ANLY 511: Assignment #7

Brody Vogel

11/8/2017

Preparation

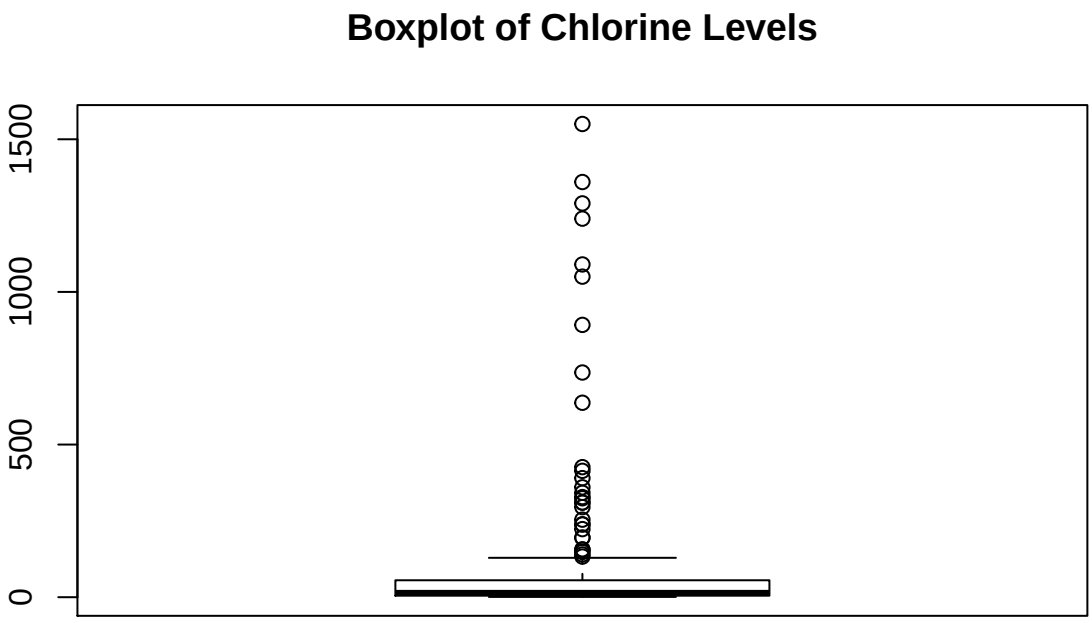
```
set.seed(1234)
```

First Problem

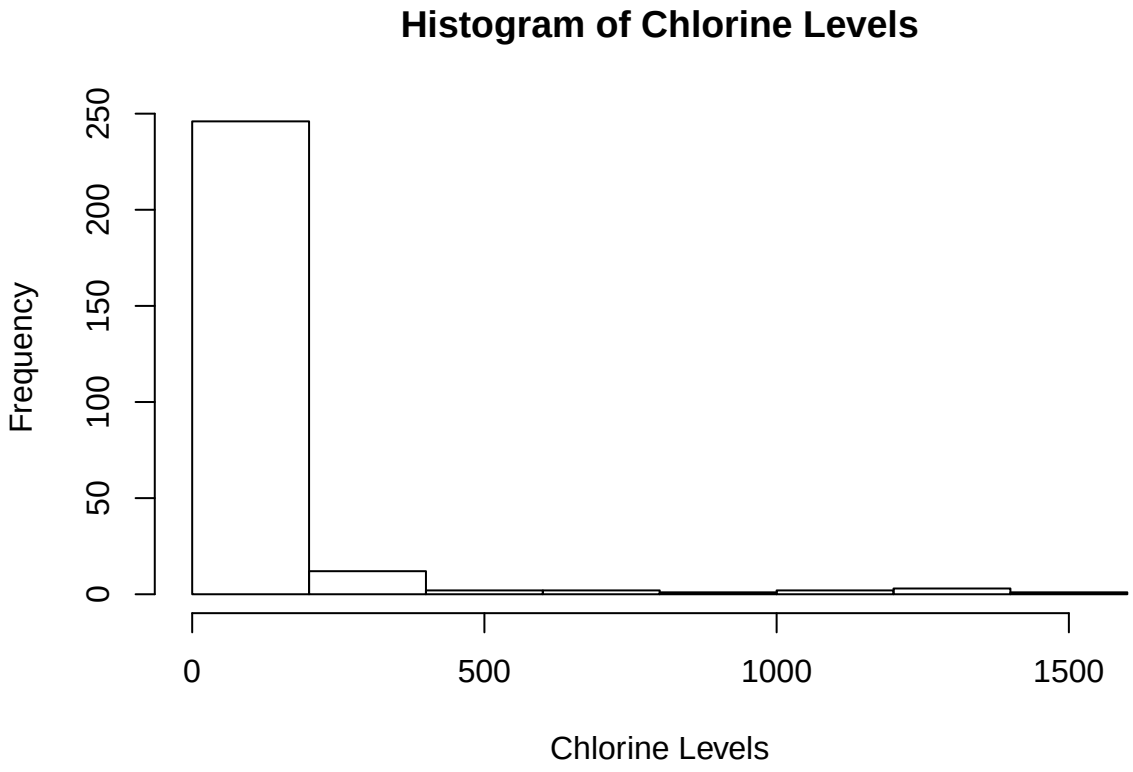
a) It looks like most of the values are in the ≈ 0 - 150 range, but there is also a significant percentage of the values in the ≈ 200 - 450 range, and a non-negligible percentage in the 500 - 1500 range; there are a lot of extreme values. It's certainly not normally distributed - may be from an exponential distribution.

b) The 90% confidence interval for the mean is [58.46, 100.51], which demonstrates the massive spread of the data.

```
### a)
bdesh <- read.csv('/Users/brodyvogel/Desktop/Data/Bangladesh.csv')
boxplot(bdesh$Chlorine, main = 'Boxplot of Chlorine Levels')
```




```
hist(bdesh$Chlorine, main = 'Histogram of Chlorine Levels', xlab = 'Chlorine Levels')
```




```
### b)
chlor <- bdesh$Chlorine
samps <- replicate(10000, mean(sample(chlor, length(chlor), replace = TRUE), na.rm = TRUE))
quantile(samps, probs = c(.05, .95))
```

```
##       5%      95%
## 58.10800 99.85186
```

Problem 1

H_0 : The median ages of the victims and survivors are not different. H_a : The median ages of the victims and survivors are different.

I first calculated 10,000 bootstrap samples of the difference in medians of samples - with replacement - from the collections of ages of survivors and victims.

I then constructed confidence intervals at the .9, .95, and .975 levels for the mean using the bootstrap samples, each of which contained 0. Thus, we cannot reject H_0; that is, we cannot say that the median ages of the victims and survivors are different.

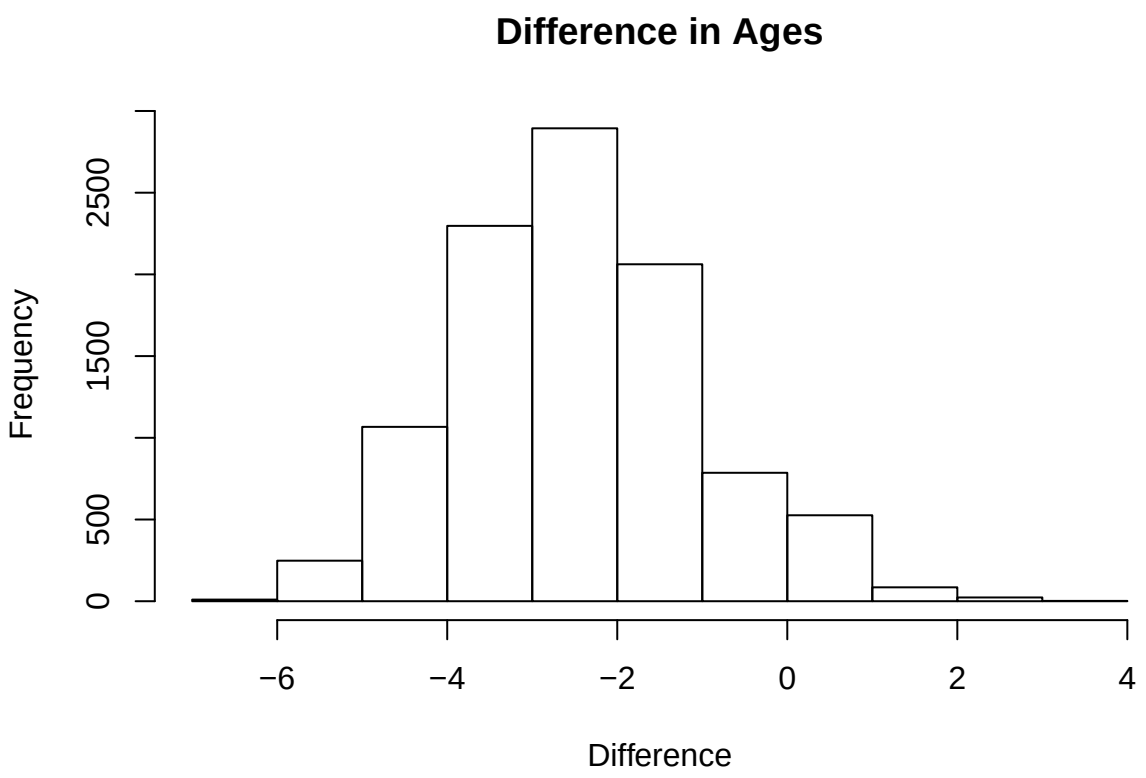
```
titc <- read.csv('/Users/brodyvogel/Desktop/Data/Titanic.csv')

survs <- titc[titc$Survived == 1, ]
victs <- titc[titc$Survived == 0, ]

survAge <- survs$Age
victsAge <- victs$Age

samp1 <- replicate(10000, median(sample(survAge, length(survAge), replace = TRUE)) - median(sample(victs
```

```r
hist(samp1, main = 'Difference in Ages', xlab = 'Difference')
```




```r
quantile(samp1, probs = c(.05, .95))
```

```
##  5% 95%
##  -4   1
```

```r
quantile(samp1, probs = c(.025, .975))
```

```
##  2.5% 97.5%
##    -5     1
```

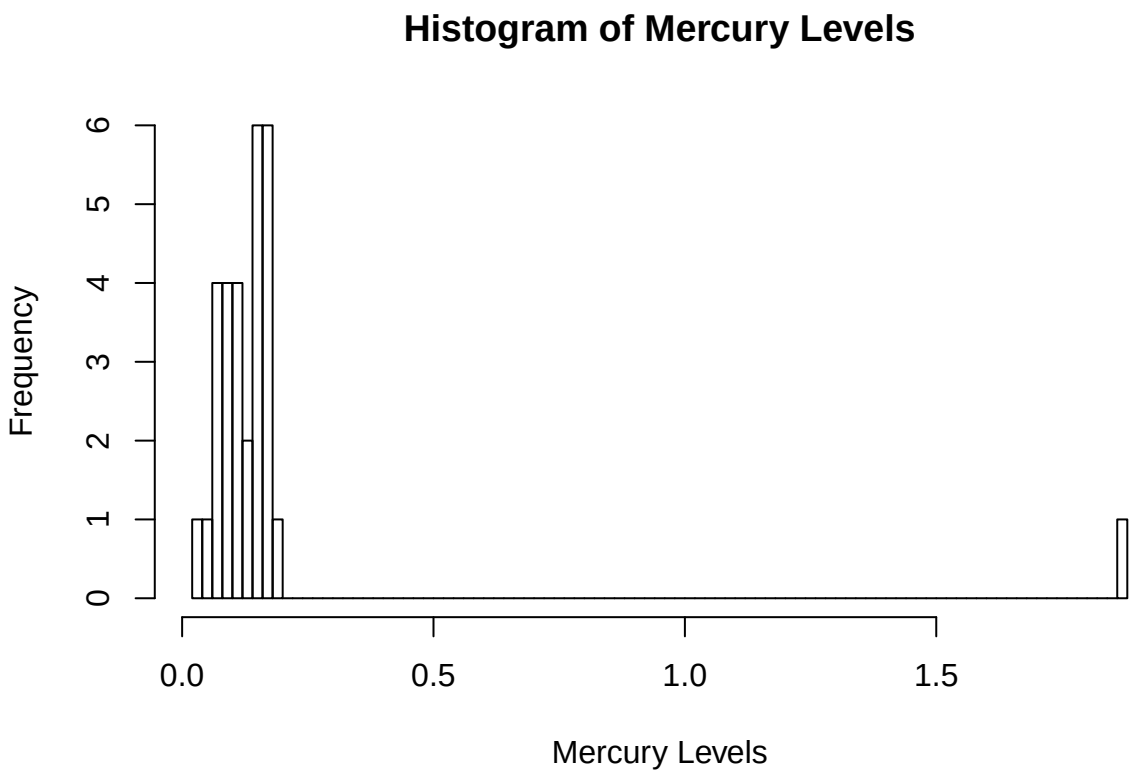
```r
quantile(samp1, probs = c(.0125, .9875))
```

```
##  1.25% 98.75%
##     -5      1
```

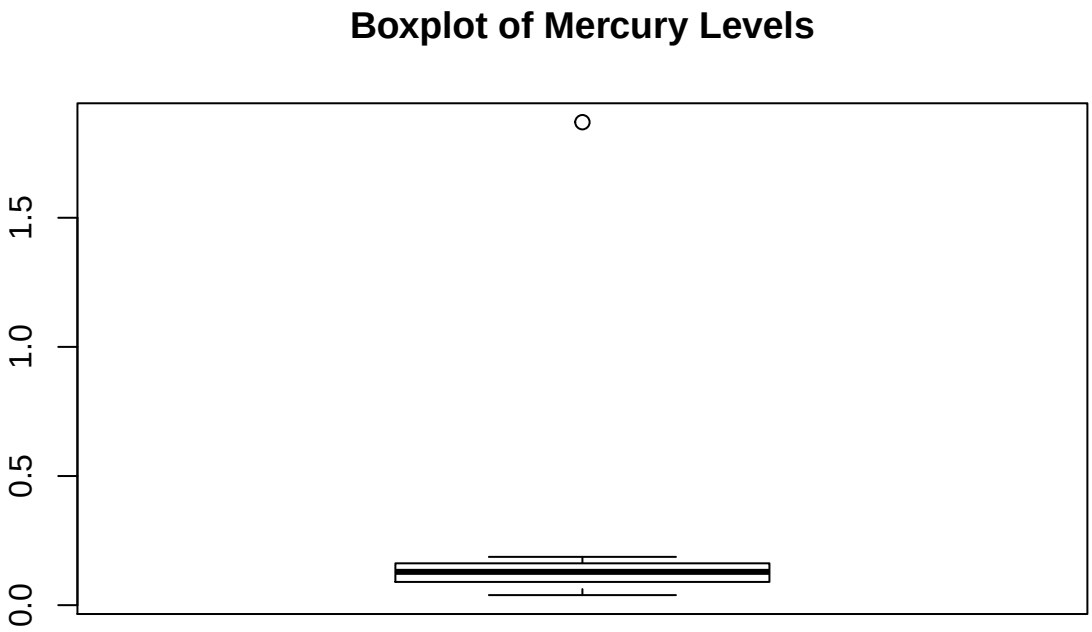
Problem 2

a) It looks like 29 of the fish had mercury levels between 0 and .5, while one poor fish fell in the 1.5 - 2 bin.

```r
fm <- read.csv('/Users/brodyvogel/Desktop/Data/FishMercury.csv')
hist(fm$Mercury, main = 'Histogram of Mercury Levels', xlab = 'Mercury Levels', breaks = 100)
```




```
boxplot(fm$Mercury, main = 'Boxplot of Mercury Levels')
```



b) Standard Error: $\approx .059$; 95% Confidence Interval: [.112, .307]

```
merc <- fm$Mercury
means <- replicate(10000, mean(sample(merc, length(merc), replace = TRUE)))

sd(means)
```

```
## [1] 0.05695653
```

```r
quantile(means, probs = c(.025, .975))
```

```
##      2.5%     97.5%
## 0.1123333 0.3060000
```

c) Standard Error: $\approx .008$; Confidence Interval: [.108, .139]

```r
merc1 <- merc[merc <= .5]

means1 <- replicate(10000, mean(sample(merc1, length(merc1), replace = TRUE)))

sd(means1)
```

```
## [1] 0.007908355
```

```r
quantile(means1, probs = c(.025, .975))
```

```
##      2.5%     97.5%
## 0.1079302 0.1386552
```

d) Removing the outlier had a noticeable effect on both the standard error and the confidence interval. With respect to the standard error, removing the outlier shrunk it from .059 down to .008; that is, there was far less spread in the distribution of bootstrapped means. Regarding the confidence interval, removing the outlier narrowed it from [.112, .307] to [.108, .139]; that is, without the outlier, our 95% confidence interval was much smaller than with it.

Problem 3

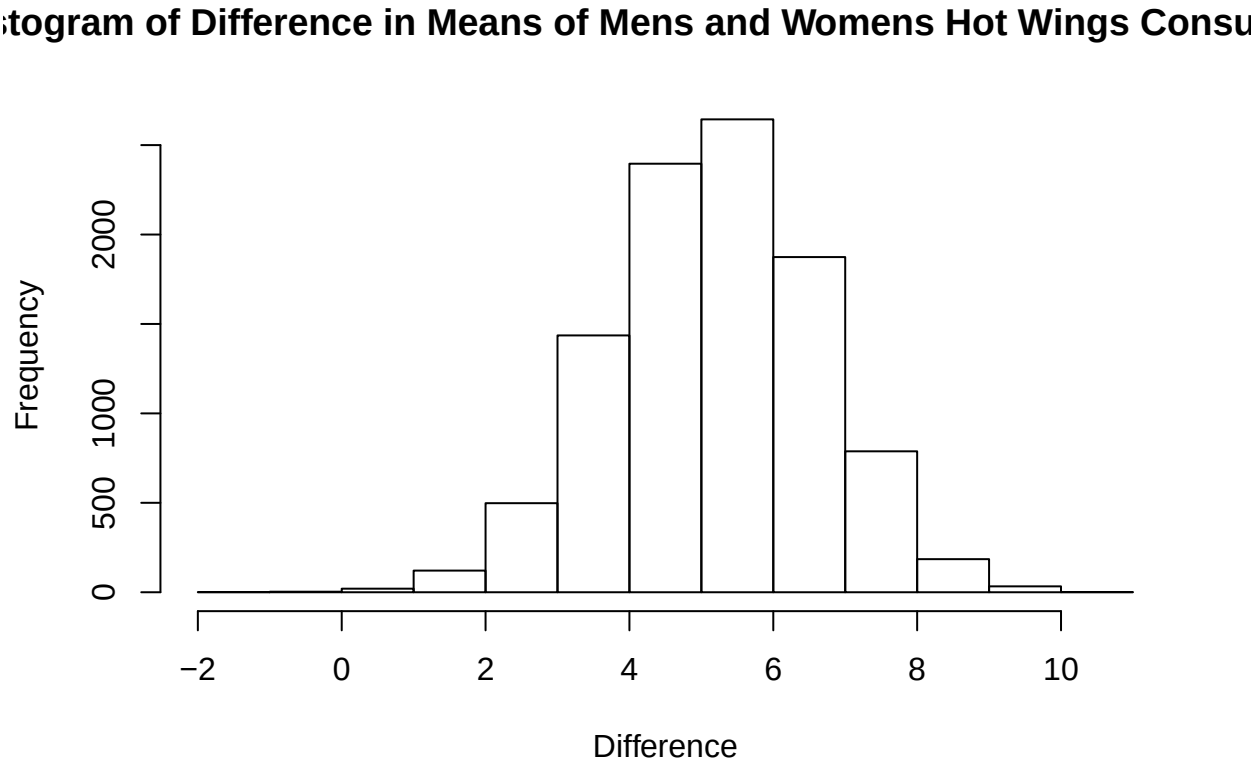
a) The bootstrap distribution of the absolute difference in means of mens' and womens' hot wings consumption has a mean of ≈ 5.19 and $\sigma \approx 1.44$. This means that we can be fairly sure that the difference between the two is not 0, which is reflected in the confidence interval at the .025 level: [1.93, 8.40] - well above 0.

```r
BW <- read.csv('/Users/brodyvogel/Desktop/Data/Beerwings.csv')

ladies <- BW[BW$Gender == 'F', ]
dudes <- BW[BW$Gender == 'M', ]

samp2 <- replicate(10000, mean(sample(dudes$Hotwings, length(dudes$Hotwings), replace = TRUE)) - mean(sa

hist(samp2, main = 'Histogram of Difference in Means of Mens and Womens Hot Wings Consumption', xlab =
```




```
mean(samp2)
```

```
## [1] 5.2108
```

```
sd(samp2)
```

```
## [1] 1.435381
```

```
quantile(samp2, probs = c(.05, .95))
```

```
##       5%      95%
## 2.866667 7.533333
```

```
quantile(samp2, probs = c(.025, .975))
```

```
##     2.5%    97.5%
## 2.400000 7.933333
```

```
quantile(samp2, probs = c(.0125, .9875))
```

```
##    1.25%   98.75%
## 1.933333 8.400000
```

b) 95% Confidence Interval: [2.40, 7.93] ; The probablity that the true difference between the means of mens' and womens' hot wings consumption is between 2.4 and 8 is .95.

```
quantile(samp2, probs = c(.025, .975))
```

```
##     2.5%    97.5%
## 2.400000 7.933333
```

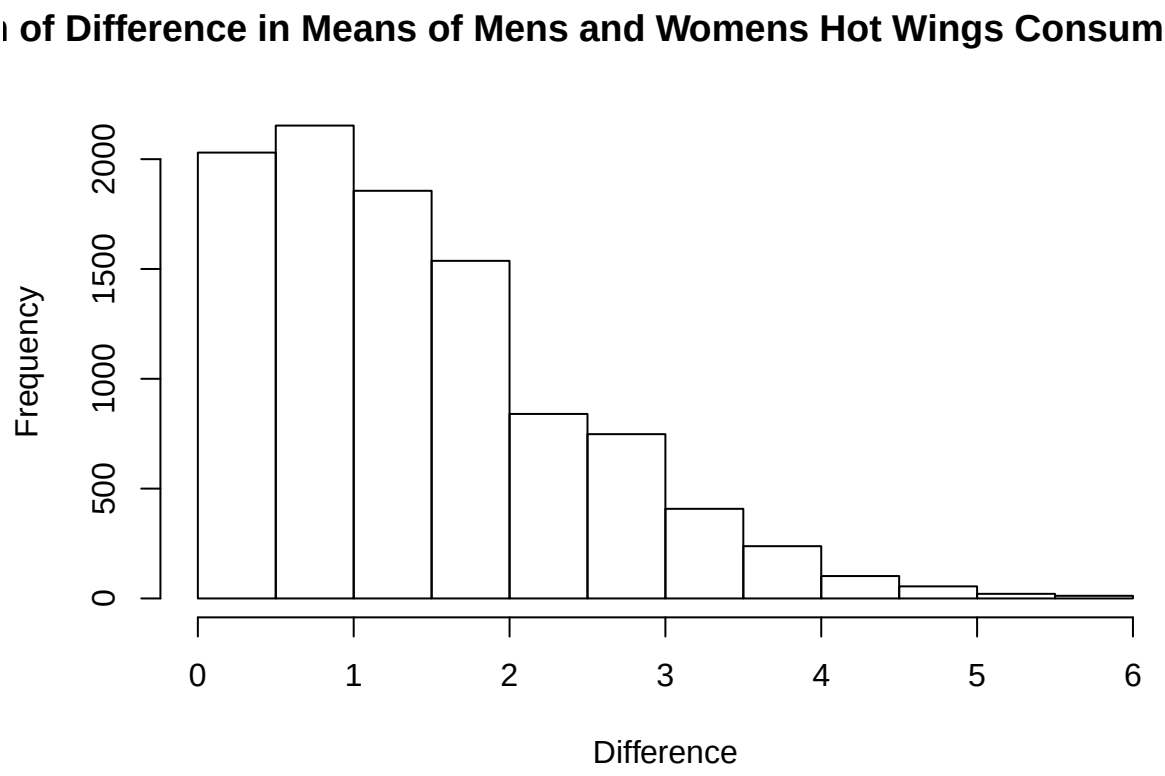
c) They differ quite a lot, because the bootstrap samples from the groups after they've alredy been partioned into men and women, whereas the permutation test samples from the entire collection of hot wings eaters. Particularly:

σs: 1.44 (bootsrap) vs. 1.03 (permutation)

means: 5.19 (bootstrap) vs 1.41 (permutation)

This coincides with theory insofar as we know that bootstrap samples do a better job of estimating the true distribution.

```
samp3 <- c()
for (x in 1:10000) {
  test <- sample(BW$ID, length(dudes$Hotwings), replace = FALSE)
  test1 <- BW$ID[-test]
  samp3 <- c(samp3, abs(mean(BW[test, ]$Hotwings) - mean(BW[test1, ]$Hotwings)))
}
hist(samp3, main = 'Histogram of Difference in Means of Mens and Womens Hot Wings Consumption (Permutati
```




```
sd(samp3)
```

```
## [1] 1.039706
```

```
mean(samp3)
```

```
## [1] 1.413213
```

Problem 4

We can't say that chocolate and vanilla ice cream do not have the same number of calories. At each of the .9, .95, and .975 confidence levels, 0 is in the confidence interval.

```
IC <- read.csv('/Users/brodyvogel/Desktop/Data/IceCream.csv')
Choc <- IC$ChocolateCalories
Van <- IC$VanillaCalories
```

```r
samps <- replicate(10000, mean(sample(Choc, length(Choc), replace = TRUE)) - mean(sample(Van, length(Va

quantile(samps, probs = c(.0125, .9875))
```

```
##     1.25%    98.75%
## -23.82083  38.61538
```

```r
quantile(samps, probs = c(.025, .975))
```

```
##      2.5%     97.5%
## -19.61795  34.46218
```

```r
quantile(samps, probs = c(.05, .95))
```

```
##        5%       95%
## -15.23205  30.07821
```

Problem 5

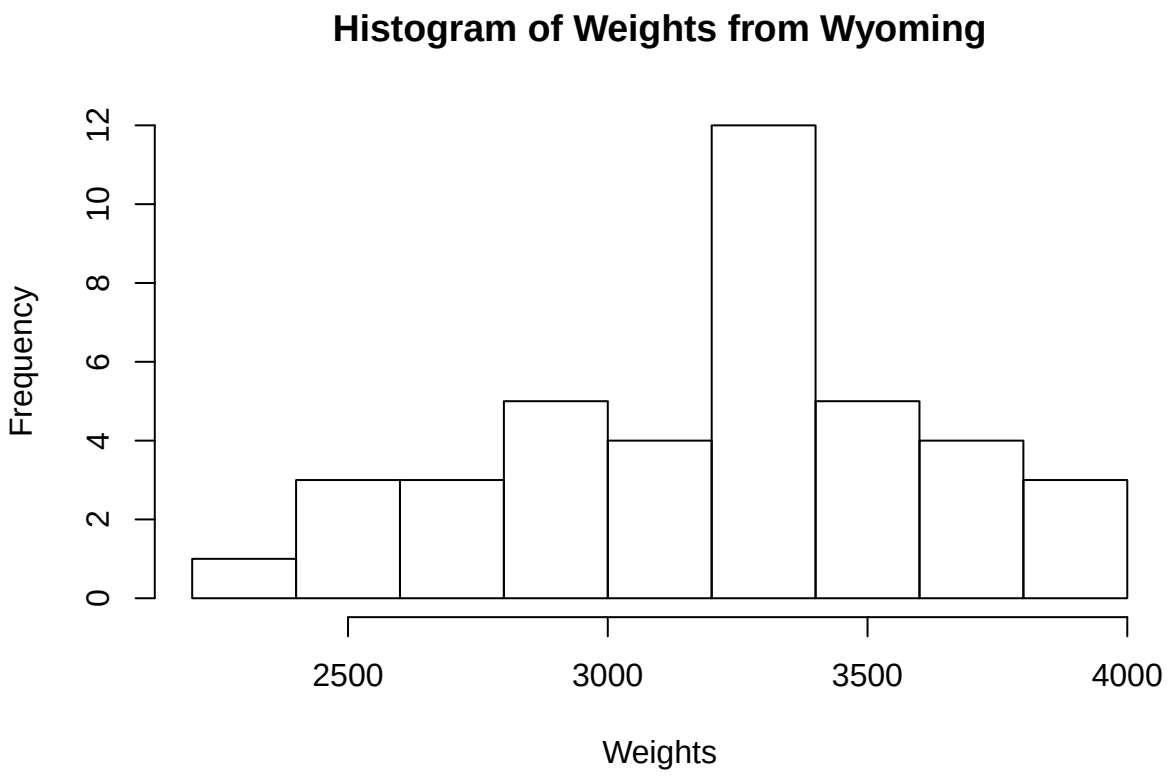
a) Neither the weights of baby girls from Wyoming nor from Arkansas appears to be normally-distributed, although that can't be said with much confidence.

It looks like girls born in Arkansas were heavier than those born in Wyoming, with a mean of 3516 compared to one of 3208, and a median of 3558 compared to one of 3278.

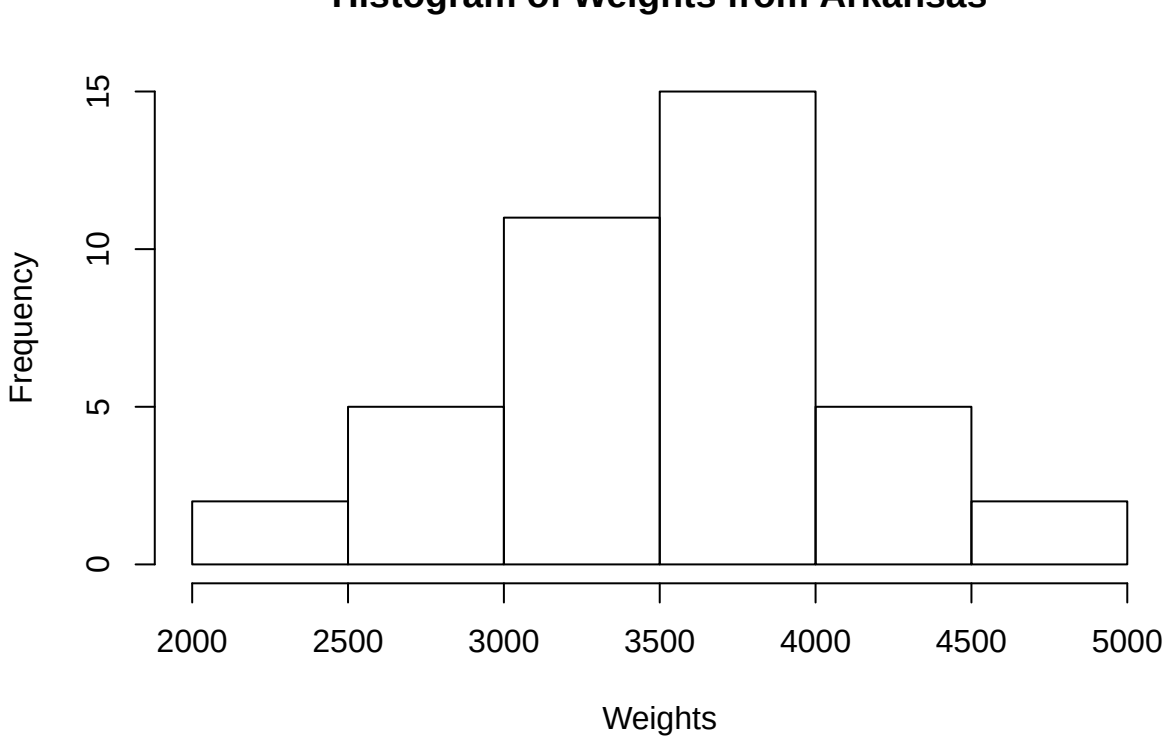
```r
girls <- read.csv('/Users/brodyvogel/Desktop/Data/Girls2004.csv')

yomin <- girls[girls$State == 'WY', ]
ark <- girls[girls$State == 'AK', ]

hist(yomin$Weight, main = 'Histogram of Weights from Wyoming', xlab = 'Weights')
```




```
hist(ark$Weight, main = 'Histogram of Weights from Arkansas', xlab = 'Weights')
```




```
summary(yomin$Weight)
```

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

```
summary(ark$Weight)
```

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

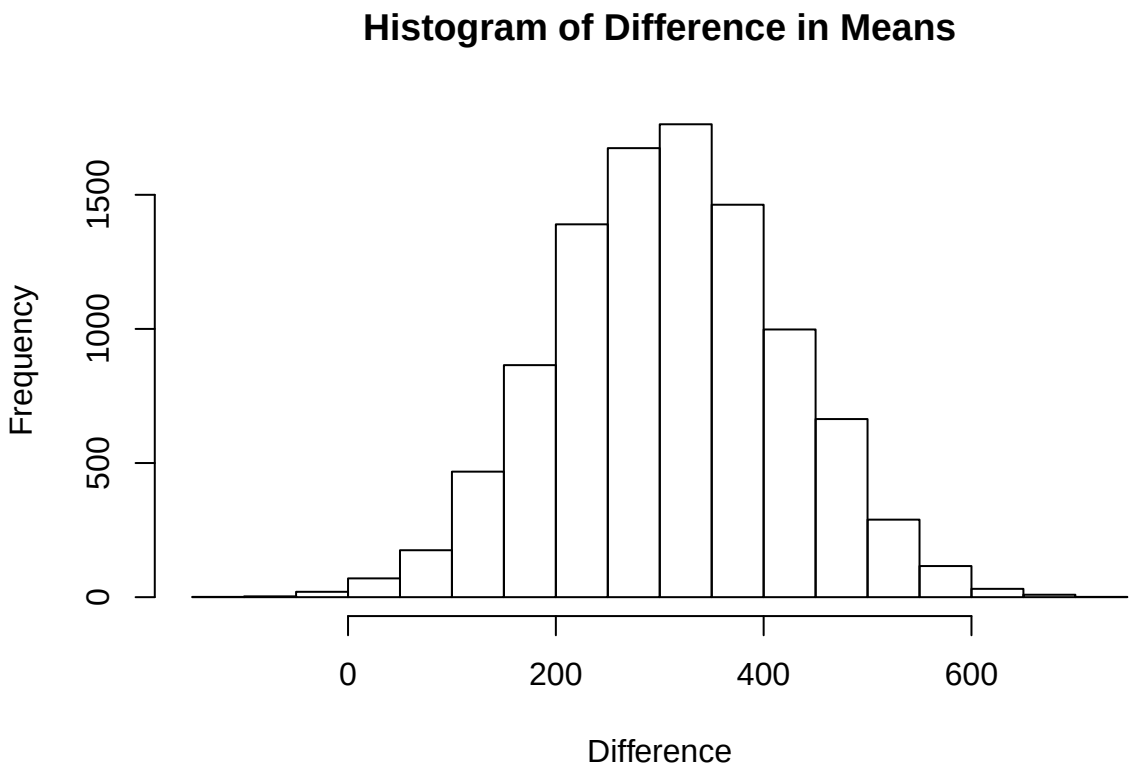
b) The bootstrap sample of differences between means of baby girls born in Arkansas and Wyoming appears to be normally-distributed around ≈ 300, with $\sigma \approx 110$. The mean difference of the bootstrap sample is ≈ 308.5, while the median is ≈ 308.7.

A 95% bootstrapped confidence interval for the difference in means of baby girls born in Arkansas and Wyoming is [87.99, 524.68]. This is a very wide interval, and says that the probability that the true difference in means of baby girls born in Arkansas and Wyoming is between 87.99 abd 524.68 is .95. Also, because the interval is far from containing 0, we can be fairly sure that there is, in fact, a difference in means.

```
yWeights <- yomin$Weight
aWeights <- ark$Weight

samp4 <- replicate(10000, mean(sample(aWeights, length(aWeights), replace = TRUE)) - mean(sample(yWeigh

hist(samp4, main = 'Histogram of Difference in Means', xlab = 'Difference')
```




```
summary(samp4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -101.5   234.4   309.8   310.4   385.6   717.2
```

```
sd(samp4)
```

```
## [1] 111.1681
```

```
quantile(samp4, probs = c(.025, .975))
```

```
##     2.5%    97.5%
```

```
##  95.3975 528.4750
```

c) bias: 308.5 - 308 = .5.

percentage of Standard Error: .5/110 ≈ .0045 %

d) From the permutation test, we can state fairly confidently that there is a difference in means of baby girls born in Arkansas and Wyoming. The permutation test produced a distribution with a mean difference of 93.78 and median of 79.40. Furthermore, 75% of the permutations produced a difference ≥ 37.

```
samp5 <- c()
for (x in 1:10000) {
  test2 <- sample(girls$ID, length(ark$Weight), replace = FALSE)
  test3 <- girls$ID[-test2]
  samp5 <- c(samp5, abs(mean(girls[test2, ]$Weight) - mean(girls[test3, ]$Weight)))
}

summary(samp5)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   38.25   79.20   93.52  134.66  451.35
```

e) This conclusion holds for the sample data only. A permutation test simulates the null distribution of the sample only, and so no information is obtained about the actual difference in means of all baby girls born in Arkansas and Wyoming. A bootstrap approach could produce a credible estimation of said difference.

Problem 6

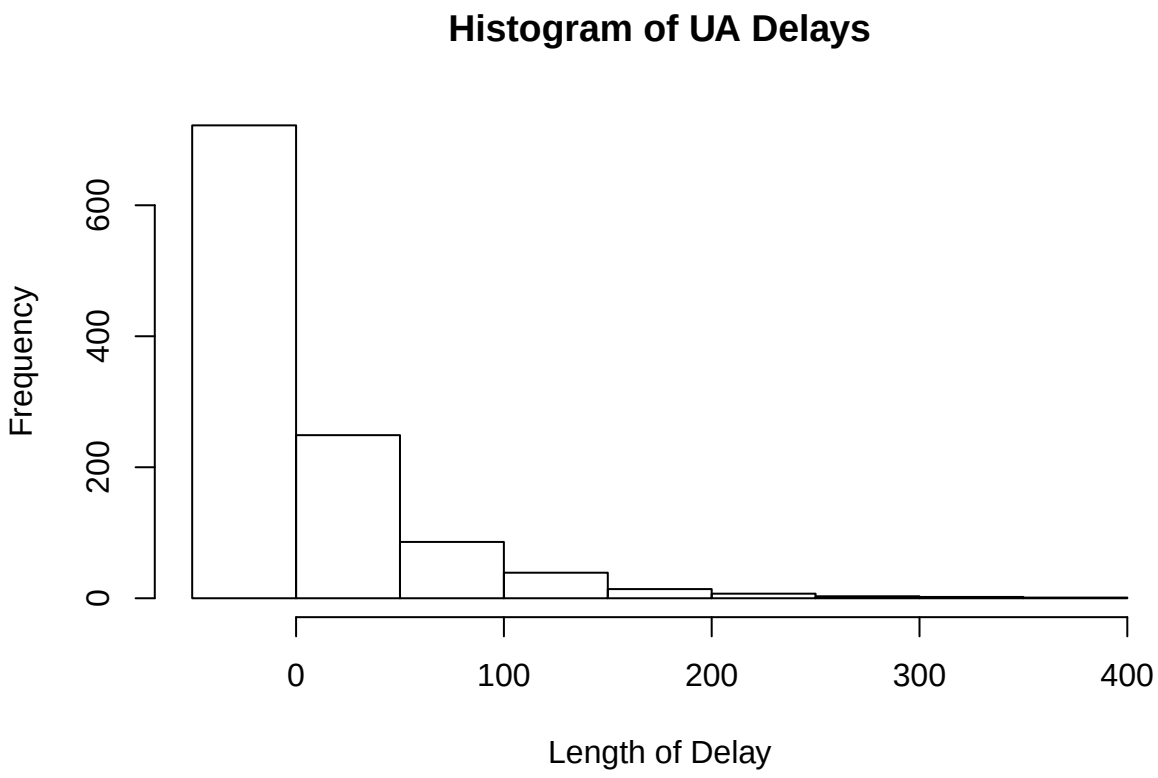
a) Both the distributions of delays on the UA and AA airlines look to be exponentially-distributed, with the majority of values being ≤ 0; that is, flights that weren't delayed.

The mean and median delays on UA were 15.98 and -1, with a standard deviation of 45.14. The mean and median delays on AA were 10.1 and -3, with a standard deviation of 40.08.

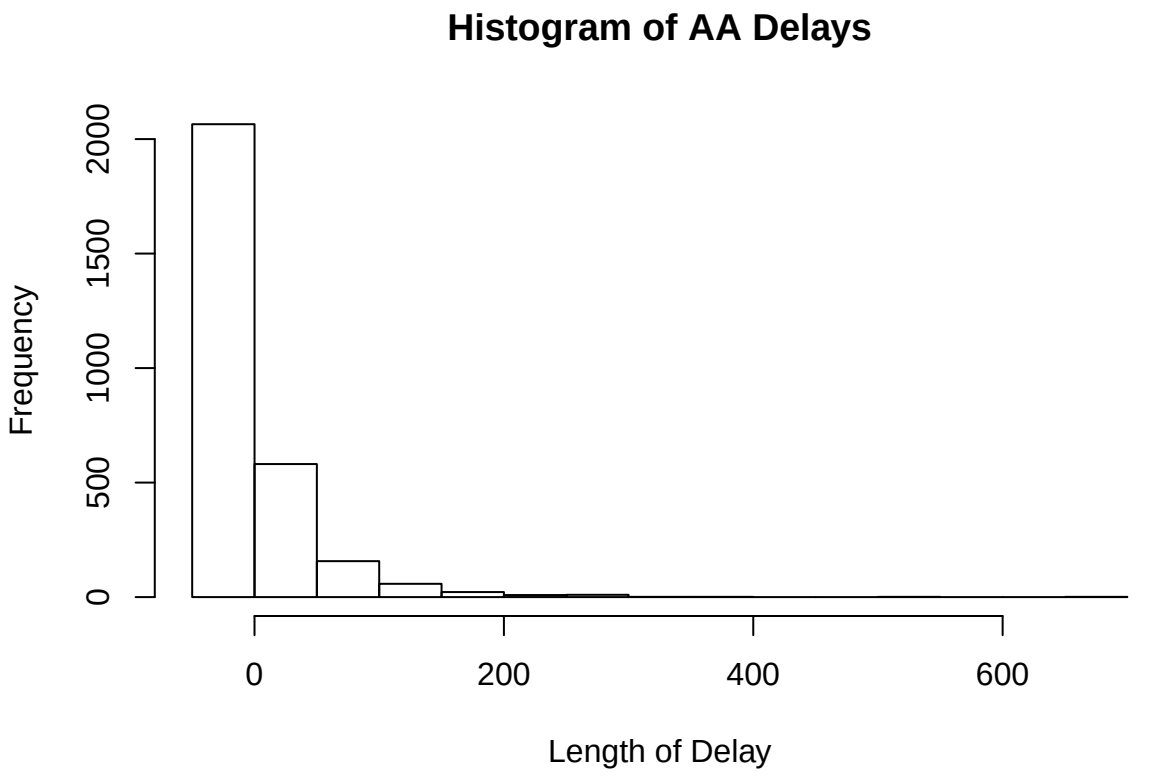
This all tells us that most flights were either on time or early, but those that were delayed tended to be delayed for a while.

```
FD <- read.csv('/Users/brodyvogel/Desktop/Data/FlightDelays.csv')
UA <- FD[FD$Carrier == 'UA', ]
AA <- FD[FD$Carrier == 'AA', ]

hist(UA$Delay, main = 'Histogram of UA Delays', xlab = 'Length of Delay')
```




```
hist(AA$Delay, main = 'Histogram of AA Delays', xlab = 'Length of Delay')
```




```
summary(UA$Delay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##  -17.00   -5.00   -1.00   15.98   12.50  377.00
```

```
summary(AA$Delay)
```

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

```
sd(UA$Delay)
```

```
## [1] 45.13895
```

```
sd(AA$Delay)
```

```
## [1] 40.08063
```

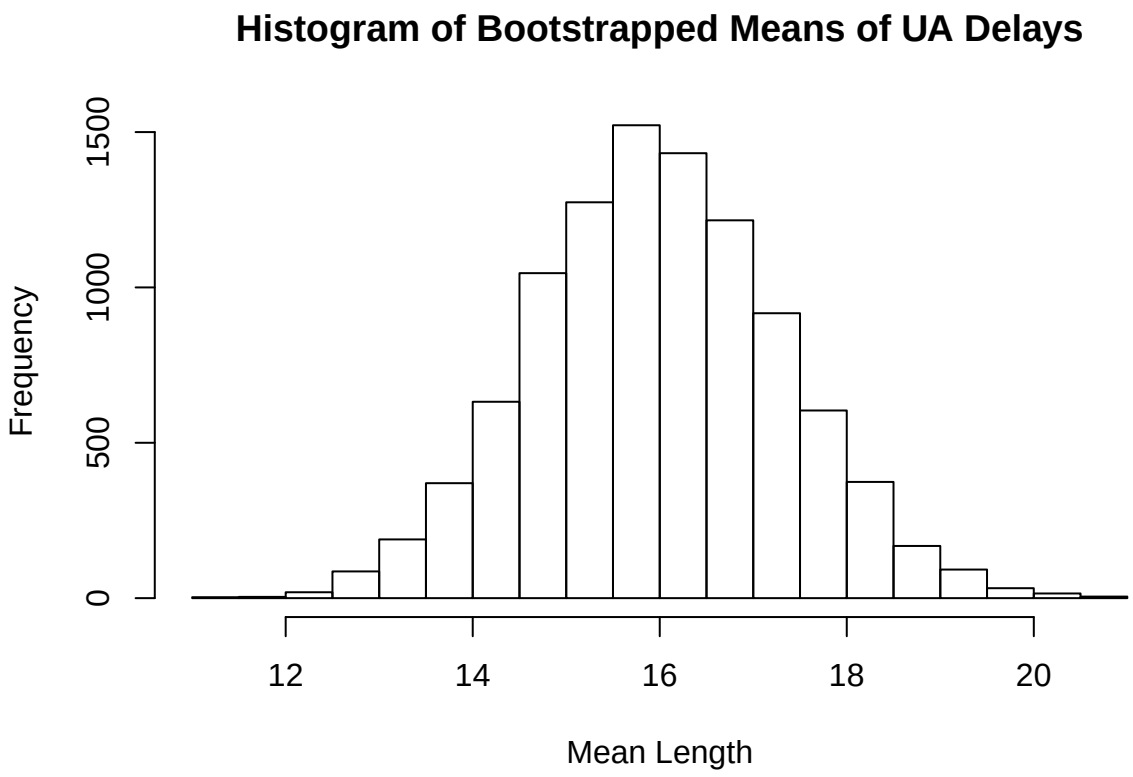
b) The bootstrapped distribution of delays on the UA airline appears to be roughly normally-distributed around ≈ 16. From the summary stats, we see that the mean of the distribution is actually ≈ 15.97, with a median of ≈ 15.94 and $\sigma \approx 1.35$.

The bootstrapped distribution of delays on the AA airline appears to be roughly normally-distributed around ≈ 10. From the summary stats, we see that the mean of the distribution is actually ≈ 10.12, with a median of ≈ 10.11 and $\sigma \approx .75$.

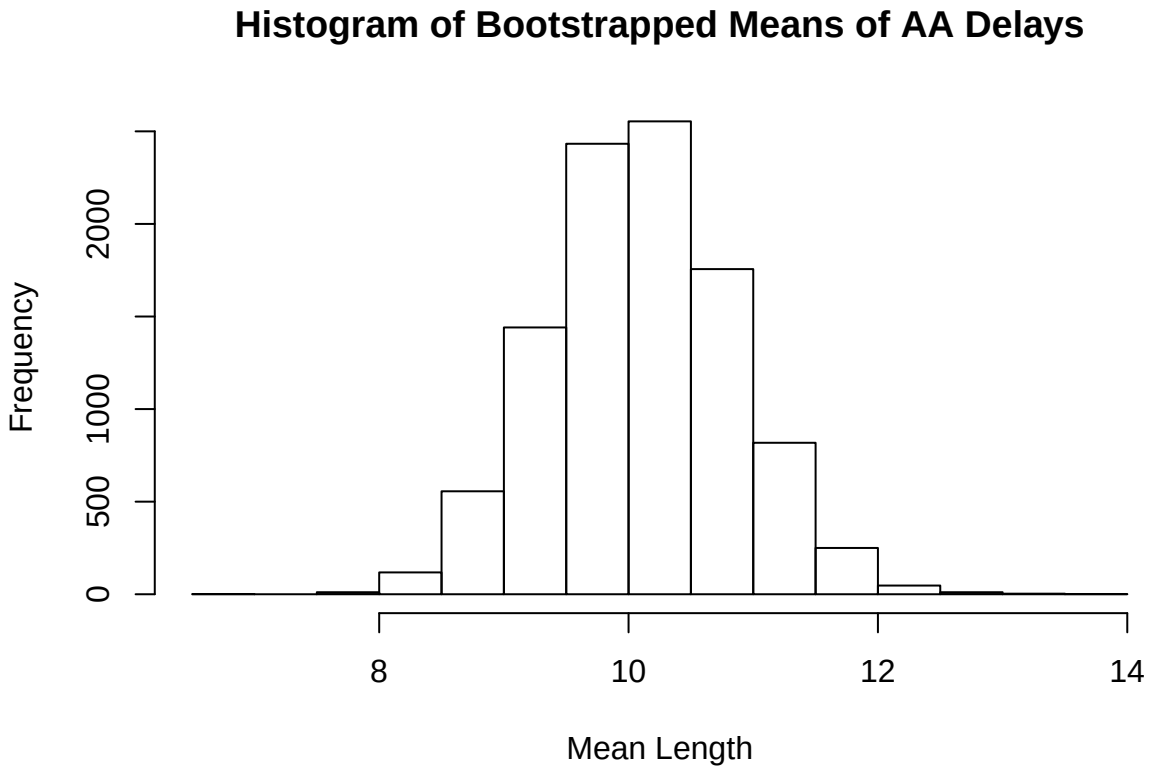
```
UAD <- UA$Delay
AAD <- AA$Delay

UASamp <- replicate(10000, mean(sample(UAD, length(UAD), replace = TRUE)))
AASamp <- replicate(10000, mean(sample(AAD, length(AAD), replace = TRUE)))

hist(UASamp, main = 'Histogram of Bootstrapped Means of UA Delays', xlab = 'Mean Length')
```




```
hist(AASamp, main = 'Histogram of Bootstrapped Means of AA Delays', xlab = 'Mean Length')
```




```
summary(UASamp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.41   15.06   15.95   15.98   16.87   20.99
```

```
summary(AASamp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.917   9.586  10.077  10.094  10.592  13.571
```

```
sd(UASamp)
```

```
## [1] 1.336137
```

```
sd(AASamp)
```

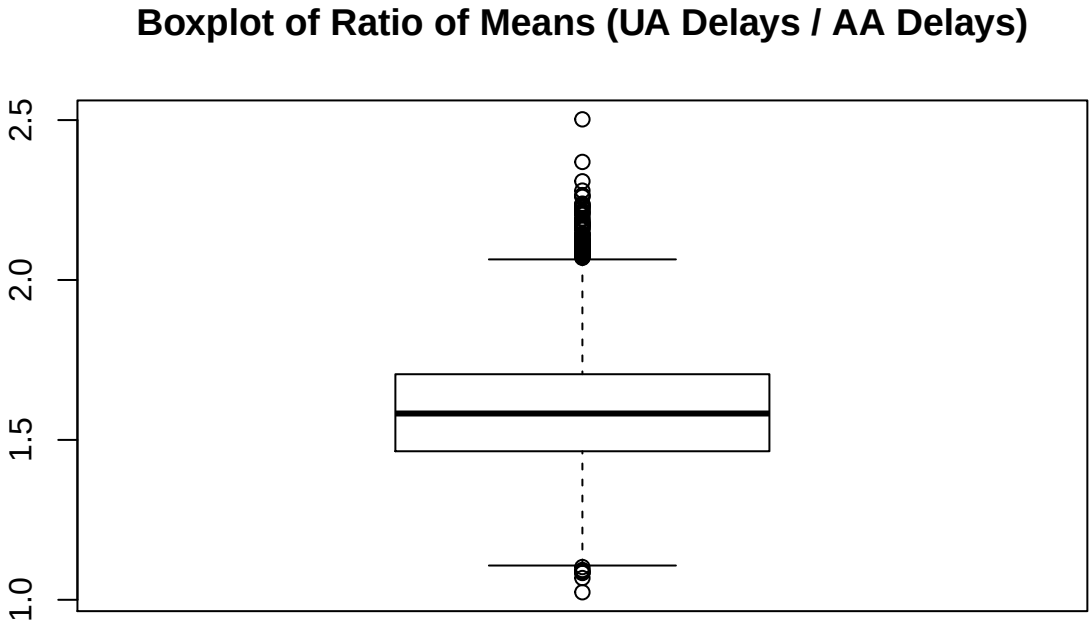
```
## [1] 0.7444942
```

c) The bootstrapped distribution of the ratio of the mean of delays on the UA airline to the mean of those on the AA airline appears to be roughly normally-distributed around ≈ 1.6. From the summary stats, we see that the mean of the distribution is actually ≈ 1.59, with a median of ≈ 1.58 and $\sigma \approx .18$.

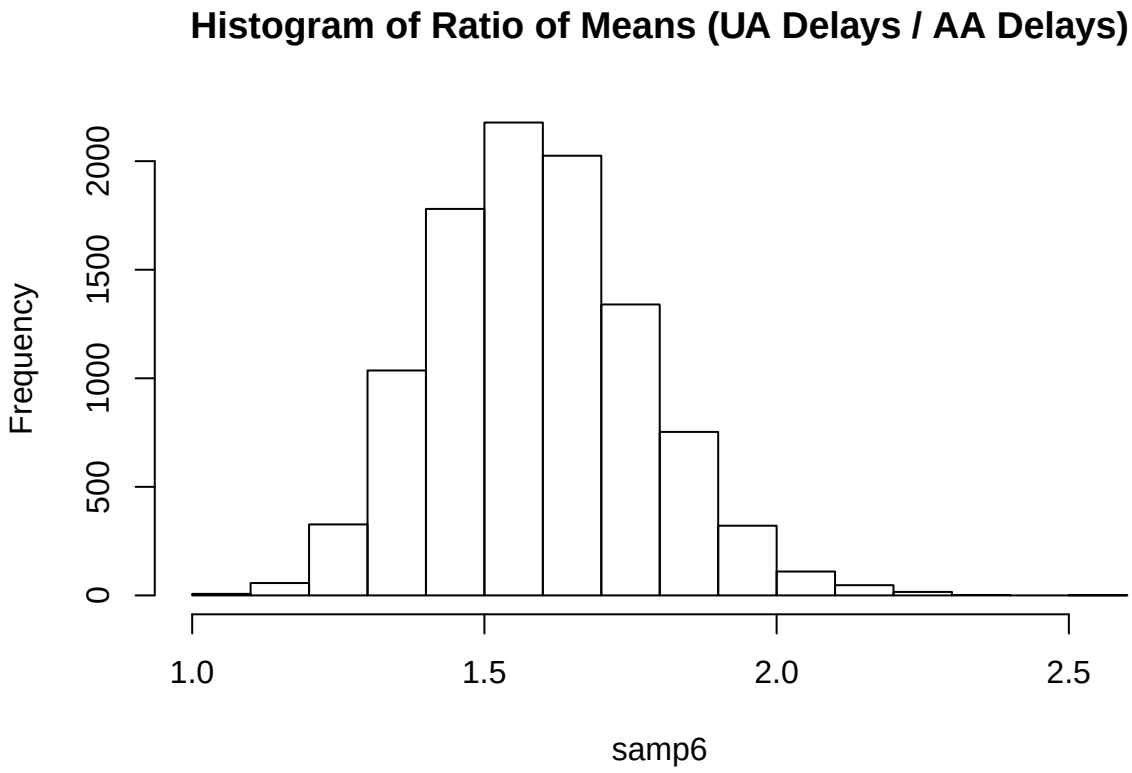
The boxplot shows that there is a significant number of outliers in the distribution of this statistic, though.

```
samp6 <- replicate(10000, mean(sample(UAD, length(UAD), replace = TRUE))/mean(sample(AAD, length(AAD), r

boxplot(samp6, main = 'Boxplot of Ratio of Means (UA Delays / AA Delays)')
```




```
hist(samp6, main = 'Histogram of Ratio of Means (UA Delays / AA Delays)')
```




```
summary(samp6)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.024   1.464   1.583   1.591   1.705   2.502
```

```
sd(samp6)
```

```
## [1] 0.179292
```

d) 95% Confidence Interval: [1.27, 1.96].

The probability that the true ratio of the mean of flight delays on the UA airline to the mean of those on the AA airline is betwen 1.27 and 1.96 is .95. Also, because 1 is not in the interval, it is very unlikely that there is, in fact, no difference.

```
quantile(samp6, probs = c(.025, .975))
```

```
##     2.5%    97.5%
## 1.270412 1.964431
```

e) bias: 1.591 - 1.582 $\approx$ .009 percentage of Standard Error: .009/.18 = .05

```
bias <- 1.591 - (mean(FD[FD$Carrier == 'UA', ]$Delay)/mean(FD[FD$Carrier == 'AA', ]$Delay))

bias
```

```
## [1] 0.008106857
```

f) I would say no. Flight delays are often caused by inclement weather, which may have corresponded to certain destinations more often than others in the test data. So, if one airline more often carried flights to these destinations with adverse weather, it'd be expected that that airline would have more, and longer, delays. So no, I don't think we could actually assume that the observations are independent.

Problem 7

a) Median = 22 ; 90th Percentile = 270

```
median(bdesh$Arsenic)
```

```
## [1] 22
```

```
quantile(bdesh$Arsenic, probs = .90)
```

```
## 90%
## 270
```

b) bias(median) = $\overline{\sigma}_{boot} - \hat{\sigma} \approx$ 23.51 - 22 = 1.51

```
ars <- bdesh$Arsenic
samp7 <- replicate(10000, median(sample(ars, length(ars), replace = TRUE)))
mean(samp7)
```

```
## [1] 23.48084
```

```
bias <- mean(samp7) - median(ars)
bias
```

```
## [1] 1.48084
```

c) bias(90th percentile) = $\overline{\sigma}_{boot} - \hat{\sigma} \approx$ 272.93 - 270 = 2.93

```
samp8 <- replicate(10000, quantile(sample(ars, length(ars), replace = TRUE), probs = .9))
mean(samp8)
```

```
## [1] 273.9174
```

```
bias = mean(samp8) - quantile(ars, probs = .9)
bias
```

```
##    90%
```

```
## 3.9174
```