# ASSIGNMENT 2

Brandon Lampe
STAT 527
Advanced Data Analysis I

September 25, 2014

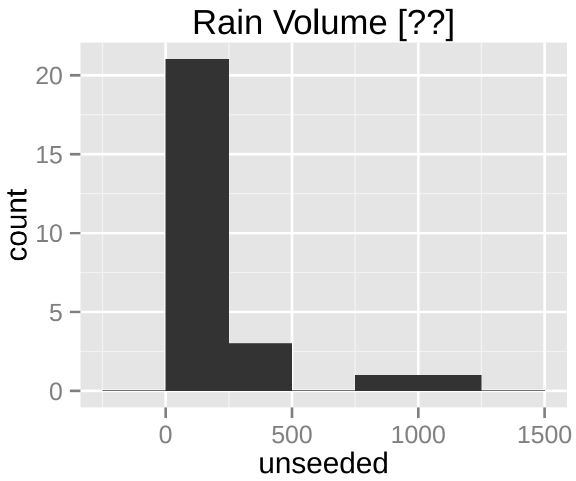
## 1 Unseeded vs seeded precipitation:

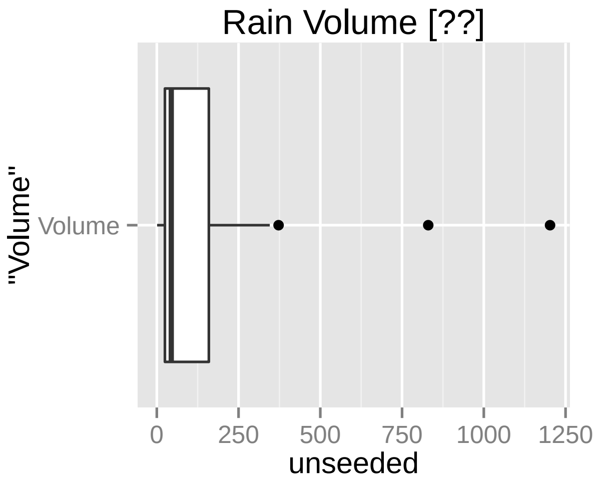
The population mean in this problem is of the total rain volume falling from the cloud base following the airplane seeding run, as measured by radar. The hypothesis is that seeding clouds can lead to increased rainfall.

```
d1 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-1.csv")
library(ggplot2)
library(grid)
library(gridExtra)
source("ADA1_FUNC.R")
```

### 1(a) Create histogram and boxplot of the unseeded days; then summarize

```
# histogram of Unseeded Precip
Precip.hist <- ggplot(d1, aes(x = unseeded))
Precip.hist <- Precip.hist + geom_histogram(binwidth = 250)
Precip.hist <- Precip.hist + labs(title = "Rain Volume [??]")

# boxplot of Unseeded Precip
Precip.box <- ggplot(d1, aes(x = "Volume", y = unseeded)) # boxplot of Precip
Precip.box <- Precip.box + geom_boxplot()
Precip.box <- Precip.box + coord_flip()
Precip.box <- Precip.box + labs(title = "Rain Volume [??]")

# plot histogram and boxplot
grid.arrange(Precip.hist, Precip.box, ncol = 2)
```





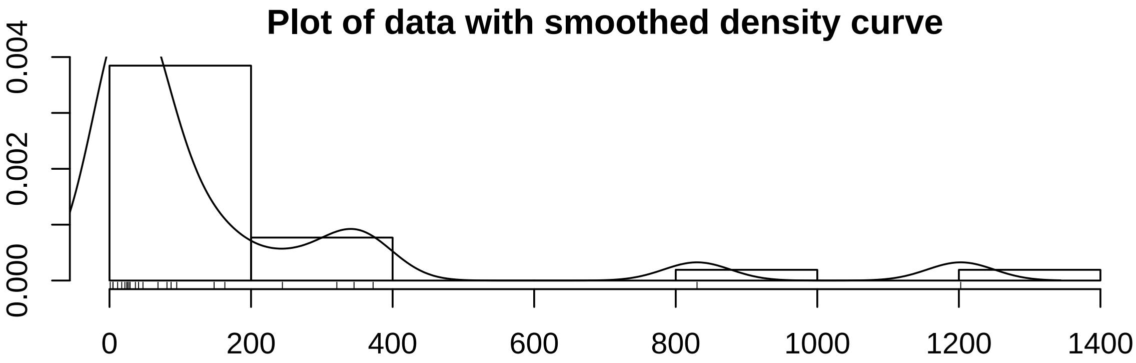
```
mean(d1$unseeded)

## [1] 164.6

median(d1$unseeded)

## [1] 44.2

sd(d1$unseeded)

## [1] 278.4

diff(fivenum(d1$unseeded)[c(2,4)]) #IQR

## [1] 138.6

fivenum(d1$unseeded)

## [1]    1.0   24.4   44.2  163.0 1202.6
```

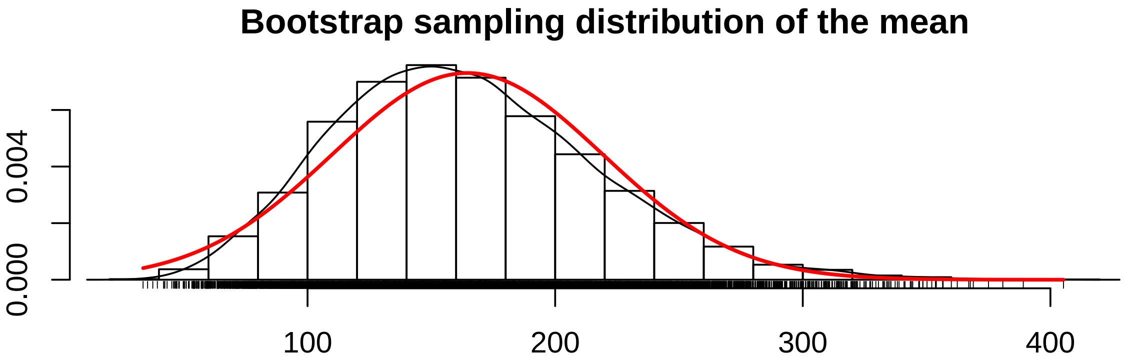
The distribution is unimodal, unsymmetric, skewed right, not normal, and has three outliers. $\bar{Y} - M = 164.6 - 44.2 = 120.4$, which is very large and consistent with observed skewness. No units were provided with the data; therefore, a quantitative analysis is difficult. The data represents volume of rain from clouds that were not seeded; therefore, this data essentially provides a measure of the volume of rain from a cumulus cloud, which often produce little to no precipitation.

## 1(b) Obtain a 95% confidence interval

```
# Manually calculate confidence interval
us.sd     <- sd(d1$unseeded)        # standard deviation unseeded
us.ct     <- length(d1$unseeded)    # count of unseeded
us.SEM    <- us.sd/sqrt(us.ct - 1)  # standard error of the mean
us.M      <- mean(d1$unseeded)      # mean
us.alpha  <- 0.05                   # alpha value, for CI = 95%

abs(qt(us.alpha/2, df = us.ct - 1)) # two sided t_crit
```

```
## [1] 2.06

us.UL     <- us.M + us.tcrit*us.SEM # upper limit of 95% CI

## Error:  object ’us.tcrit’ not found

us.LL     <- us.M - us.tcrit*us.SEM # lower limit of 95% CI

## Error:  object ’us.tcrit’ not found

# Use function to calculate confidence interval
t.test(d1$unseeded, mu = us.M) # Student's t-test function

##
##  One Sample t-test
##
## data:  d1$unseeded
## t = 0, df = 25, p-value = 1
## alternative hypothesis: true mean is not equal to 164.6
## 95 percent confidence interval:
##   52.13 277.05
## sample estimates:
## mean of x
##     164.6

#peform bootstrap sampling to determine if the distribution is normal
bs.one.samp.dist(d1$unseeded)
```





The 95% Confidence Interval = 164.6 ± 112.5. Therefore, with 95% confidence the mean of unseeded rainfall volume will be betwee 52 and 278.

The assumptions for the t-test and the corresponding confidence interval are:

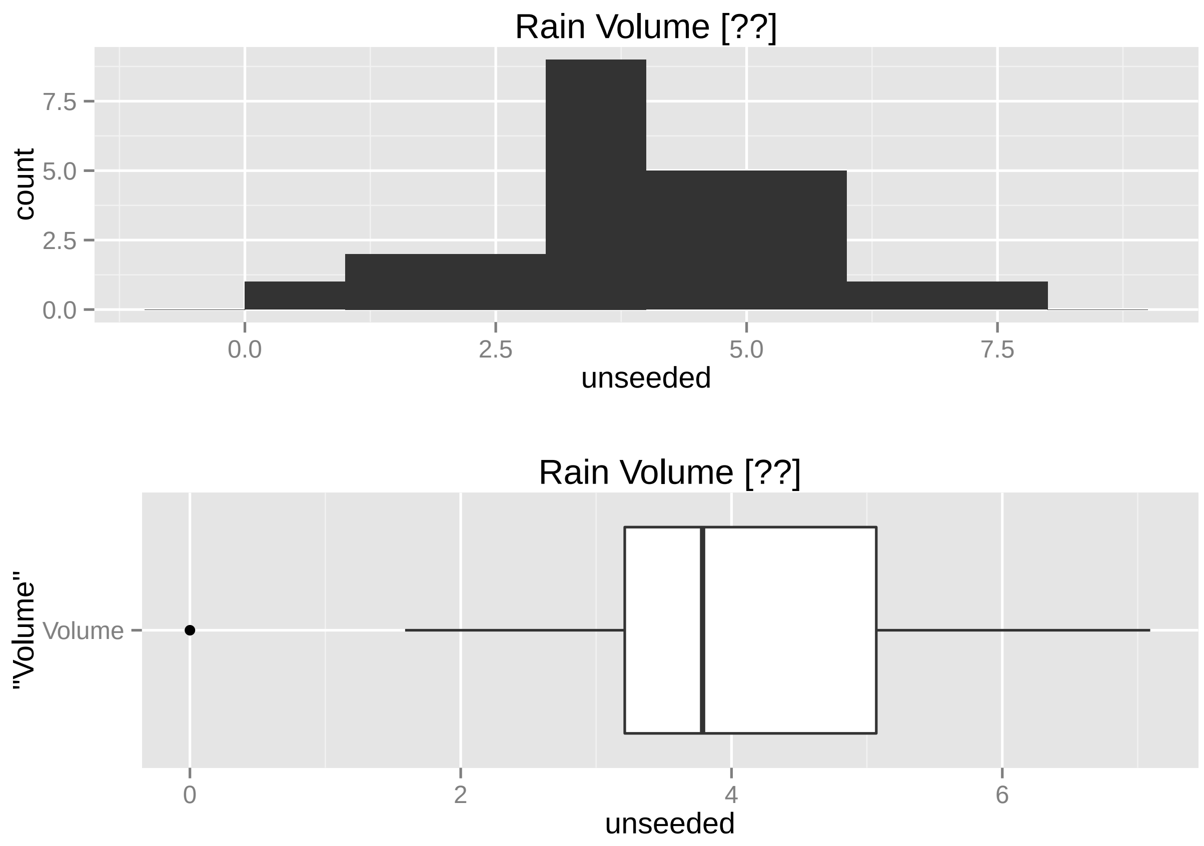
- Data are a random sample from the population
- Population frequency curve is normal

The population of cumulus clouds were randomly seeded by the meachanism. However, using the bootstrop method, the data are skewed right and not normally distributed, which is contrary to the underlying assumpionts used to calculated the confidence interval. The normality assumption was not met.

## 1(c) Transform the data by taking the log() of each value

```
logd1 <- log(d1) # natural log of data

logd1.hist <- ggplot(logd1, aes(x = unseeded))
logd1.hist <- logd1.hist + geom_histogram(binwidth = 1)
logd1.hist <- logd1.hist + labs(title = "Rain Volume [??]")

# boxplot of Unseeded logd1
logd1.box <- ggplot(logd1, aes(x = "Volume", y = unseeded)) # boxplot of logd1
logd1.box <- logd1.box + geom_boxplot()
logd1.box <- logd1.box + coord_flip()
logd1.box <- logd1.box + labs(title = "Rain Volume [??]")
```

```
# plot histogram and boxplot
grid.arrange(logd1.hist, logd1.box, ncol = 1)
```

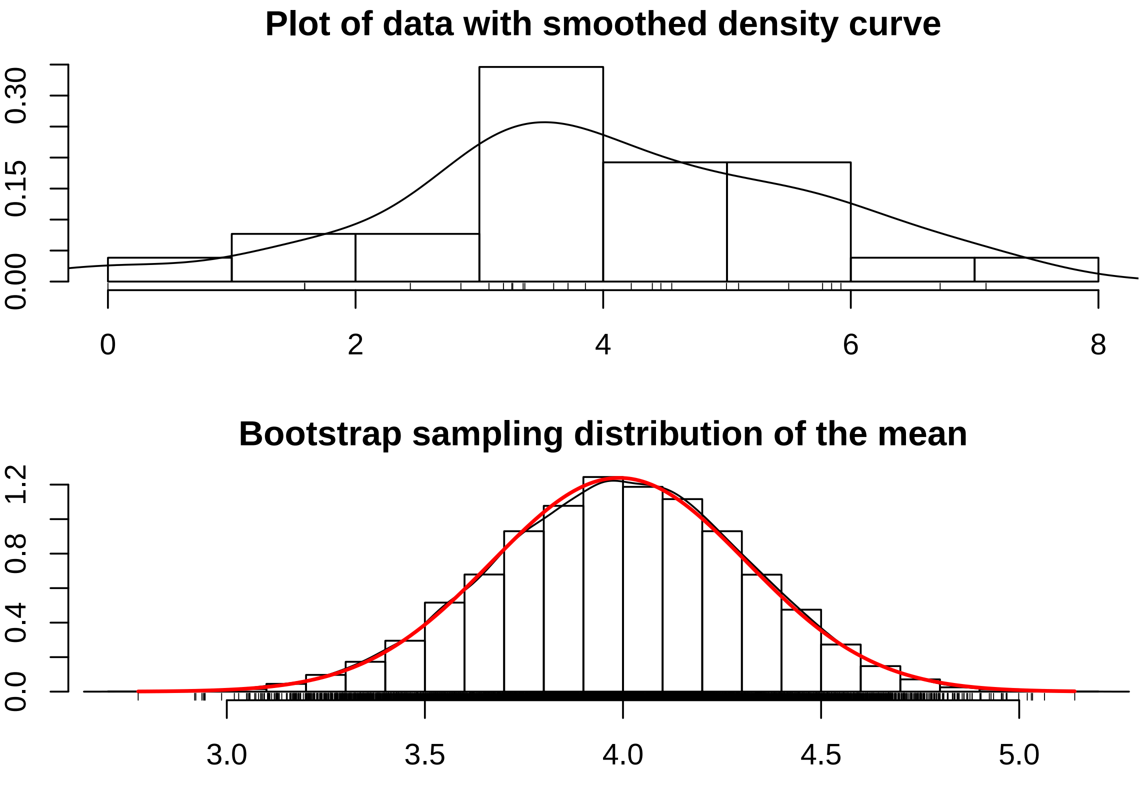


```
mean(logd1$unseeded)

## [1] 3.99

median(logd1$unseeded)

## [1] 3.786

sd(logd1$unseeded)

## [1] 1.642

diff(fivenum(logd1$unseeded)[c(2,4)]) #IQR

## [1] 1.899

fivenum(logd1$unseeded)

## [1] 0.000 3.195 3.786 5.094 7.092
```

The transformed distribution is unimodal, symmetric, close to normal, with one outlier. $\bar{Y} - M = 3.99 - 3.79 = 0.2$ is fairly small relative to the variability, which is consistent with the observed symmetry.

### 1(d) Obtain a 95% confidence interval for the mean log()

```
# Use function to calculate confidence interval
abs(qt(.025, df = us.ct - 1)) # two sided t_crit

## [1] 2.06

t.test(logd1$unseeded) # Student's t-test function

##
##  One Sample t-test
##
## data:  logd1$unseeded
## t = 12.39, df = 25, p-value = 3.59e-12
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3.327 4.654
## sample estimates:
## mean of x
##      3.99

#peform bootstrap sampling to determine if the distribution is normal
bs.one.samp.dist(logd1$unseeded)
```



The 95% confidence interval of the log of the mean unseeded precipitation data is from 3.4 to 4.7. Yes, the assumptions for this method appear to be appropriate because the data were randomly sampled and the Bootstrap sampling of the data are normally distributed.

## 2 TCL

```
d2 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-2.csv")
```

### 2(a)

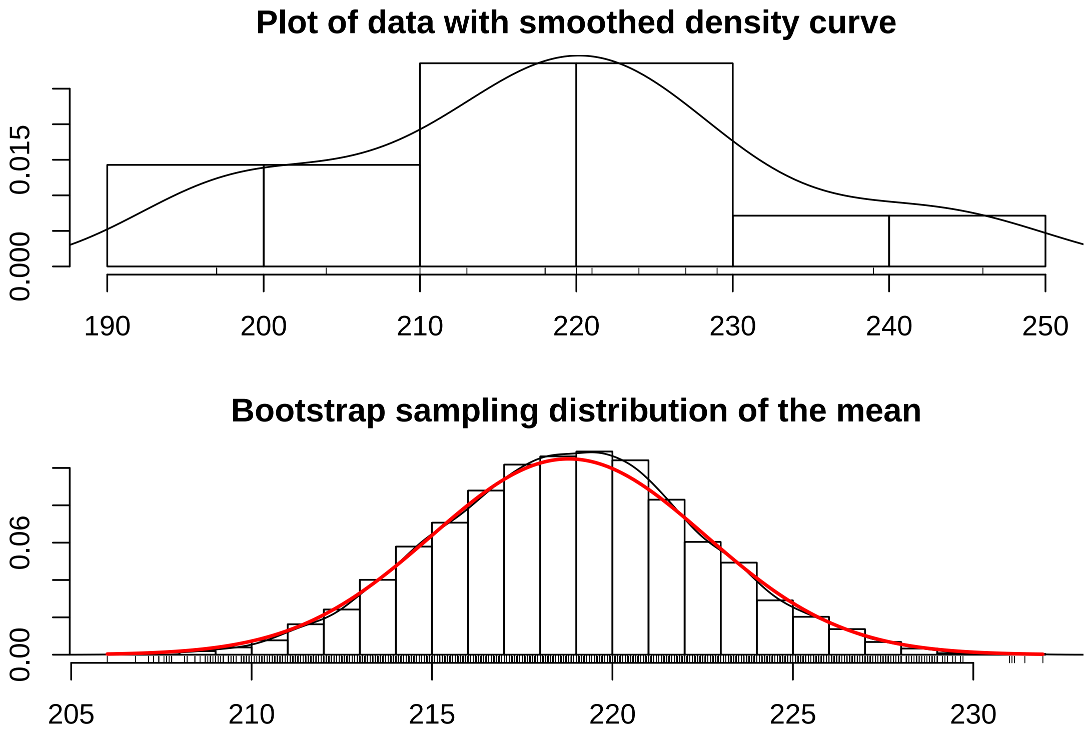
(A) **define the population parameter:** The population parameter is the mean TCL of young adult males on the Kaiser plan. $\mu$ = mean TCL of young adult males on the Kaiser plan.

(B) **state the hypothesis:** The hypothesis is that the TCL of young adult males on the Kaiser plan is equal to the mean TCL of all adult males in the United States, which is 210. $H_0 : \mu = 210$ against $H_A : \mu \neq 210$.

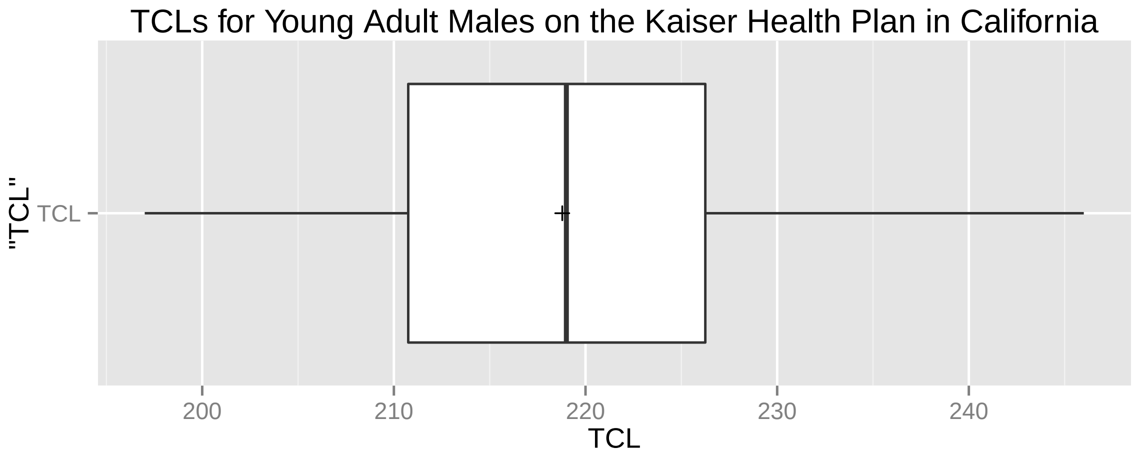
(C) **state assumptions and how they will be assessed:** The assumptions for this analysis are that the data are normally distributed and they data were randomly sampled. Normality will be assessed using the bootstrap method where 10,000 proxy samples are randomly drawn from the full sample. The mean of each proxy sample is then calculated and plotted against a normal distribution curve. If the distribution of the bootstrap (proxy) samples appears normal, then the normality assumption is valid. No information is provided regarding the randomness of the sample population; therefore, I will assume it was randomly obtained.

(D) **evaluate assumptions based on graphical summaries:**

```
bs.one.samp.dist(d2$TCL)       # plot histogram and frequency density curve
```



```
# plot box plot
d2.box <- ggplot(d2, aes(x = "TCL", y = TCL)) # boxplot of d2
d2.box <- d2.box + geom_boxplot()
d2.box <- d2.box + coord_flip()
d2.box <- d2.box + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
d2.box <- d2.box + labs(title = "TCLs for Young Adult Males on the Kaiser Health Plan in Californi
d2.box
```



Based on the histogram of the sample data, the data are unimodal, symmetric, and nearly normal. Based on the bootstrap distribution, the sample data are normally distributed. The boxplot of sample data shows that no outliers exist, the tails are approximately equal, and the mean is very near the meadian.

(E) **discuss the test and the decision:**

```
ybar <- mean(d2$TCL)# mean of sample
s <- sd(d2$TCL)      # standard deviation
n <- length(d2$TCL) # observations in sampe
SEM <- s/sqrt(n)     # standard error of the mean
df <- n - 1          # degrees of freedom
tcrit <-qt(.025, df)# t_crit
M <- median(d2$TCL)# median

210 - tcrit*SEM

## [1] 218.2

210 + tcrit * SEM

## [1] 201.8

d2.t <- t.test(d2, mu = 210)  # t-test summary
d2.t

##
##  One Sample t-test
##
```

```
## data:  d2
## t = 2.309, df = 13, p-value = 0.038
## alternative hypothesis: true mean is not equal to 210
## 95 percent confidence interval:
##  210.6 227.0
## sample estimates:
## mean of x
##     218.8
```

The sample summaries are $n = 14, \bar{Y} = 218.8, s = 14.2$ and $SE_{\bar{Y}} = 3.80$. $\bar{Y} - M = 218.8 - 219 = 0.02$ is very small relative to the variability, which is consistent with the observed symmetry.

I reject $H_0$ in favor of $H_A$. This test was performed at a 5% level (i.e. $\alpha = 0.05$), and the p-value (0.038) is $\leq \alpha$ and accordingly $t_s(2.309) \geq t_{crit}(2.16)$. The data suggest that $\mu \neq 210$.

### 2(b)

```
t.test(d2, mu = 210)

##
##  One Sample t-test
##
## data:  d2
## t = 2.309, df = 13, p-value = 0.038
## alternative hypothesis: true mean is not equal to 210
## 95 percent confidence interval:
##  210.6 227.0
## sample estimates:
## mean of x
##     218.8

IQR(d2$TCL)

## [1] 15.5
```

With 95% confidence, the mean TCLs of young males on the Kaiser Health plan are between 211 and 227, which does not include $H_0$. That is $\bar{Y} \pm t_{0.05} SE_{\bar{Y}} = (211, 227)$. Therefore, I reject $H_0$ in favor of $H_A$ because $\mu \neq 210$.

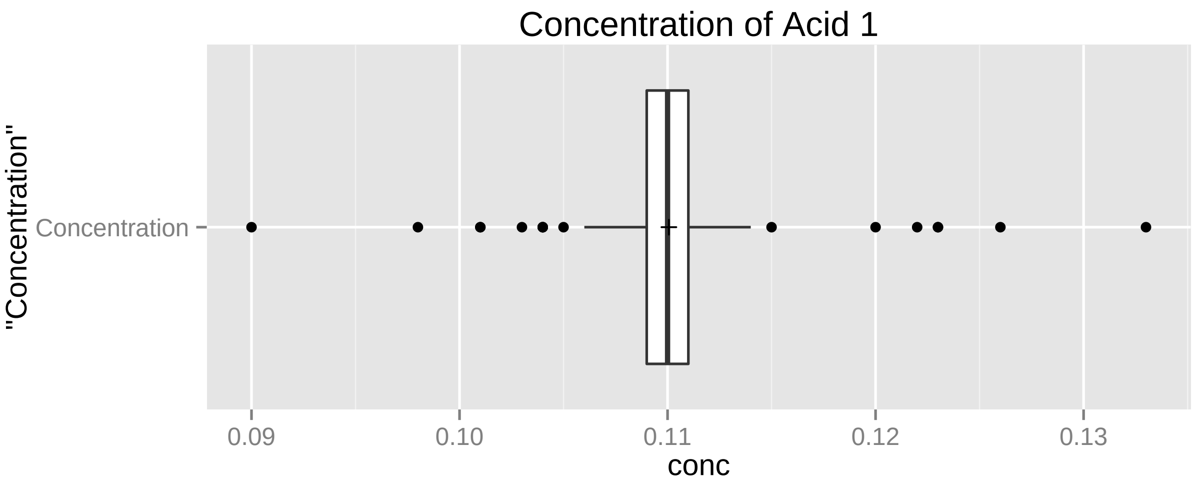
## 3 Acid

```
d3 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-3.csv")
acid1 <- subset(d3,exper =="Acid1", select = c(conc,exper))
acid2 <- subset(d3,exper =="Acid2", select = c(conc,exper))
```

The population parameter is the acidity of the solution. The hypothesis is that the class was ”biased” and thought the acidity was either less or greater than it actually was.

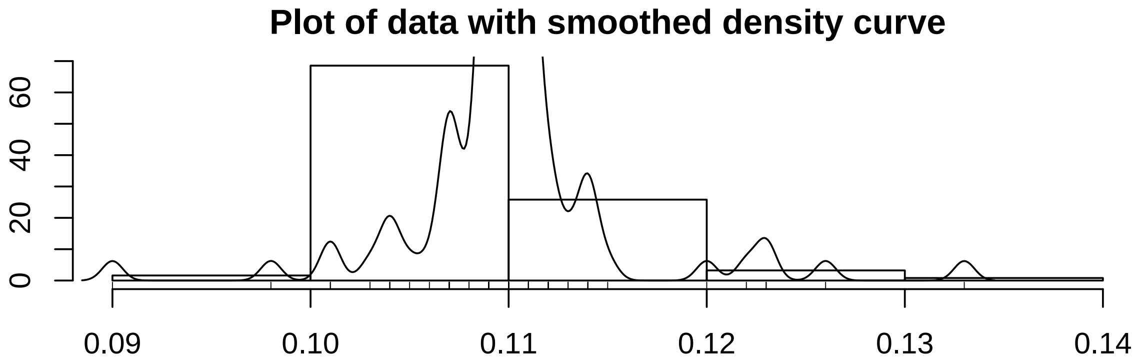
### 3(a) Acid 1

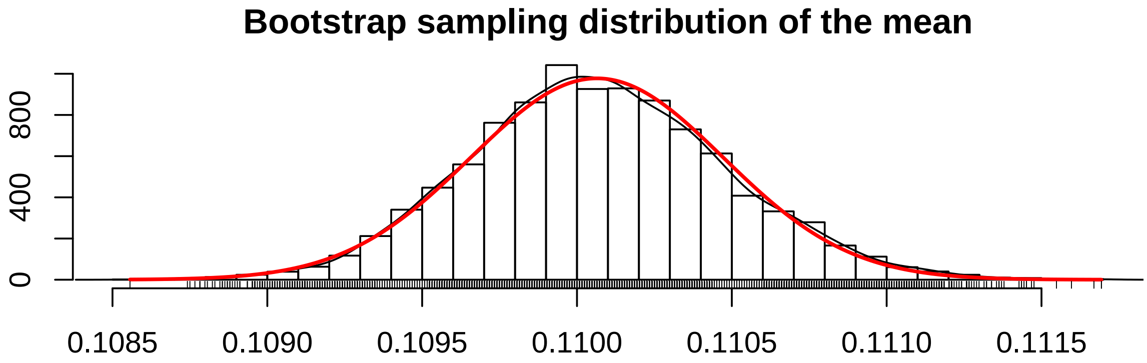
```
# boxplot of Unseeded acid1
acid1.box <- ggplot(acid1, aes(x = "Concentration", y = conc)) # boxplot of acid1
acid1.box <- acid1.box + geom_boxplot()
acid1.box <- acid1.box + coord_flip()
acid1.box <- acid1.box + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
acid1.box <- acid1.box + labs(title = "Concentration of Acid 1")
acid1.box
```



```
acid1.t <- t.test(acid1$conc, mu = 0.110)  # t-test summary
acid1.t

##
##  One Sample t-test
##
## data:  acid1$conc
## t = 0.1581, df = 123, p-value = 0.8746
## alternative hypothesis: true mean is not equal to 0.11
## 95 percent confidence interval:
##  0.1093 0.1109
## sample estimates:
## mean of x
##    0.1101

acid1.ybar <- mean(acid1$conc)  # mean of sample
acid1.s <- sd(acid1$conc)       # standard deviation
acid1.n <- length(acid1$conc)   # observations in sampe
acid1.SEM <- s/sqrt(acid1.n)    # standard error of the mean
acid1.df <- acid1.n - 1         # degrees of freedom
acid1.tcrit <-qt(.025, acid1.df)# t_crit
acid1.M <- median(acid1$conc)   # median
acid1.IQR <- IQR(acid1$conc)
```

```
bs.one.samp.dist(acid1$conc)      # plot histogram and frequency density curve
```





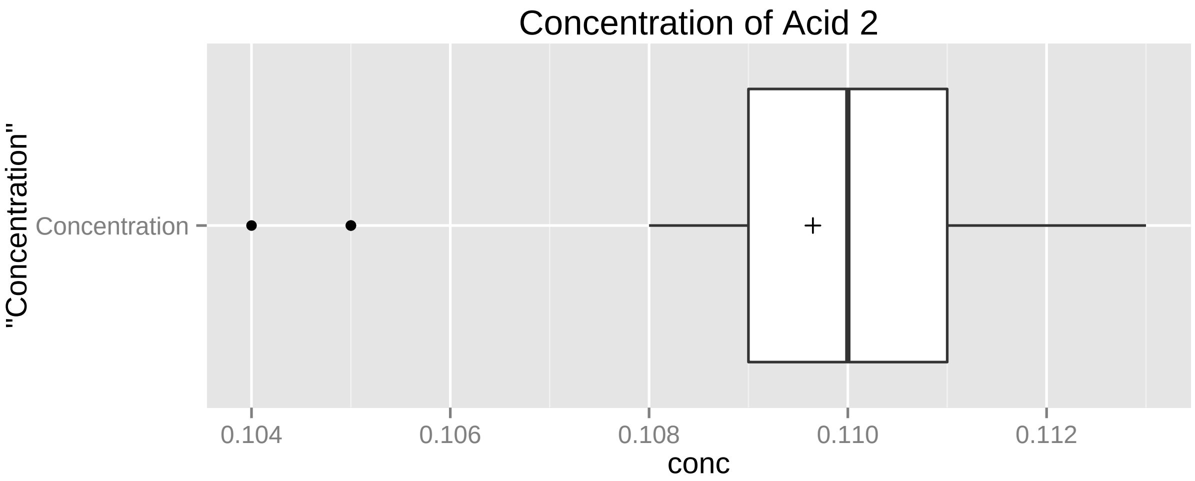
- **define the population parameter:** The population parameter is the mean measured acidity of the solution ($\mu$ = mean acidity of solution).
- **state the hypothesis:** The hypothesis is that the students did not have a bias ($H_0 : \mu = 0.110$ against $H_A : \mu \neq 0.110$).
- **state assumptions and how they will be assessed:** The assumptions for this analysis are that the data are normally distributed. Also, I will assume the entire population was sampled.
- Based on the histogram of the sample data, the data are unimodal, symmetric, and nearly normal. Based on the bootstrap distribution, the sample data are normally distributed. The boxplot of sample data shows outliers do exist, the tails are approximately equal, and the mean is essentially the meadian.

  The sample summaries are $n = 124, \bar{Y} = 0.110, s = 0.005$ and $SE_{\bar{Y}} = 1.28$. $\bar{Y} - M = 0.110 - 0.110 = 0$ is very small relative to the variability, which is consistent with the observed symmetry.
- I fail to reject $H_0$ in favor of $H_A$. This test was performed at a 5% level (i.e. $\alpha = 0.05$), and the p-value (0.8747) is $\geq \alpha$ and accordingly $t_s(0.1581) \leq t_{crit}(1.979)$. The data suggest that $\mu = 0.110$. Additionally, I am 95% confident that the population mean is $0.110 \pm 8e - 4$, and the students were not biased for experiment 1.

## 3(b) Acid 2

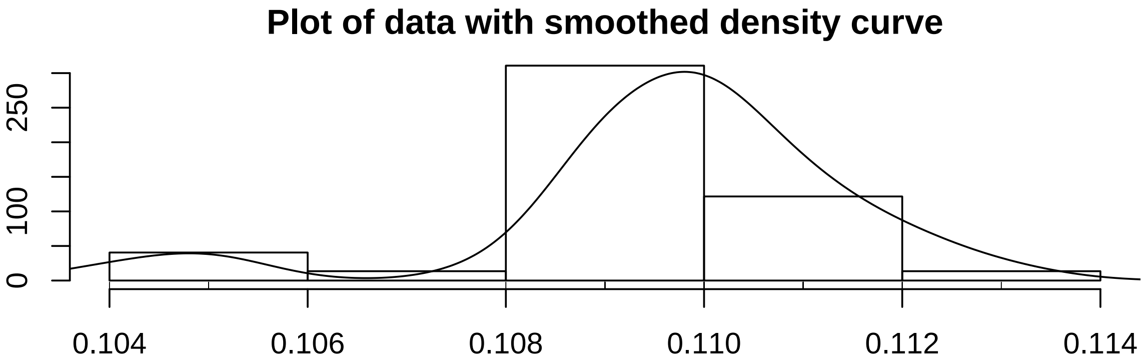
```
# boxplot of Unseeded acid1
acid2.box <- ggplot(acid2, aes(x = "Concentration", y = conc)) # boxplot of acid2
acid2.box <- acid2.box + geom_boxplot()
```

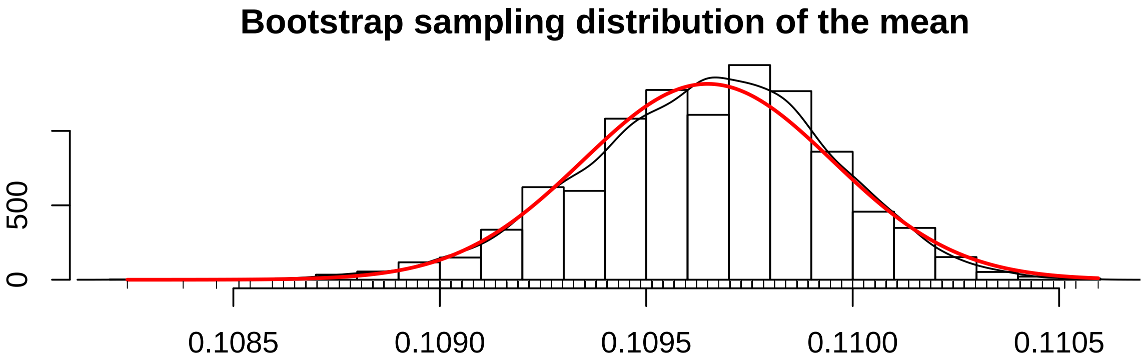
```
acid2.box <- acid2.box + coord_flip()
acid2.box <- acid2.box + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
acid2.box <- acid2.box + labs(title = "Concentration of Acid 2")
acid2.box
```



```
acid2.t <- t.test(acid2$conc, mu = 0.110)  # t-test summary
acid2.t

##
##  One Sample t-test
##
## data:  acid2$conc
## t = -1.159, df = 36, p-value = 0.2541
## alternative hypothesis: true mean is not equal to 0.11
## 95 percent confidence interval:
##  0.1090 0.1103
## sample estimates:
## mean of x
##    0.1096

acid2.ybar <- mean(acid2$conc)  # mean of sample
acid2.s <- sd(acid2$conc)       # standard deviation
acid2.n <- length(acid2$conc)   # observations in sampe
acid2.SEM <- s/sqrt(acid2.n)    # standard error of the mean
acid2.df <- acid2.n - 1         # degrees of freedom
acid2.tcrit <-qt(.025, acid2.df)# t_crit
acid2.M <- median(acid2$conc)   # median
acid2.IQR <- IQR(acid2$conc)
```

```
bs.one.samp.dist(acid2$conc)      # plot histogram and frequency density curve
```





- **define the population parameter:** The population parameter is the mean measured acidity of the solution ($\mu =$ mean acidity of solution).
- **state the hypothesis:** The hypothesis is that the students did not have a bias ($H_0 : \mu = 0.110$ against $H_A : \mu \neq 0.110$).
- **state assumptions and how they will be assessed:** The assumptions for this analysis are that the data are normally distributed, and that the entire population was sampled.
- Based on the histogram of the sample data, the data are unimodal, skewed left, and not normal. Based on the bootstrap distribution, the sample are normally distributed. The boxplot of sample data shows that outliers do exist, the tails are not equal, and the mean noticeably different than the meadian.

  The sample summaries are $n = 37, \bar{Y} = 0.1096, s = 0.002$ and $SE_{\bar{Y}} = 2.34$. $\bar{Y} - M = 0.1096 - 0.110 = -4e - 4$, which is orders of magnitude smaller than the IQR.
- I fail to reject $H_0$ in favor of $H_A$. This test was performed at a 5% level (i.e. $\alpha = 0.05$), and the p-value (0.8747) is $\geq \alpha$ and accordingly $t_s(0.1581) \leq t_{crit}(1.979)$. In a valid t-test, this results would indicate the null hypothesis should be rejected, but because the normality assumption was not met, I choose to accept the null hypothesis.

  $\mu = 0.110$. Additionally, I am 95% confident that the population mean is $0.1096 \pm 0.00065$, and the students were not biased for experiment 2.