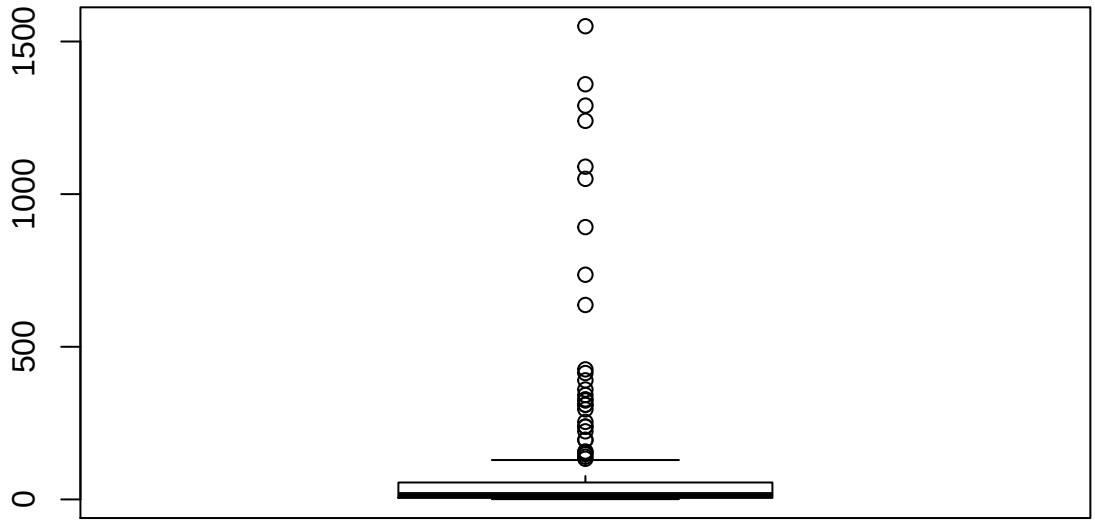
HW7

Yigao Li

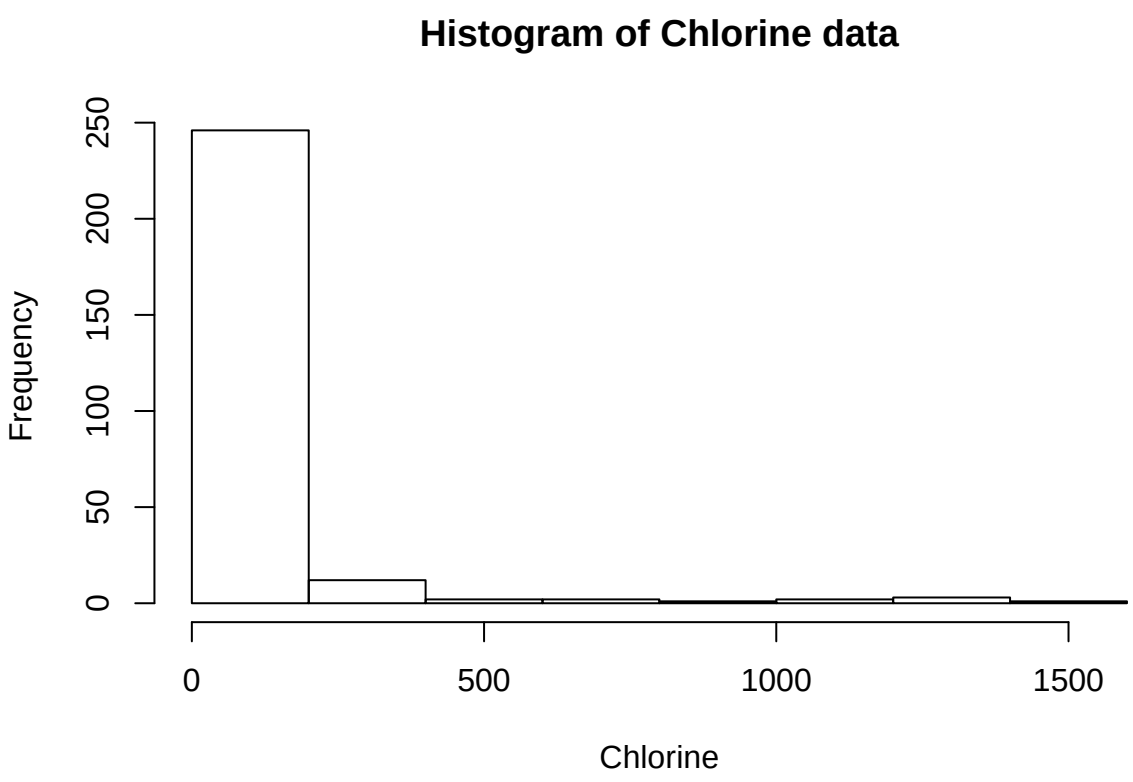
November 12, 2017

Problem 0

```
bangladesh <- read.csv("D:/Courses/ANLY 511/Bangladesh.csv")
boxplot(bangladesh$Chlorine)
```




```
hist(bangladesh$Chlorine, main = "Histogram of Chlorine data", xlab = "Chlorine")
```



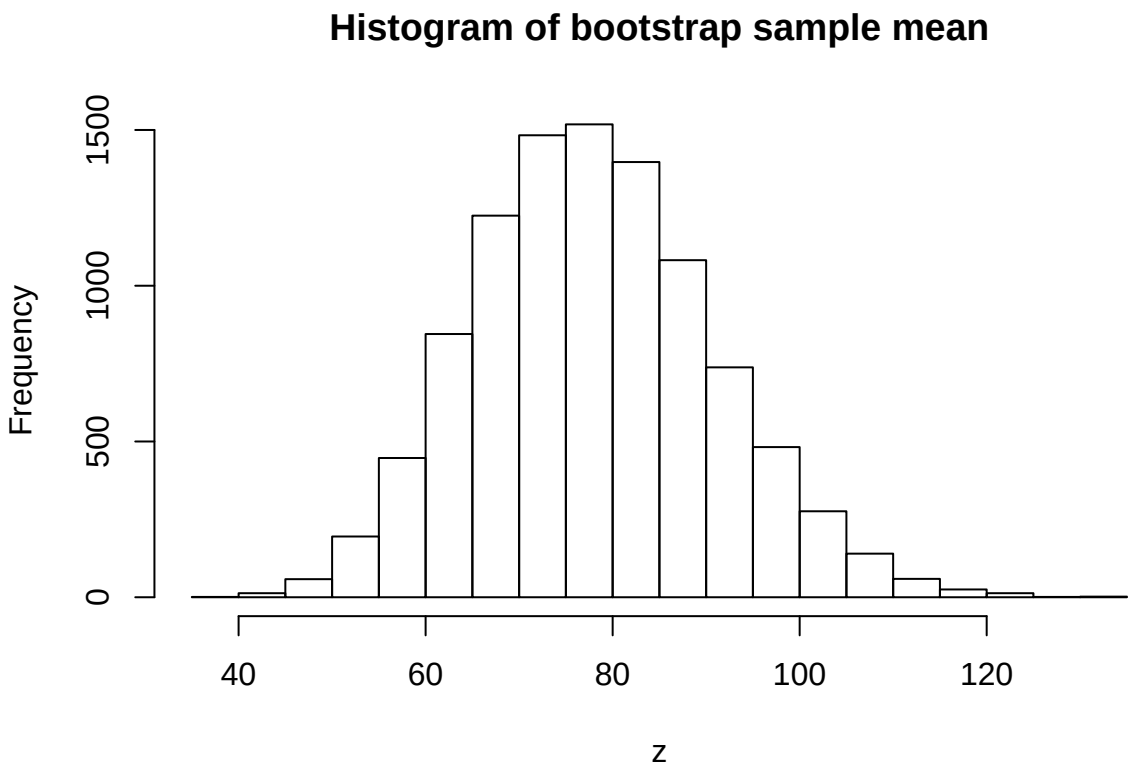

```
chlorine <- bangladesh$Chlorine[!is.na(bangladesh$Chlorine)]
cat("(a) From the boxplot and histogram, Chlorine approximately follows Exponential Distribution")
```

```
## (a) From the boxplot and histogram, Chlorine approximately follows Exponential Distribution
```

```
cat("with mean", mean(chlorine))
```

```
## with mean 78.08401
```

```
n <- length(chlorine)
N <- 10000
z <- replicate(N, mean(sample(chlorine, n, replace = TRUE)))
hist(z, main = "Histogram of bootstrap sample mean")
```




```
#test.stat <- qnorm(0.95)
test.stat <- qt(0.95, df = N-1)
lower.bound <- mean(z) - test.stat*sd(z)
upper.bound <- mean(z) + test.stat*sd(z)
cat("(b) 90% confidence interval for the mean is [", lower.bound, " , ", upper.bound, "]", sep = '')
```

```
## (b) 90% confidence interval for the mean is [56.96495 , 99.06011]
```

Problem 1

```
titani <- read.csv("D:/Courses/ANLY 511/Titanic.csv")
age.victim <- titani$Age[titani$Survived == 0]
age.survivor <- titani$Age[titani$Survived == 1]
nv <- length(age.victim)
ns <- length(age.survivor)
N = 10000
age.victim.bootstrap <- replicate(N, median(sample(age.victim, nv, replace = TRUE)))
age.survivor.bootstrap <- replicate(N, median(sample(age.survivor, ns, replace = TRUE)))
varv <- var(age.victim.bootstrap)/N
vars <- var(age.survivor.bootstrap)/N
test.stat <- (mean(age.victim.bootstrap) - mean(age.survivor.bootstrap))/sqrt(varv + vars)
degree <- (varv^2 + vars^2)^2/(varv^2/(N-1) + vars^2/(N-1))
crit <- qt(0.975, df = degree)
test.stat > crit
```

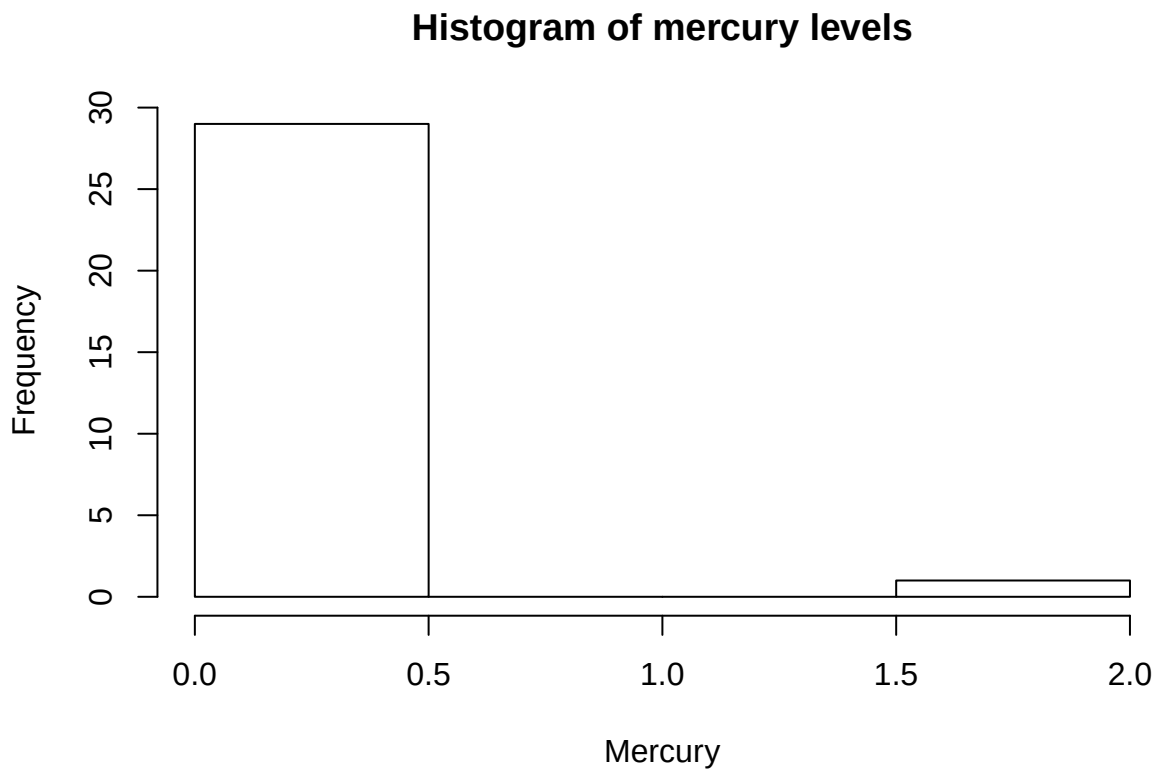
```
## [1] FALSE
```

Hypothesis Testing
H_0: Median ages of the victims and survivors are the same.
H_1: Median ages of the victims and survivors are different.
Deal with Titanic data and split age into 2 data for victims and survivors separately
Use Bootstrap to get 10000 samples of each group and calculate their median differences
To test our hypothesis, we use Two Sample T-test.
Since test statistics is less than critical value at significance level 0.05, we fail to reject H_0. There's not enough evidence to show that median ages of the victims and survivors are the same.

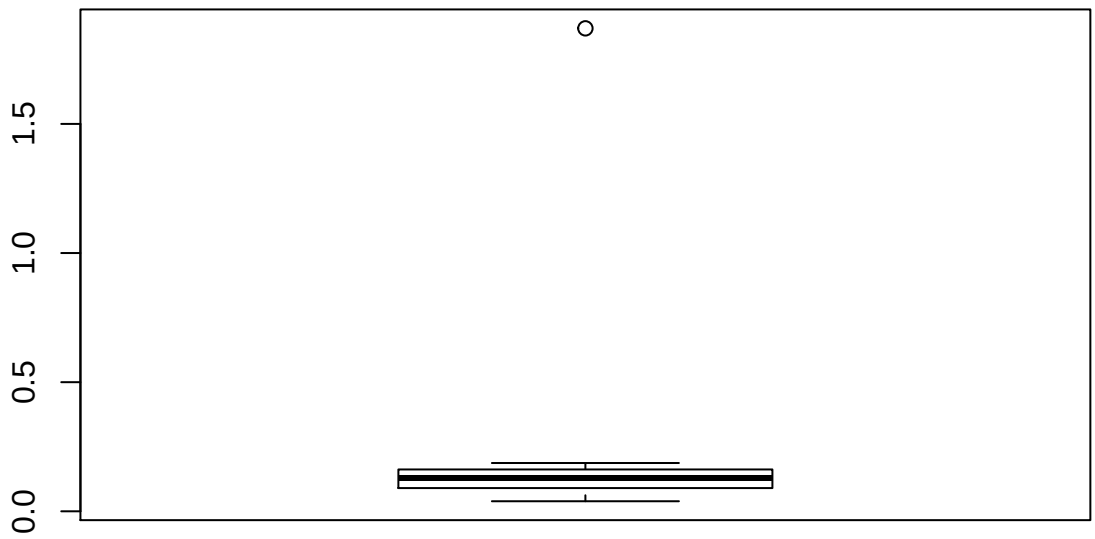
Problem 2

(a)

```
fishmercury <- read.csv("D:/Courses/ANLY 511/FishMercury.csv")
hist(fishmercury$Mercury, main = "Histogram of mercury levels", xlab = "Mercury")
```




```
boxplot(fishmercury$Mercury)
```



From plots, we observe that except for 1 observation greater than 1.5, all other observations are between 0 and 0.2.

(b)

```
mercury <- fishmercury$Mercury
n <- length(mercury)
N <- 10000
mercury.bootstrap <- replicate(N, mean(sample(mercury, n, replace = TRUE)))
standard.error <- sd(mercury.bootstrap)
cat("Bootstrap standard error is", standard.error)
```

```
## Bootstrap standard error is 0.05792895
```

```
test.stat <- qt(0.975, df = N-1)
lower.bound <- mean(mercury) - test.stat*standard.error
upper.bound <- mean(mercury) + test.stat*standard.error
cat("95% confidence interval for the mean is [", lower.bound, " , ", upper.bound, "]", sep = '')
```

```
## 95% confidence interval for the mean is [0.06831426 , 0.2954191]
```

(c)

```
mercury <- mercury[mercury < 1]
n <- length(mercury)
N <- 10000
mercury.bootstrap <- replicate(N, mean(sample(mercury, n, replace = TRUE)))
standard.error <- sd(mercury.bootstrap)
cat("Bootstrap standard error is", standard.error)
```

```
## Bootstrap standard error is 0.00788731
```

```
test.stat <- qt(0.975, df = N-1)
lower.bound <- mean(mercury) - test.stat*standard.error
upper.bound <- mean(mercury) + test.stat*standard.error
cat("95% confidence interval for the mean is [", lower.bound, " , ", upper.bound, "]", sep = '')
```

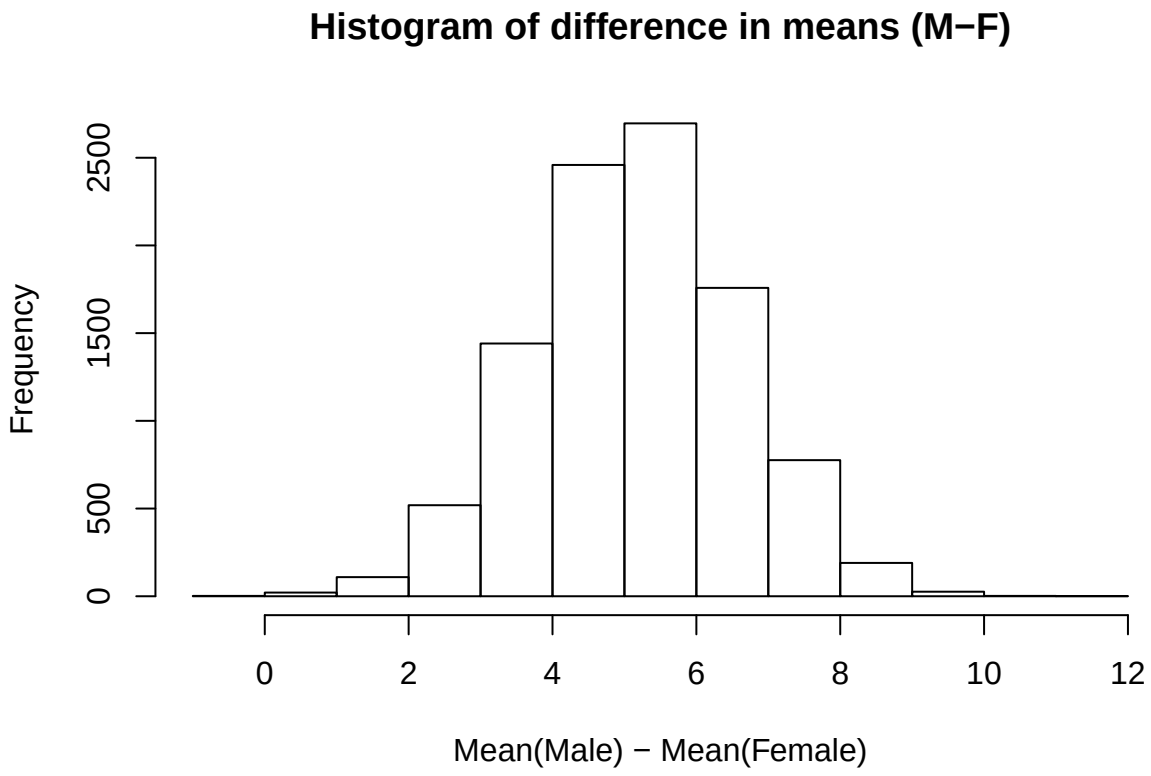
```
## 95% confidence interval for the mean is [0.1081945 , 0.1391159]
```

(d) After removing the outlier, standard error decreased from approximately 0.060 to 0.008, and confidence interval gets narrorwer.

Problem 3

(a)

```
beerwings <- read.csv("D:/Courses/ANLY 511/Beerwings.csv")
wingsM <- beerwings$Hotwings[beerwings$Gender == "M"]
wingsF <- beerwings$Hotwings[beerwings$Gender == "F"]
nM <- length(wingsM)
nF <- length(wingsF)
N <- 10000
wingsM.bootstrap <- replicate(N, mean(sample(wingsM, nM, replace = TRUE)))
wingsF.bootstrap <- replicate(N, mean(sample(wingsF, nF, replace = TRUE)))
wings.diff <- wingsM.bootstrap - wingsF.bootstrap
hist(wings.diff, main = "Histogram of difference in means (M-F)", xlab = "Mean(Male) - Mean(Female)")
```



Bootstrap distribution approximately follows normal distribution with mean between 5 and 6.

(b)

```
test.stat <- qt(0.975, df = N-1)
lower.bound <- mean(wings.diff) - test.stat*sd(wings.diff)
upper.bound <- mean(wings.diff) + test.stat*sd(wings.diff)
cat("95% confidence interval for the mean difference is [", lower.bound, " , ", upper.bound, "]",
    sep = '')
```

```
## 95% confidence interval for the mean difference is [2.396123 , 7.98329]
```

We are 95% confident that difference of men and women consuming hot wings lies in the above interval.

(c)

Bootstrap sample data inside each category with replacement, while permutation distribution permutate categories to each value without replacement.

Problem 4

```
icecream <- read.csv("D:/Courses/ANLY 511/IceCream.csv")
vanilla <- icecream$VanillaCalories
choco <- icecream$ChocolateCalories
n <- dim(icecream)[1]
N <- 10000
vanilla.bootstrap <- replicate(N, mean(sample(vanilla, n, replace = TRUE)))
choco.bootstrap <- replicate(N, mean(sample(choco, n, replace = TRUE)))
calorie.diff <- choco.bootstrap - vanilla.bootstrap
test.stat <- qt(0.975, df = N-1)
lower.bound <- mean(calorie.diff) - test.stat*sd(calorie.diff)
upper.bound <- mean(calorie.diff) + test.stat*sd(calorie.diff)
cat("95% confidence interval for the mean difference is [", lower.bound, " , ", upper.bound, "]",
    sep = '')
```

```
## 95% confidence interval for the mean difference is [-19.02618 , 33.52246]
```

Since 0 is in confidence interval, we fail to reject H_0. There is not enough evidence to show that vanilla and chocolate ice creams have the same amount of calories.

Problem 5

(a)

```
girls <- read.csv("D:/Courses/ANLY 511/Girls2004.csv")
weight.wy <- girls$Weight[girls$State == "WY"]
weight.ak <- girls$Weight[girls$State == "AK"]
print("Summary statistics for weight of baby girls born in Wyoming")
```

```
## [1] "Summary statistics for weight of baby girls born in Wyoming"
```

```
summary(weight.wy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2212    2934    3278    3208    3515    3995
```

```r
print("Summary statistics for weight of baby girls born in Arkansas")
```

```
## [1] "Summary statistics for weight of baby girls born in Arkansas"
```

```r
summary(weight.ak)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2182    3170    3558    3516    3926    4592
```

From the summary statistics, in general, baby girls born in Wyoming weight less than those born in Arkansas.

(b)

```r
nWY <- length(weight.wy)
nAK <- length(weight.ak)
N <- 10000
wy.bootstrap <- replicate(N, mean(sample(weight.wy, nWY, replace = TRUE)))
ak.bootstrap <- replicate(N, mean(sample(weight.ak, nAK, replace = TRUE)))
weight.diff <- ak.bootstrap - wy.bootstrap
test.stat <- qt(0.975, df = N-1)
lower.bound <- mean(weight.diff) - test.stat*sd(weight.diff)
upper.bound <- mean(weight.diff) + test.stat*sd(weight.diff)
cat("95% confidence interval for the mean difference is [", lower.bound, " , ", upper.bound, "]",
    sep = '')
```

```
## 95% confidence interval for the mean difference is [89.88007 , 526.5588]
```

We are 95% confident that mean difference of weights between baby girls born in Wyoming and Arkansas is in the above interval.

(c)

```r
theta.hat <- abs(mean(weight.ak) - mean(weight.wy))
bias <- abs(theta.hat - mean(weight.diff))
cat("Bootstrap estimate of the bias is", bias)
```

```
## Bootstrap estimate of the bias is 0.2305525
```

```r
cat("It represents", bias/sd(weight.diff), "of the bootstrap standard error.")
```

```
## It represents 0.002069848 of the bootstrap standard error.
```

(d)

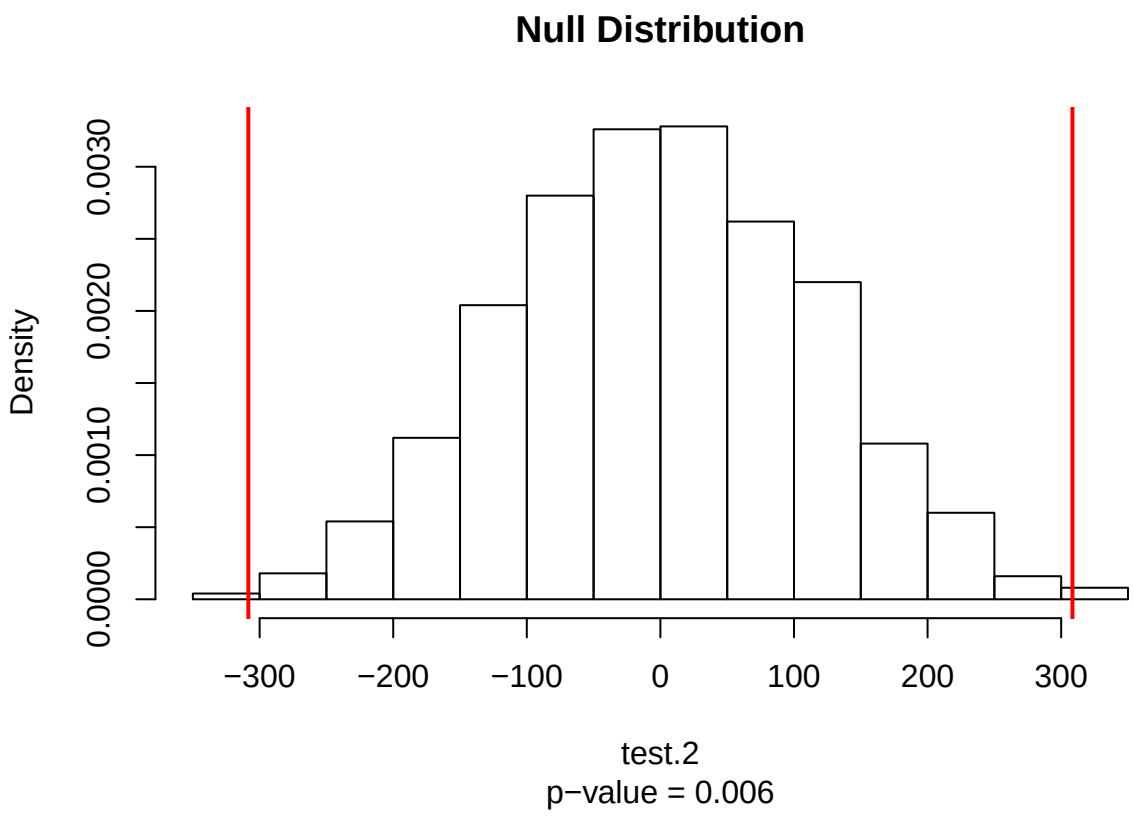
```r
mytest.2 <- function(mydf){
  agg <- aggregate(Weight ~ State, data = mydf, FUN = mean)
  return(agg$Weight[1] - agg$Weight[2])
}

permute.sample.2 <- function(mydf){
  n <- dim(mydf)[1]
  mydf$State <- mydf$State[sample(n, n, replace = F)]
  return(mytest.2(mydf))
}

girls.permute <- girls
N <- 1000
test.2 <- replicate(N, permute.sample.2(girls.permute))
cat("Difference in mean weights is", mean(test.2))
```

```
## Difference in mean weights is 0.3611
```

```
hist(test.2, main = "Null Distribution", prob = T,
     sub = paste("p-value =", mean(abs(test.2) > mytest.2(girls))))
abline(v = mytest.2(girls), col = 2, lwd = 2)
abline(v = -mytest.2(girls), col = 2, lwd = 2)
```



(e)

Because p-value is less than significance level 0.05, we reject H_0. It is statistically significant that there is no difference between weight of baby girls born in Wyoming and Arkansas.

Problem 6

(a)

```
flightdelays <- read.csv("D:/Courses/ANLY 511/FlightDelays.csv")
flightdelays.ua <- flightdelays$Delay[flightdelays$Carrier == "UA"]
flightdelays.aa <- flightdelays$Delay[flightdelays$Carrier == "AA"]
print("Summary statistics for flight delay lengths for UA flights")
```

```
## [1] "Summary statistics for flight delay lengths for UA flights"
```

```
summary(flightdelays.ua)
```

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

```r
print("Summary statistics for flight delay lengths for AA flights")
```

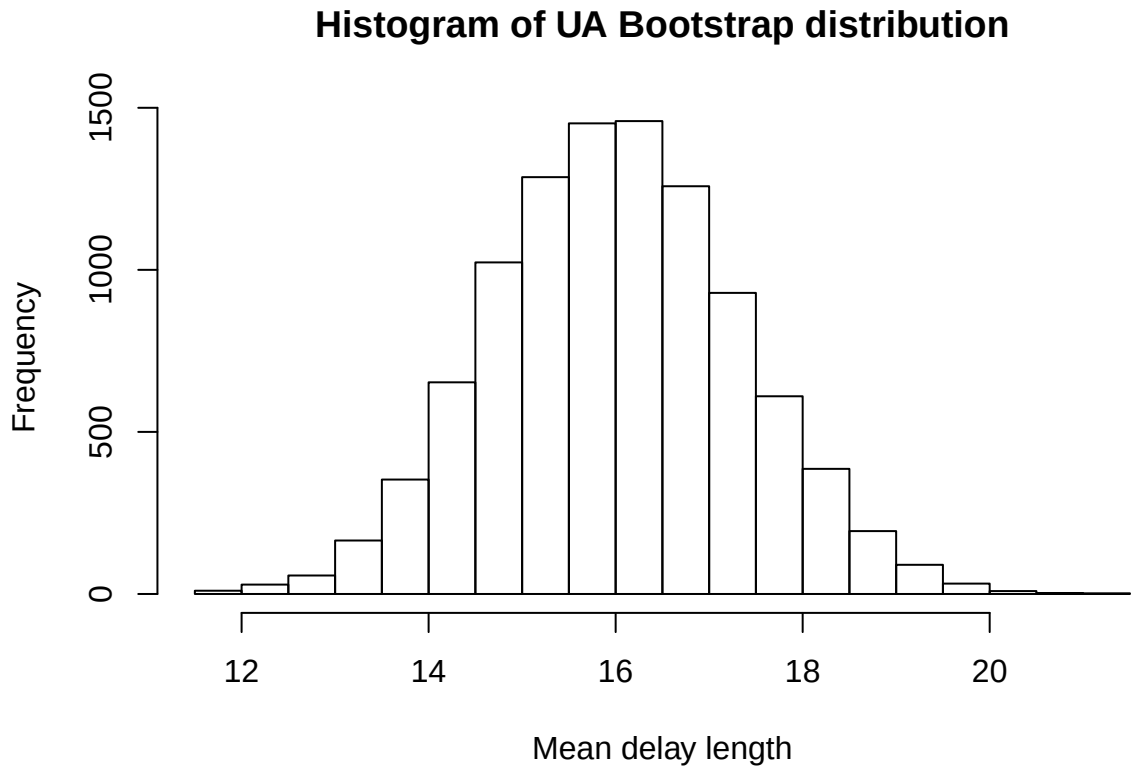
```
## [1] "Summary statistics for flight delay lengths for AA flights"
```

```r
summary(flightdelays.aa)
```

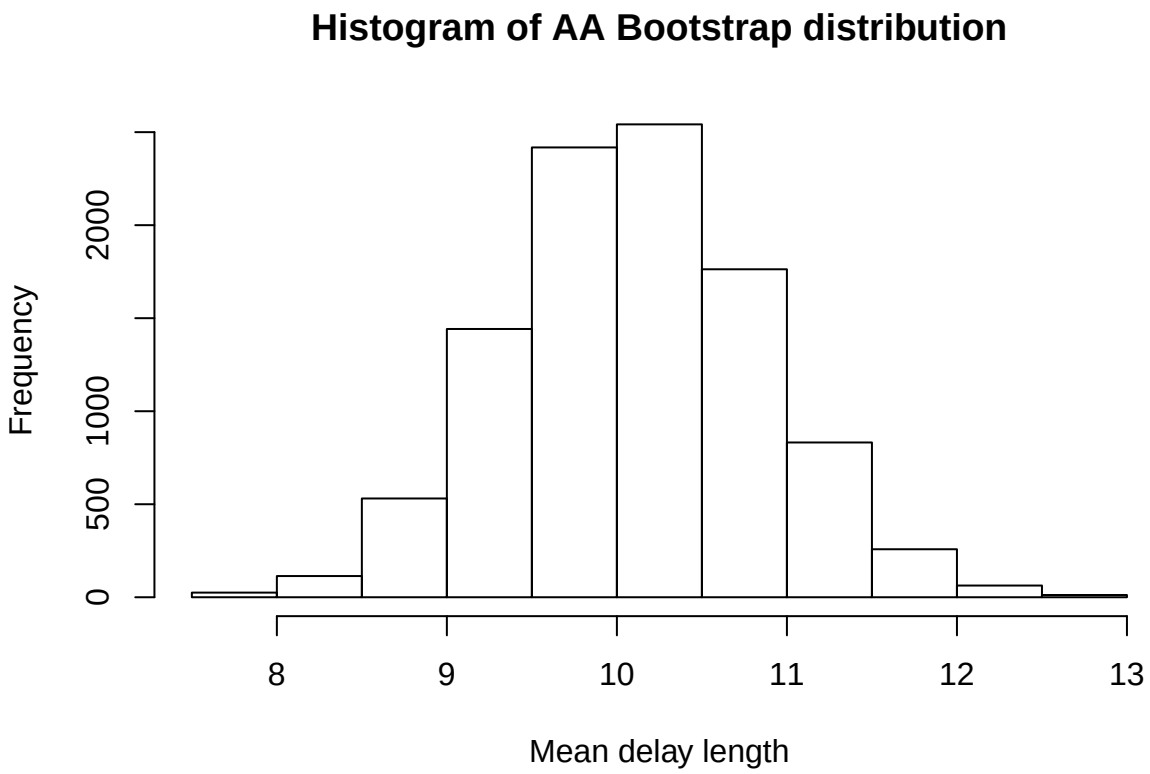
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -19.0    -6.0    -3.0    10.1     4.0   693.0
```

(b)

```r
nUA <- length(flightdelays.ua)
nAA <- length(flightdelays.aa)
N <- 10000
ua.bootstrap <- replicate(N, mean(sample(flightdelays.ua, nUA, replace = TRUE)))
aa.bootstrap <- replicate(N, mean(sample(flightdelays.aa, nAA, replace = TRUE)))
hist(ua.bootstrap, main = "Histogram of UA Bootstrap distribution", xlab = "Mean delay length")
```



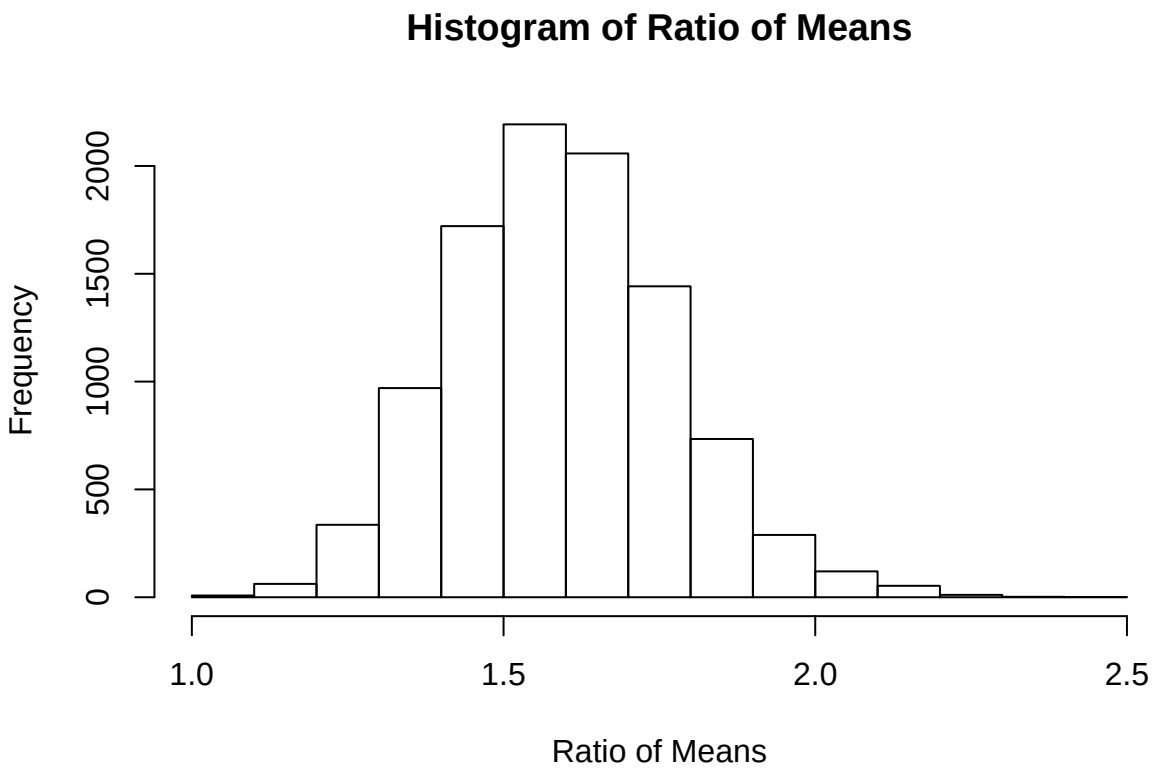

```r
hist(aa.bootstrap, main = "Histogram of AA Bootstrap distribution", xlab = "Mean delay length")
```



Bootstrap distribution for UA is approximately normal distribution with mean around 16 and for AA is also normally distributed with mean around 10.

(c)

```
ratioofmeans <- ua.bootstrap/aa.bootstrap
hist(ratioofmeans, main = "Histogram of Ratio of Means", xlab = "Ratio of Means")
```



Ratio of mean approximately follows Cauchy distribution ($\frac{normal}{normal}$). It's a bell shape curve.

(d)

```
test.stat <- qt(0.975, df = N-1)
lower.bound <- mean(ratioofmeans) - test.stat*sd(ratioofmeans)
upper.bound <- mean(ratioofmeans) + test.stat*sd(ratioofmeans)
cat("95% confidence interval for the ratio of means is [", lower.bound, " , ", upper.bound, "]",
    sep = '')
```

```
## 95% confidence interval for the ratio of means is [1.243833 , 1.943362]
```

We are 95% confident that the ratio of means is in the above interval.

(e)

```
theta.hat <- mean(flightdelays.ua)/mean(flightdelays.aa)
bias <- theta.hat - mean(ratioofmeans)
cat("Bootstrap estimate of the bias is", bias)
```

```
## Bootstrap estimate of the bias is -0.01070411
```

```
cat("\nIt represents", bias/sd(ratioofmeans), "of the bootstrap standard error.")
```

```
##
## It represents -0.05998948 of the bootstrap standard error.
```

(f)

No, observations are not perfectly independent. Because some flights share the same airplane, once a flight was delayed, its continuing flight's delay will depend on its previous flight.

Problem 7

(a)

```
arsenic <- bangladesh$Arsenic
summary(arsenic)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.5     6.0    22.0   125.3   109.0  2400.0
```

```
cat("90th percentile is", quantile(arsenic, probs = 0.9))
```

```
## 90th percentile is 270
```

(b)

```
n <- length(arsenic)
N <- 10000
arsenic.bootstrap <- replicate(N, median(sample(arsenic, n, replace = TRUE)))
theta.hat <- median(arsenic)
bias <- theta.hat - mean(arsenic.bootstrap)
cat("Bias of the median is", bias)
```

```
## Bias of the median is -1.54898
```

(c)

```
arsenic.bootstrap.90 <- replicate(N, quantile(sample(arsenic, n, replace = TRUE), probs = 0.9))
theta.hat <- quantile(arsenic, probs = 0.9)
bias <- theta.hat - mean(arsenic.bootstrap.90)
cat("Bias of the 90th percentile is", bias)
```

```
## Bias of the 90th percentile is -2.4608
```