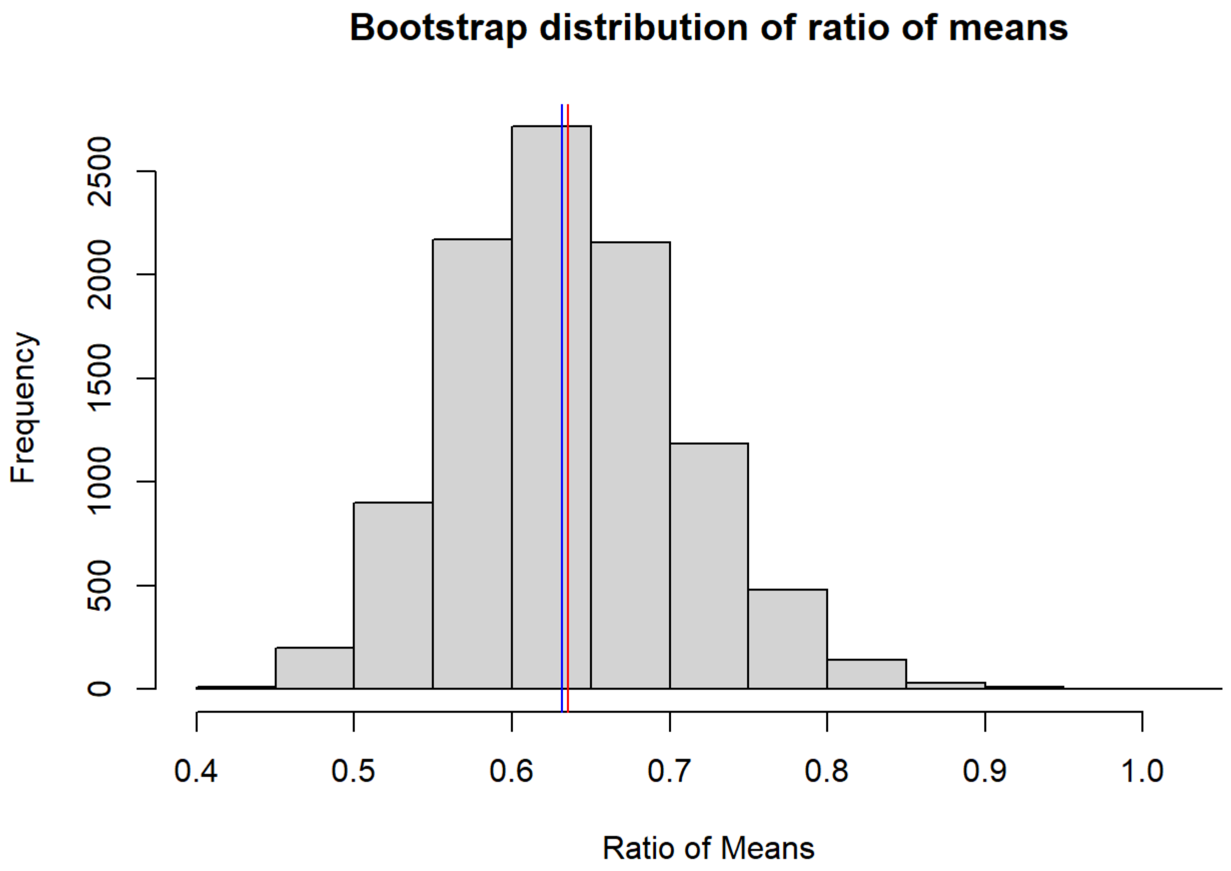
```
(se <- sd(United.mean))
```

```
## [1] 1.359441
```

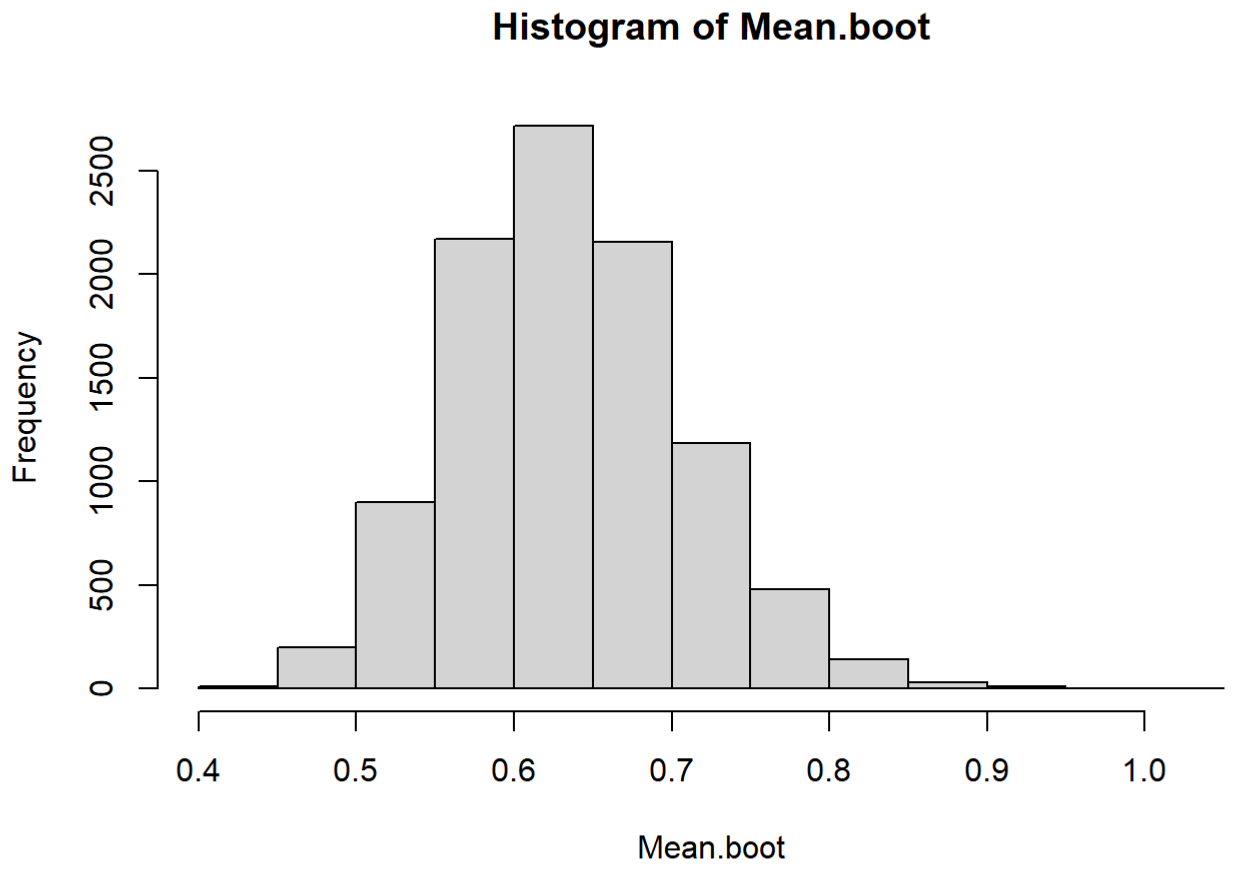
Normal distribution. Summary stats coded above

(d).

```r
B<-10^4
Mean.boot <- numeric(B)
#set.seed(100)
for (i in 1:B)
{
  American.boot <- sample(AmericanDelay, n, replace = TRUE)
  United.boot <- sample(UnitedDelay, m, replace = TRUE)
  Mean.boot[i] <- mean(American.boot)/mean(United.boot)
}
hist(Mean.boot, main = "Bootstrap distribution of ratio of means",xlab="Ratio of Means")
abline(v = mean(Mean.boot), col = "red")
abline(v = mean(AmericanDelay)/mean(UnitedDelay), col = "blue")
```



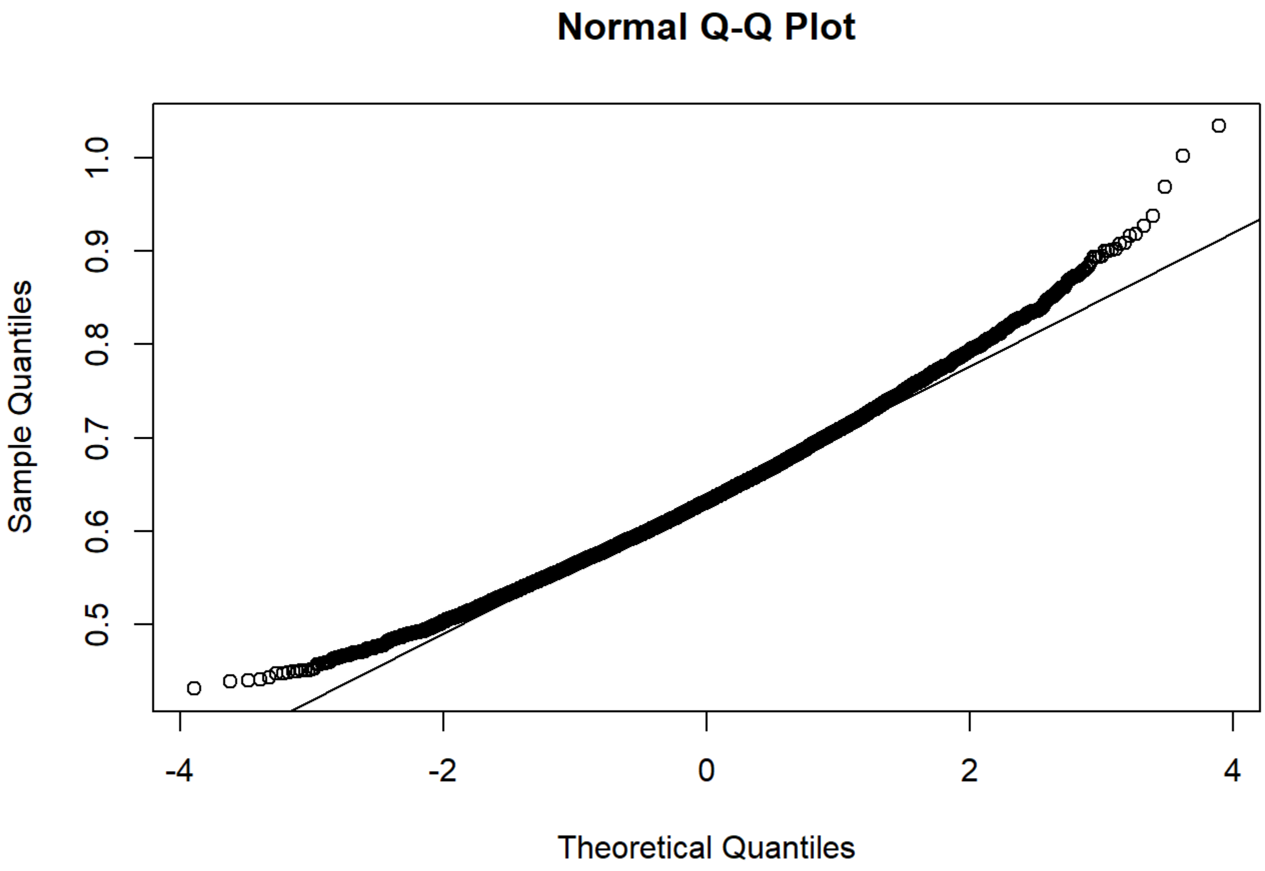
```r
hist(Mean.boot)
```



```
sd(Mean.boot)
```

```
## [1] 0.07267821
```

```
qqnorm(Mean.boot)
qqline(Mean.boot)
```



```
quantile(Mean.boot, c(.025, .975))
```

```
##      2.5%     97.5% 
## 0.5053343 0.7898412
```

Find that it is slightly skewed to the right. With 95% confidence, the mean of American flights delayed is 0.5-0.79 times that of the mean of United flights delayed

(e).

```
bias<-mean(Mean.boot)-(mean(AmericanDelay)/mean(UnitedDelay))
bias/sd(Mean.boot)
```

```
## [1] 0.05490186
```

The bootstrap estimate of error is 0.064, which represents 6.4% of the bootstrap standard error.

(f). Yes, the observations are independent of each other, as the means of each carrier’s delays follows a normal distribution.