ASSIGNMENT 3

Brandon Lampe
STAT 527
Advanced Data Analysis I

October 7, 2014

1 Cloud seeding:

1(a) (10 pts) Carefully check the assumption of normality of the data on the original scale by describing the shape of the data distribution and the sampling distribution of the mean (using the bootstrap). You need to do the seeded and unseeded days separately.

```
# Load libraries and source functions for analysis
library(ggplot2)
library(grid)
library(gridExtra)
library(reshape2)
library(car)
source("ADA1_FUNC.R")

# read in cloud seeding data
d1.initial <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-1.csv")
LogS <- log(d1.initial$seeded)             # log of seeded data
LogUS <- log(d1.initial$unseeded)          # log of unseeded data
d1 <- data.frame(cbind(d1.initial,LogUS, LogS ))   # all data
```

```
# histogram of Unseeded Precip
Precip.hist <- ggplot(d1, aes(x = unseeded))
Precip.hist <- Precip.hist + geom_histogram(binwidth = 100,color = "black",
                                            fill = "white")
Precip.hist <- Precip.hist + labs(title = "Precipitation, Unseeded")

# boxplot of Unseeded Precip
Precip.box <- ggplot(d1, aes(x = "Precipitation",y = unseeded)) # boxplot of Precip
Precip.box <- Precip.box + geom_boxplot()
Precip.box <- Precip.box + coord_flip()
Precip.box <- Precip.box + stat_summary(fun.y = mean, geom = "point",
                                        shape = 3, size = 4)
```

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

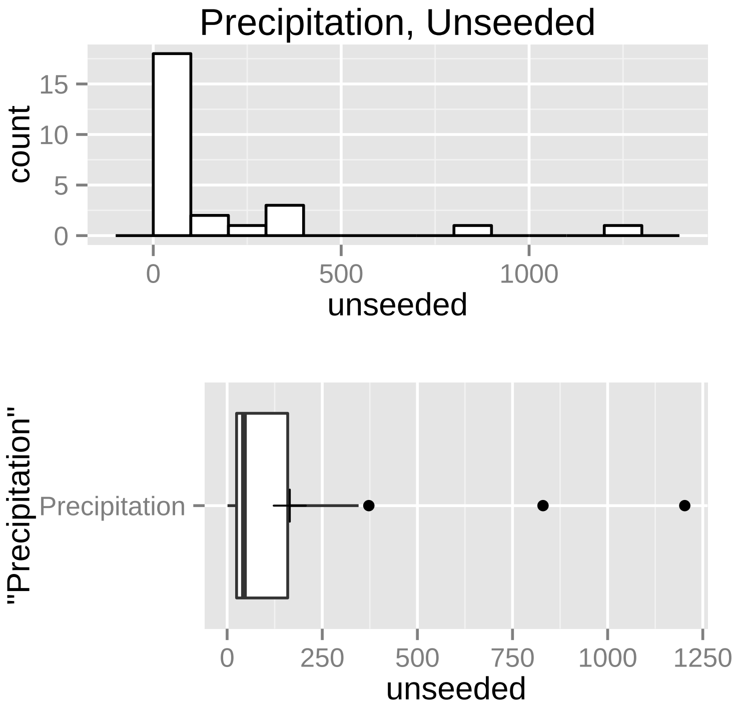


Figure 1: Histogram and Boxplot of Unseeded Data Distribution

```
# qq plot for unseeded data
qqPlot(d1$unseeded, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Precipitation")
```

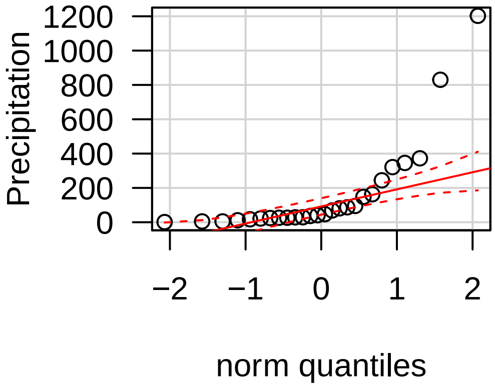
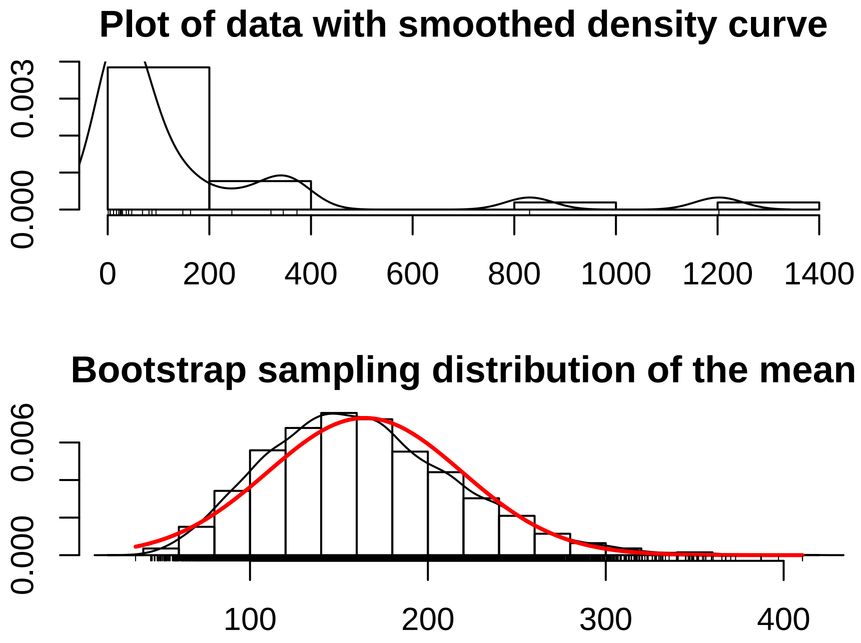


Figure 2: QQ Plot of Unseeded Data Distribution.

- Unseeded Data Distribtion: The unseeded data distribution is unimodal, skewed right, and contains outliers. The mean of the data is outside of the IQR. This shape occurs because only positive valued data are possible. The QQ plot of the data shows the data are not normally distributed, as they deviate substantially from the line representing normality and more than 5% of the data are outside

```
unseed.samp <- bs.one.samp.dist(d1$unseeded)  # bootstrap of unseeded data
```



```
unseed.samp.df <- data.frame(unseed.samp)
```

Figure 3: Resuts of Sampling Distribution from Bootstrap Function of Unseeded Data.

```
# histogram of Unseeded Precip
Samp.hist <- ggplot(unseed.samp.df, aes(x = unseed.samp))
Samp.hist <- Samp.hist + geom_histogram(binwidth = 10,color = "black",
                                            fill = "white")
Samp.hist <- Samp.hist + labs(title = "Precipitation, Bootstrap Sample of Unseeded")

# boxplot of Unseeded Precip
Samp.box <- ggplot(unseed.samp.df, aes(x = "Precipitation",y = unseed.samp)) # boxplot of Precip
Samp.box <- Samp.box + geom_boxplot()
Samp.box <- Samp.box + coord_flip()
Samp.box <- Samp.box + stat_summary(fun.y = mean, geom = "point",
                                        shape = 3, size = 4)
```

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

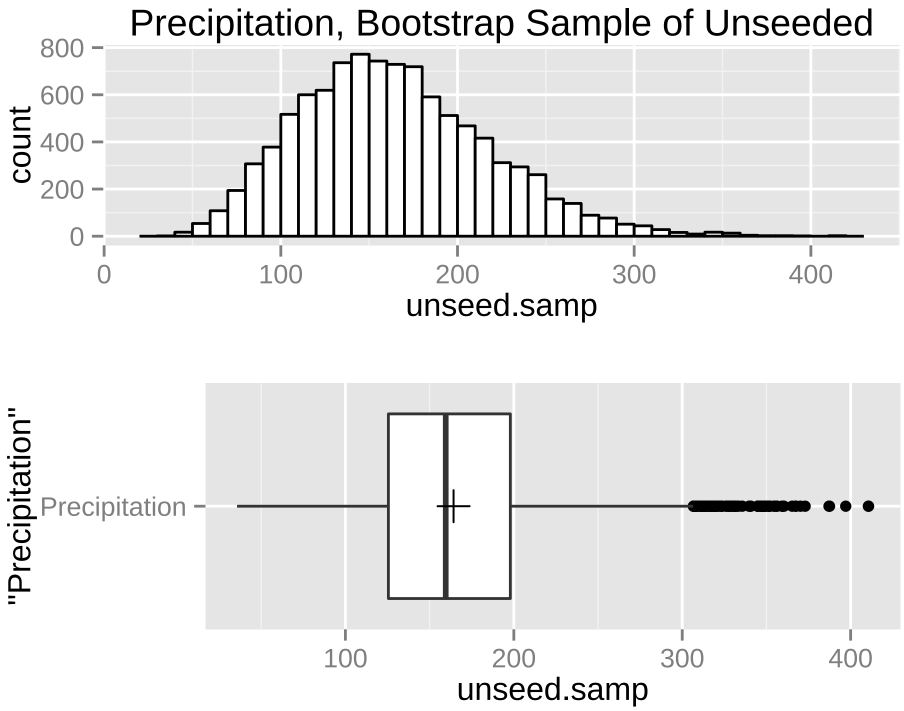


Figure 4: Histogram and Boxplot of Bootstrap Sample Distribution of Unseeded Data.

```
# qq plot for bootstrap sampling distribution of unseeded data
qqPlot(unseed.samp, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Precipitation")
```

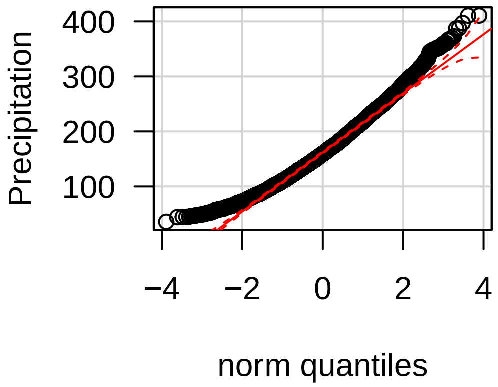


Figure 5: QQ Plot of Bootstrap Sampling Distribution of Unseeded Data.

- Unseeded Bootsrap Sample Distribtion: The sampled distribution is unimodal, skewed right, and contains outliers. The mean of the data is near the mean. The QQ plot of the sample data shows they are not normally distributed; as they deviate substantially from the line representing normality and far more than 5% of the data are outside of the limits.

```
# histogram of Seeded Precip
Precip.hist.seed <- ggplot(d1, aes(x = seeded))
Precip.hist.seed <- Precip.hist.seed + geom_histogram(binwidth = 100,color = "black",
                                            fill = "white")
Precip.hist.seed <- Precip.hist.seed + labs(title = "Precipitation, Seeded")

# boxplot of Seeded Precip
Precip.box.seed <- ggplot(d1, aes(x = "Precipitation",y = seeded)) # boxplot of Precip
Precip.box.seed <- Precip.box.seed + geom_boxplot()
Precip.box.seed <- Precip.box.seed + coord_flip()
Precip.box.seed <- Precip.box.seed + stat_summary(fun.y = mean, geom = "point",
                                        shape = 3, size = 4)
```

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

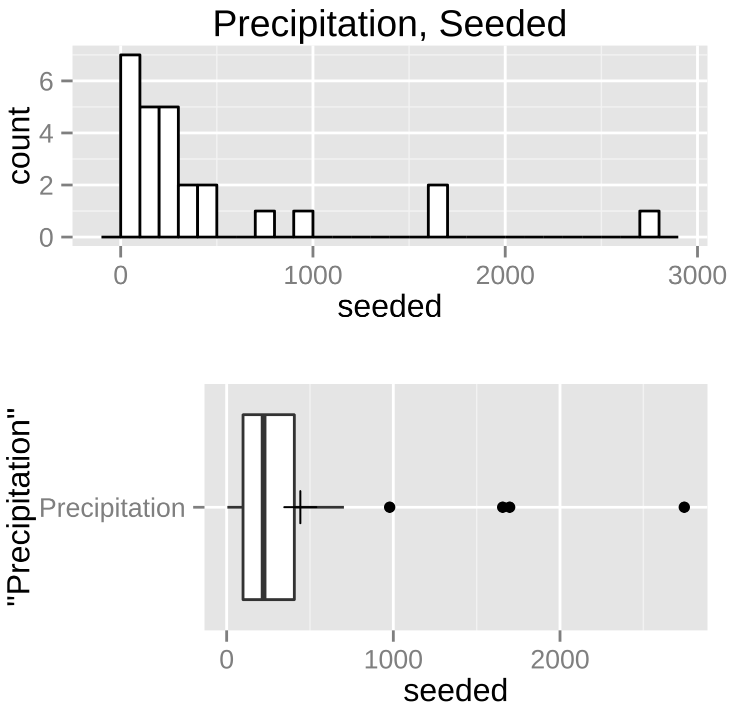


Figure 6: Histogram and Boxplot of Seeded Data Distribution.

```
# qq plot for Seeded data
# las = 1 : turns labels on y-axis to read horizontally
# id.n = n : labels n most extreme observations, and outputs to console
# id.cex = 1 : is the size of those labels
# lwd    =  1 : line width
qqPlot(d1$seeded, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Precipitation")
```

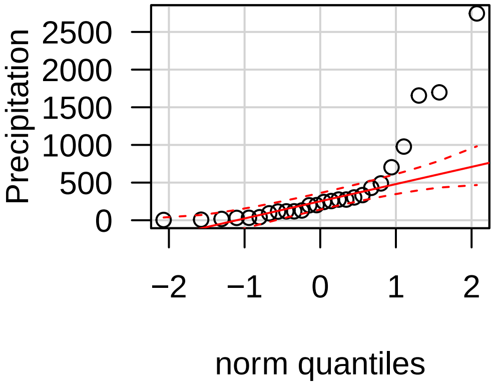
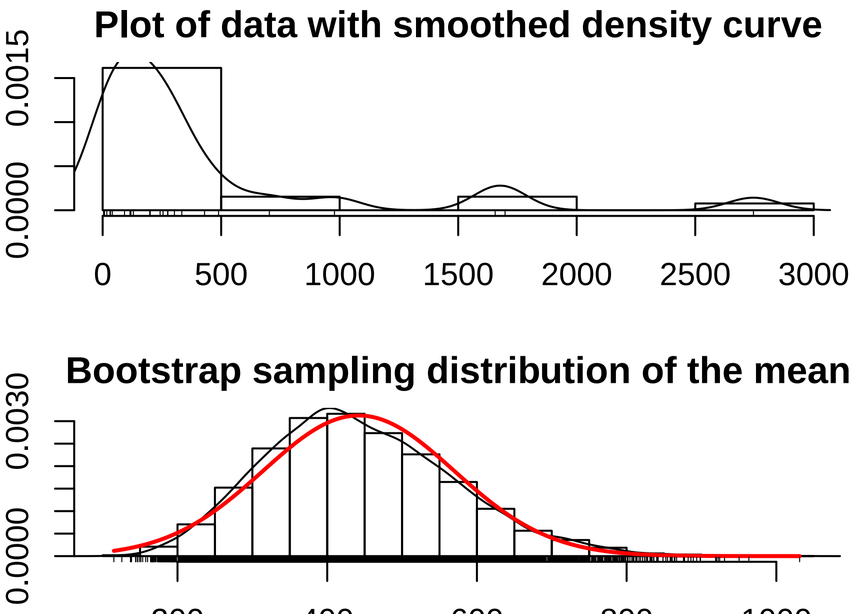


Figure 7: QQ Plot of Seeded Data Distribution.

- Seeded Data Distribtion: The seeded data distribution is unimodal, skewed right, and contains outliers. The mean of the data is outside of the IQR. This shape occurs because only positive valued data are possible. The QQ plot of the data shows the data are not normally distributed, as they deviate substantially from the line representing normality and more than 5% of the data are outisde the limits.

```r
seed.samp <- bs.one.samp.dist(d1$seeded)  # bootstrap of seeded data
```



```r
seed.samp.df <- data.frame(seed.samp)
```

Figure 8: Resuts of Sampling Distribution from Bootstrap Function of Seeded Data.

```r
# histogram of Unseeded Precip
Samp.hist.seed <- ggplot(seed.samp.df, aes(x = seed.samp))
Samp.hist.seed <- Samp.hist.seed + geom_histogram(binwidth = 100,color = "black",
                                         fill = "white")
Samp.hist.seed <- Samp.hist.seed + labs(title = "Precipitation, Bootstrap Sample of Seeded")

# boxplot of Seeded Precip
Samp.box.seed <- ggplot(seed.samp.df, aes(x = "Precipitation",y = seed.samp)) # boxplot of Precip
Samp.box.seed <- Samp.box.seed + geom_boxplot()
Samp.box.seed <- Samp.box.seed + coord_flip()
Samp.box.seed <- Samp.box.seed + stat_summary(fun.y = mean, geom = "point",
                                        shape = 3, size = 4)
```

```
# plot histogram and boxplot
grid.arrange(Samp.hist.seed, Samp.box.seed, nrow = 2)
```

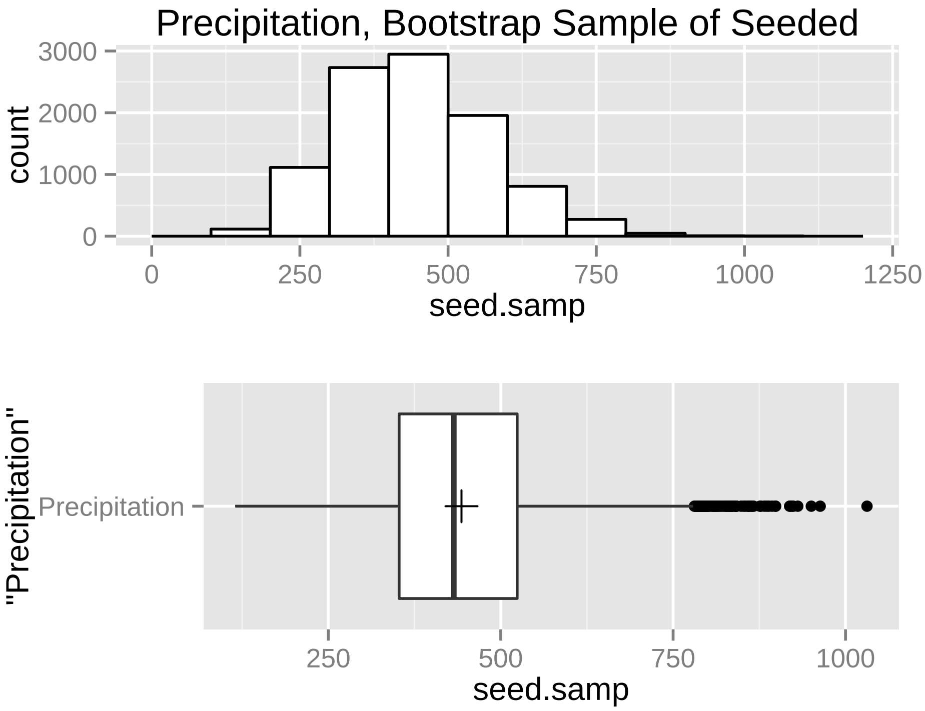


Figure 9: Histogram and Boxplot of Bootstrap Sample Distribution of Seeded Data.

```
# qq plot for bootstrap sampling distribution of seeded data
qqPlot(seed.samp, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Precipitation")
```

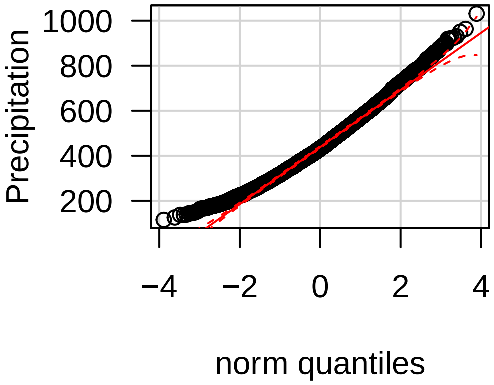


Figure 10: QQ Plot of Bootstrap Sampling Distribution of Seeded Data.

- Seeded Bootsrap Sample Distribtion: The sampled distribution is unimodal, skewed right, and contains outliers. The mean of the data is near the mean. The QQ plot of the sample data shows they are not normally distributed, as more than 5% of the sample data are outiside of the limits.

1(b) (10 pts) Repeat the previous question for the log-transformed data.

```
# histogram of LogUS Precip
Precip.hist <- ggplot(d1, aes(x = LogUS))
Precip.hist <- Precip.hist + geom_histogram(binwidth = 1,color = "black",
                                            fill = "white")
Precip.hist <- Precip.hist + labs(title = "Precipitation, Log Unseeded")

# boxplot of Log Unseeded Precip
Precip.box <- ggplot(d1, aes(x = "Precipitation",y = LogUS)) # boxplot of Precip
Precip.box <- Precip.box + geom_boxplot()
Precip.box <- Precip.box + coord_flip()
Precip.box <- Precip.box + stat_summary(fun.y = mean, geom = "point",
                                        shape = 3, size = 4)

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

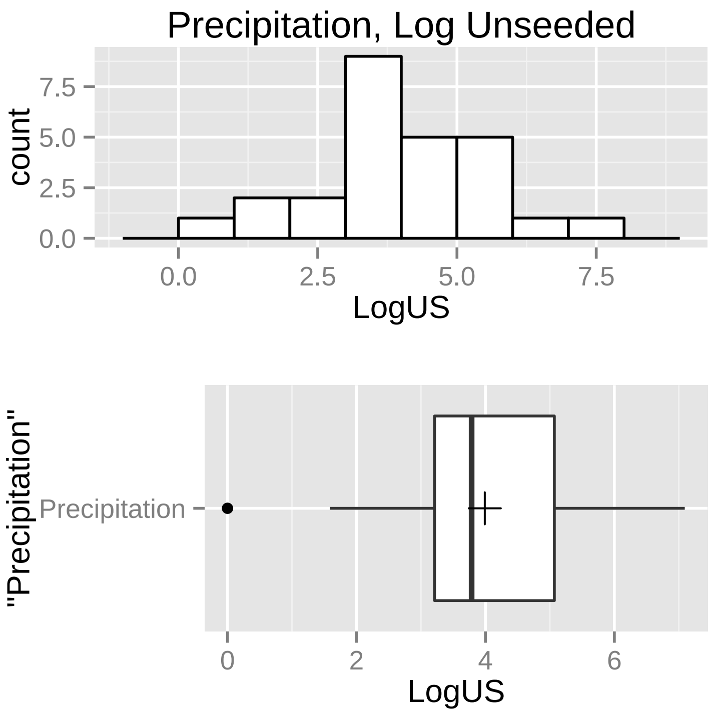


Figure 11: Histogram and Boxplot of Log-Transformed Unseeded Data Distribution.

```
# qq plot for LogUS data
qqPlot(d1$LogUS, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Precipitation")
```

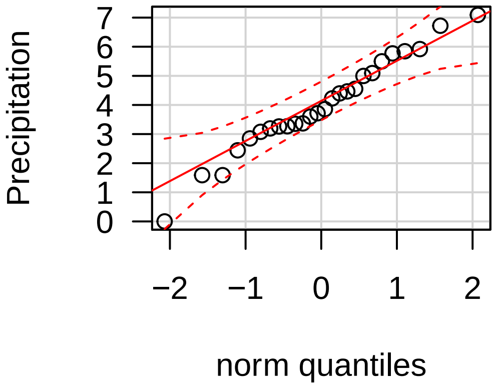
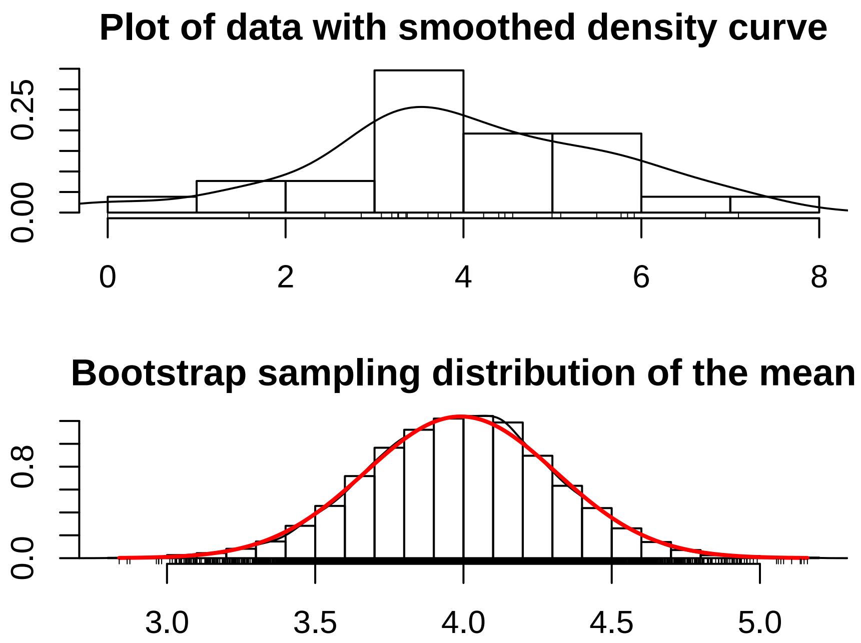


Figure 12: QQ Plot of Log-Transformed Unseeded Data Distribution.

- Log-Transformed Unseeded Data Distribtion: The data distribution is unimodal, normal, and contains one outlier. The mean of the data is inside of the IQR. The distribution does not appear to be affected by the one sidedness of the untransformed data. The QQ plot of the data shows the data are normally distributed, as they closely follow the line representing normality and none of the data are outside of the limits.

```
LogUS.samp <- bs.one.samp.dist(d1$LogUS)  # bootstrap of LogUS data
```



```
LogUS.samp.df <- data.frame(LogUS.samp)
```

Figure 13: Resuts of Sampling Distribution from Bootstrap Function of Log-Transformed Unseeded Data.

```
# histogram of LogUS Precip
Samp.hist.LogUS <- ggplot(LogUS.samp.df, aes(x = LogUS.samp))
Samp.hist.LogUS <- Samp.hist.LogUS + geom_histogram(binwidth = .25,color = "black",
                                            fill = "white")
Samp.hist.LogUS <- Samp.hist.LogUS +
  labs(title = "Precipitation, Bootstrap Sample of Log Unseeded")

# boxplot of LogUS Precip
Samp.box.LogUS <- ggplot(LogUS.samp.df, aes(x = "Precipitation",y = LogUS.samp))
Samp.box.LogUS <- Samp.box.LogUS + geom_boxplot()
Samp.box.LogUS <- Samp.box.LogUS + coord_flip()
Samp.box.LogUS <- Samp.box.LogUS + stat_summary(fun.y = mean, geom = "point",
                                        shape = 3, size = 4)

# plot histogram and boxplot
grid.arrange(Samp.hist.LogUS, Samp.box.LogUS, nrow = 2)
```

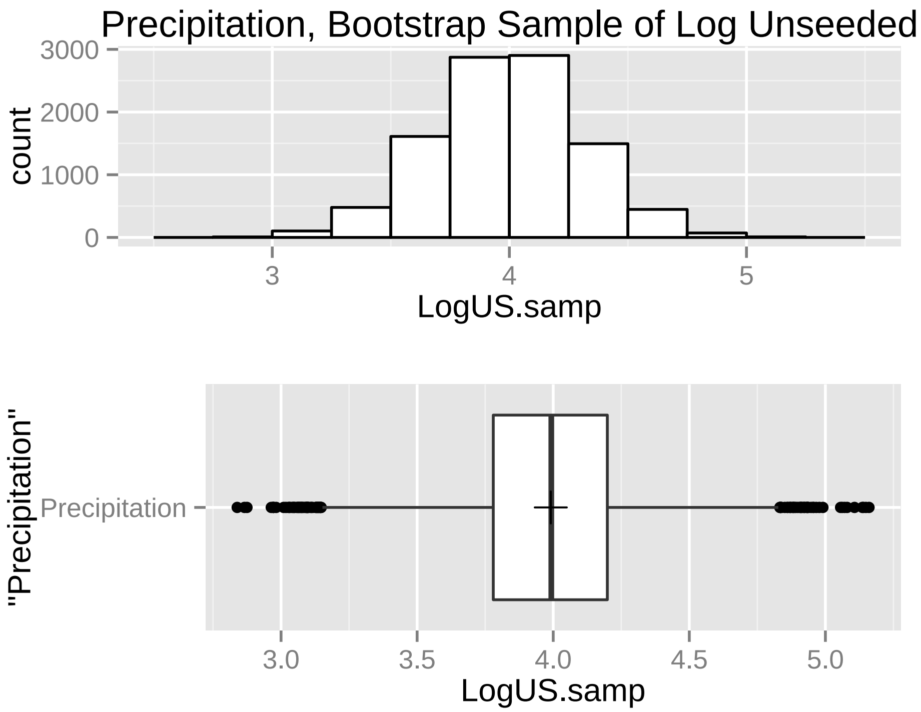


Figure 14: Histogram and Boxplot of Bootstrap Sample Distribution of Log-Transformed Unseeded Data.

```r
# qq plot for bootstrap sampling distribution of LogUS data
qqPlot(LogUS.samp, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Precipitation")
```

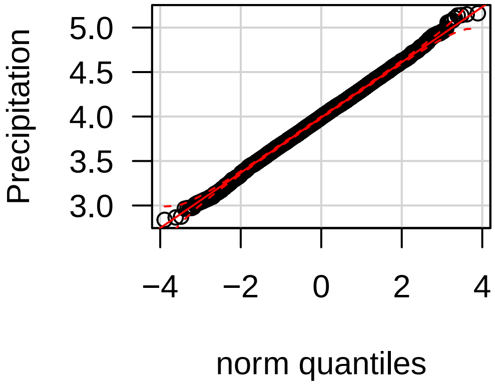


Figure 15: QQ Plot of Bootstrap Sampling Distribution of Log-Transformed Unseeded Data.

- Log-Transformed Unseeded Bootsrap Sample Distribtion: The sampled distribution is unimodal, normal, and contains outliers. The mean of the data is nearly the median. The distribution does not appear to be affected by the one sidedness of the untransformed data. The QQ plot of the data shows the data are normally distributed, as they closely follow the line representing normality.

```r
# histogram of Log Seeded Precip
Precip.hist.LogS <- ggplot(d1, aes(x = LogS))
Precip.hist.LogS <- Precip.hist.LogS + geom_histogram(binwidth = 1,color = "black",
                                                     fill = "white")
Precip.hist.LogS <- Precip.hist.LogS + labs(title = "Precipitation, Log Seeded")

# boxplot of Log Seeded Precip
Precip.box.LogS <- ggplot(d1, aes(x = "Precipitation",y = LogS)) # boxplot of Precip
Precip.box.LogS <- Precip.box.LogS + geom_boxplot()
Precip.box.LogS <- Precip.box.LogS + coord_flip()
Precip.box.LogS <- Precip.box.LogS + stat_summary(fun.y = mean, geom = "point",
                                                 shape = 3, size = 4)
```

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

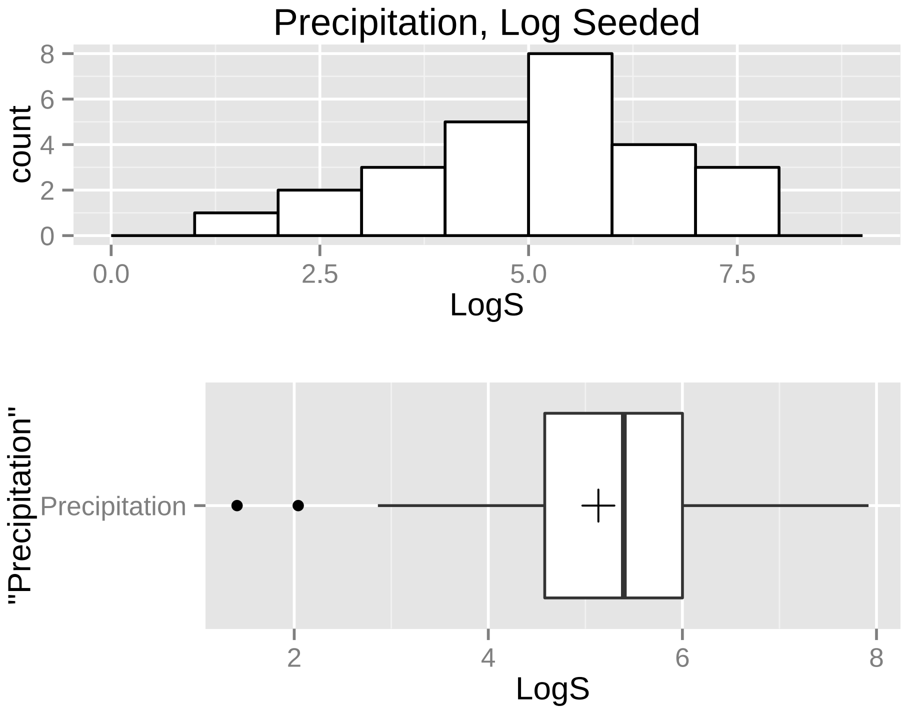


Figure 16: Histogram and Boxplot of Log-Transformed Seeded Data Distribution.

```
# qq plot for Log Seeded data
qqPlot(d1$LogS, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Precipitation")
```

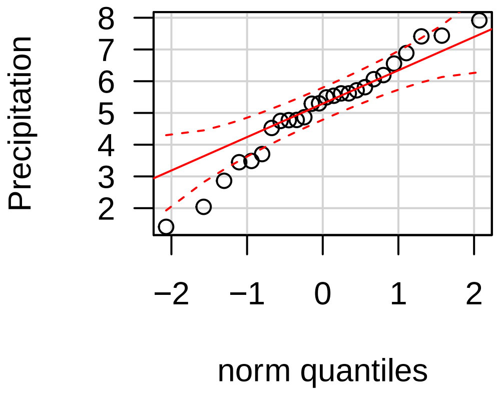


Figure 17: QQ Plot of Log-Transformed Seeded Data Distribution.

- Log-Transformed Seeded Data Distribtion: The distribution is unimodal, has equilength tails, and contains two outliers. The mean of the data is inside the IQR. The QQ plot of the data shows the data are not normally distributed, as over 25% of the data are outside of the CI limits of a normal distribution. If these data were normal, I would exect only 5% to be beyone these limits.

```
LogS.samp <- bs.one.samp.dist(d1$LogS)  # bootstrap of Log seeded data
```

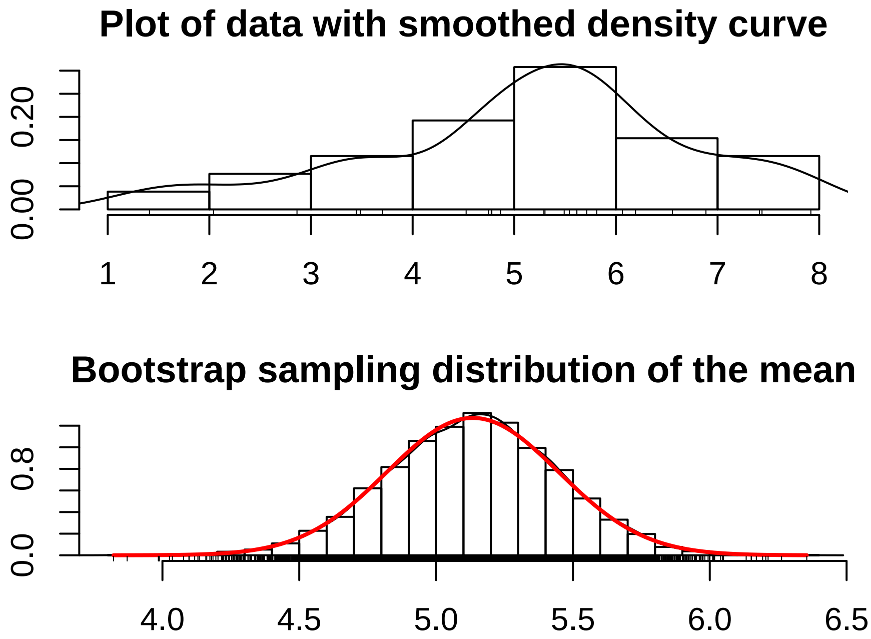


Figure 18: Resuts of Sampling Distribution from Bootstrap Function of Log-Transformed Seeded Data.

```
# histogram of LogUS Precip
LogS.samp.df <- data.frame(LogS.samp)
Samp.hist.LogS <- ggplot(LogS.samp.df, aes(x = LogS.samp))
Samp.hist.LogS <- Samp.hist.LogS + geom_histogram(binwidth = .25,color = "black",
                                         fill = "white")
Samp.hist.LogS <- Samp.hist.LogS +
  labs(title = "Precipitation, Bootstrap Sample of Log Seeded")

# boxplot of Log Seeded Precip
Samp.box.LogS <- ggplot(LogS.samp.df,
                        aes(x = "Precipitation",y = LogS.samp)) # boxplot of Precip
Samp.box.LogS <- Samp.box.LogS + geom_boxplot()
Samp.box.LogS <- Samp.box.LogS + coord_flip()
Samp.box.LogS <- Samp.box.LogS + stat_summary(fun.y = mean, geom = "point",
                                        shape = 3, size = 4)
```

```
# plot histogram and boxplot
grid.arrange(Samp.hist.LogS, Samp.box.LogS, nrow = 2)
```

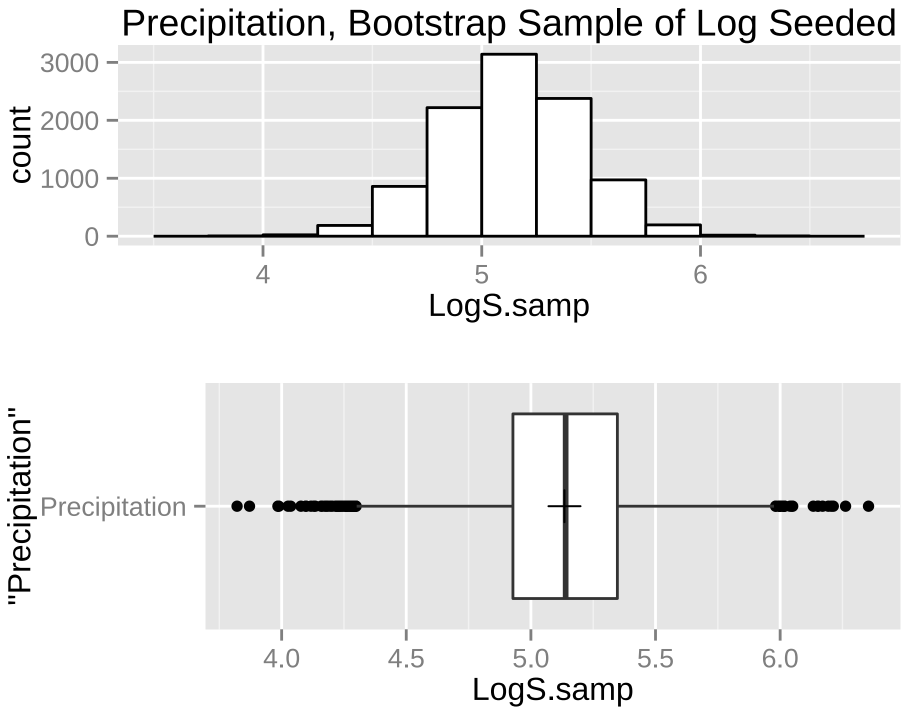


Figure 19: Histogram and Boxplot of Bootstrap Sample Distribution of Log-Transformed Seeded Data.

```
# qq plot for bootstrap sampling distribution of Log seeded data
qqPlot(LogS.samp, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Precipitation")
```

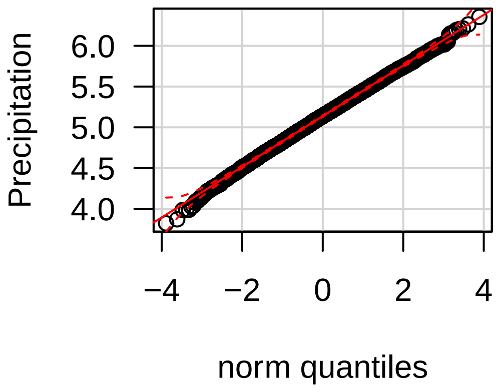


Figure 20: QQ Plot of Bootstrap Sampling Distribution of Log-Transformed Seeded Data.

- Log-Transformed Seeded Bootsrap Sample Distribtion: The sampled distribution is unimodal, normal, and contains outliers. The tails are equilength, and the mean of the data is near the mean. The QQ plot of the sample data shows they are normally distributed.

1(c) (20 pts) Compare the groups using two-sample t-procedures. Choose the most appropriate scale (natural or log units) in which to perform this analysis.

- Scale for analysis: The log-transfered data will be used for the two sample analysis. The standard deviation between the two samples varies by less than 3%; therefore, the pooled variance method.
- Definition of Population Parameters: $\mu = \mu_1 - \mu_2 =$ The population parameter is the difference in the population mean precipitation volume between unseeded and seeded clouds.
- Hypothesis: Is it plausible that the difference in the population mean precipitation volume is different from zero. In notation: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$

```
# summary of statistics
m1 <- mean(LogUS)
s1 <- sd(LogUS)
n1 <- length(LogUS)
m2 <- mean(LogS)
s2 <- sd(LogS)
n2 <- length(LogS)

c(m1, s1, n1) #Unseeded statistics

## [1]  3.990  1.642 26.000

c(m2, s2, n2) #Seeded Statistics

## [1]  5.134  1.600 26.000

# Two sample T Test test with pooled variance
d1.c.t <- t.test(LogUS, LogS, var.equal = TRUE)
d1.c.t

##
##  Two Sample t-test
##
## data:  LogUS and LogS
## t = -2.544, df = 50, p-value = 0.01408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.0467 -0.2409
## sample estimates:
## mean of x mean of y
##     3.990     5.134
```

- Summary: The pooled analysis suggests that the difference in the mean precipitation volume is different than zero. The t-statistic was -2.5 and two-sided p-value was 1.4×10^{-2} therefore, because the p-value is less than 0.05, I reject the Null hypothesis ($H_0 : \mu_1 - \mu_2 = 0$) in favor of the Alternative hypothesis ($H_A : \mu_1 - \mu_2 \neq 0$). The Difference in the population mean precipitation volume between the unseeded and seeded clouds are different.

  With 95% confidence, the difference in the poulation mean precipitation volume ($\mu_1 - \mu_2$) is between -2.0 and -0.2. That is, I am 95% confident that the population mean precipitation volume for seeded clouds μ_2 exceeds the population mean precipiation volume for unseeded clouds μ_1 by betwen 0.2 and 2.0 units.

```r
#### Visual comparison of whether sampling distribution is close to Normal via Bootstrap
bs.two.samp.diff.dist(LogS, LogUS)
```

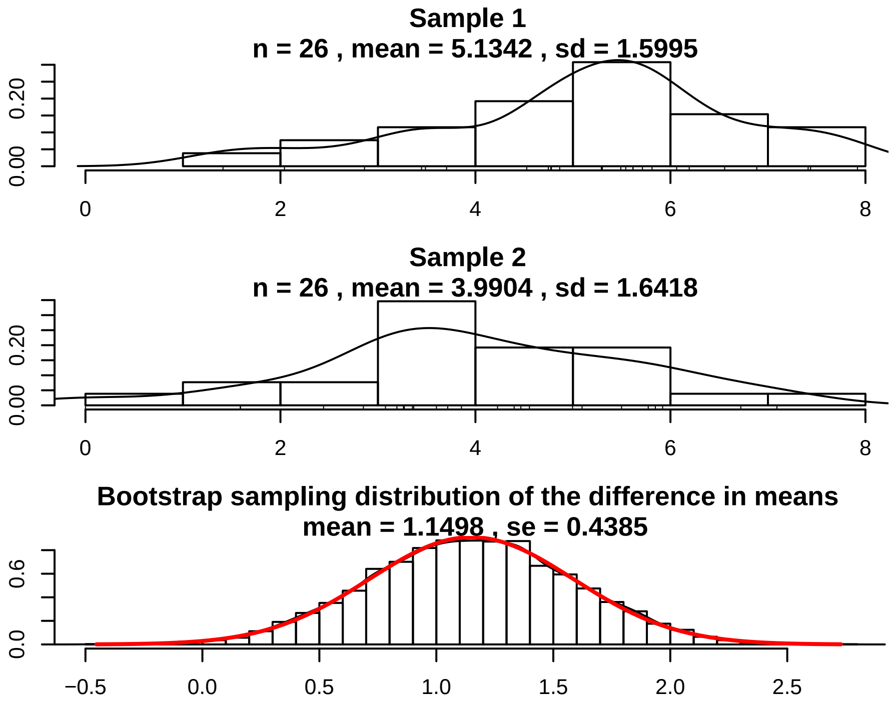


Figure 21: Bootstrap Sampling Distribution of Pooled Log-Transformed Precipitation Data.

- Assumptions: The pooled variance method assumes the populations have normal frequency curves, with equal population standard deviations. As shown above, the distribution of difference in means is very close to normal.

2 Acid

```r
d2 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-3.csv")
acid1 <- subset(d2,exper =="Acid1", select = c(conc,exper))
acid2 <- subset(d2,exper =="Acid2", select = c(conc,exper))
```

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

2(a) (10 pts) Check the normality assumption for both experiments as in problem 1 above.

2(a).1 Acid 1

```r
# histogram of acid1
acid1.hist <- ggplot(acid1, aes(x = conc))
acid1.hist <- acid1.hist + geom_histogram(binwidth = .001)

# boxplot of 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)
```

```
# plot histogram and boxplot
grid.arrange(acid1.hist, acid1.box, nrow = 2)
```

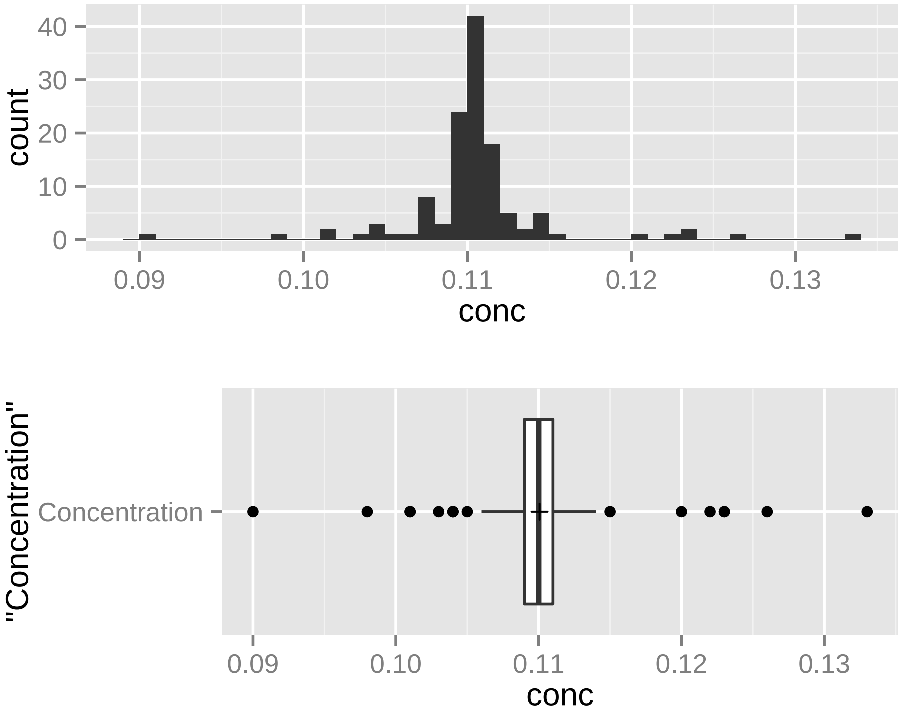


Figure 22: Histogram and Boxplot of Acid 1 Concentration.

```
# qq plot for acid 1 concentration data
qqPlot(acid1$conc, las = 1, id.n = 0, id.cex = 1, lwd = 1, ylab = "Acid 1 Concentration")
```

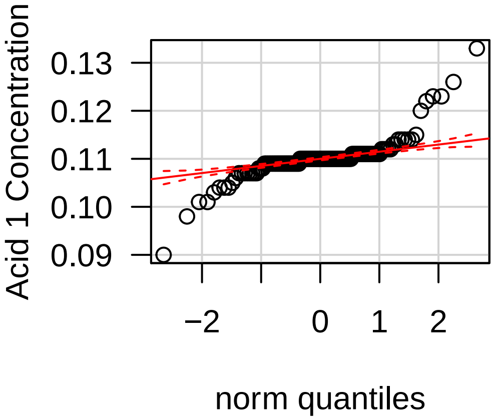


Figure 23: QQ Plot of Acid 1 Concentration Data Distribution.

- Acid 1 Data Distribtion: The Acid 1 data distribution is unimodal, normal, and contains outliers. The mean of the data is inside of the IQR and the tails are nearly equal length. The QQ plot of the data shows the are not normally distributed, as they deviate substantially from the line representing normality and more than 5% of the data are outside of the limits.

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

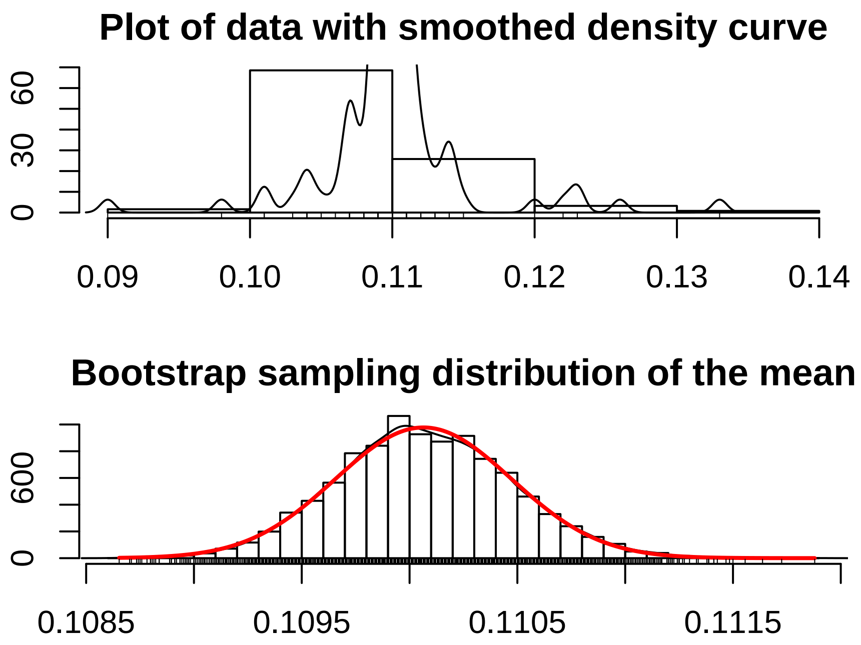


Figure 24: Bootstrap Sample Distribution of Acid 1.

```
acid1.boot.df <- data.frame((acid1.boot))
acid1.boot.df$Acid <- rep("Acid.boot",length(acid1.boot))

# histogram of acid1
acid1.hist.boot <- ggplot(acid1.boot.df, aes(x = X.acid1.boot.))
acid1.hist.boot <- acid1.hist.boot + geom_histogram(binwidth = .0001)

# boxplot of acid1
acid1.box.boot <- ggplot(acid1.boot.df, aes(x = "Concentration", y = X.acid1.boot.)) # boxplot of acid1
acid1.box.boot <- acid1.box.boot + geom_boxplot()
acid1.box.boot <- acid1.box.boot + coord_flip()
acid1.box.boot <- acid1.box.boot + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
```

```r
# plot histogram and boxplot
grid.arrange(acid1.hist.boot, acid1.box.boot, nrow = 2)
```

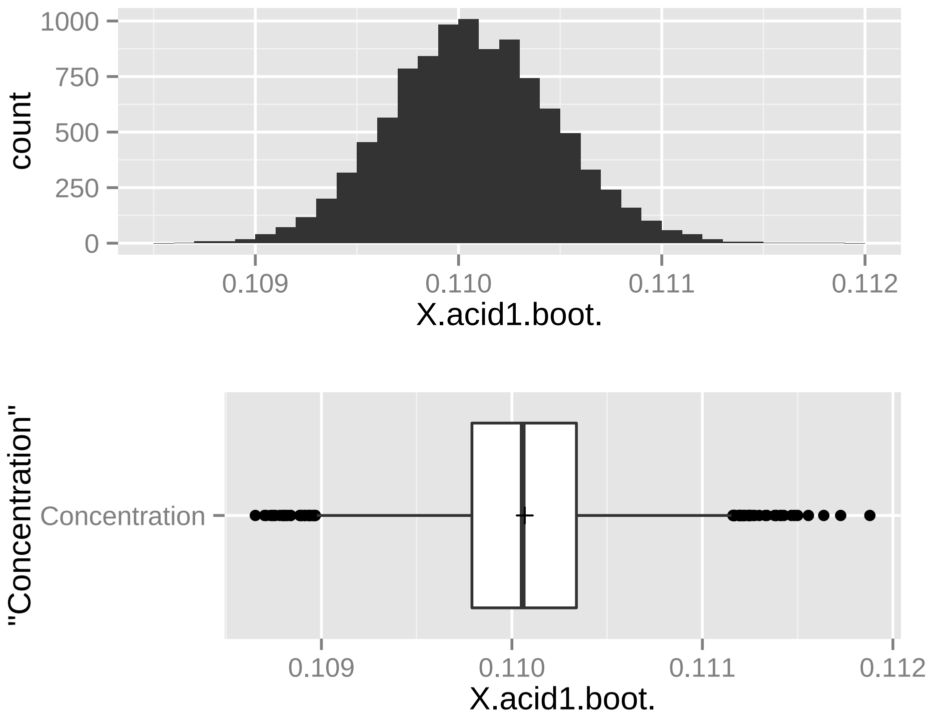


Figure 25: Histogram and Boxplot of Acid 1 Bootstrap Sample Distribution.

```r
# qq plot for acid 1 concentration data
qqPlot(acid1.boot, las = 1, id.n = 0, id.cex = 1, lwd = 1)
```

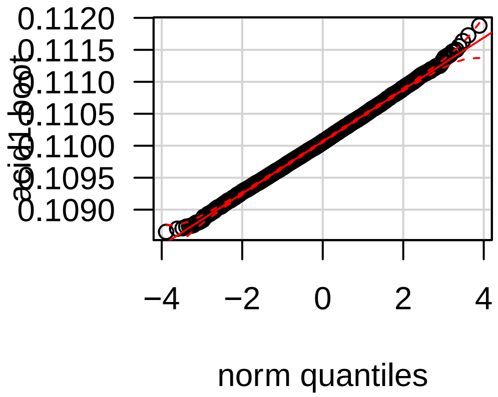


Figure 26: QQ Plot of Acid 1 Concentration Bootstrap Sample Distribution.

- Acid 1 Bootstrap Sample Distribtion: The Acid 1 bootstrap sample distribution is unimodal, normal, and contains outliers. The mean of the data is inside of the IQR and the tails are nearly equal length. The QQ plot of the data shows the are normally distributed, as they follow the line representing normality very closely.

2(a).2 Acid 2

```r
# histogram of acid2
acid2.hist <- ggplot(acid2, aes(x = conc))
acid2.hist <- acid2.hist + geom_histogram(binwidth = .001)

# boxplot of acid2
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)
```

```r
# plot histogram and boxplot
grid.arrange(acid2.hist, acid2.box, nrow = 2)
```

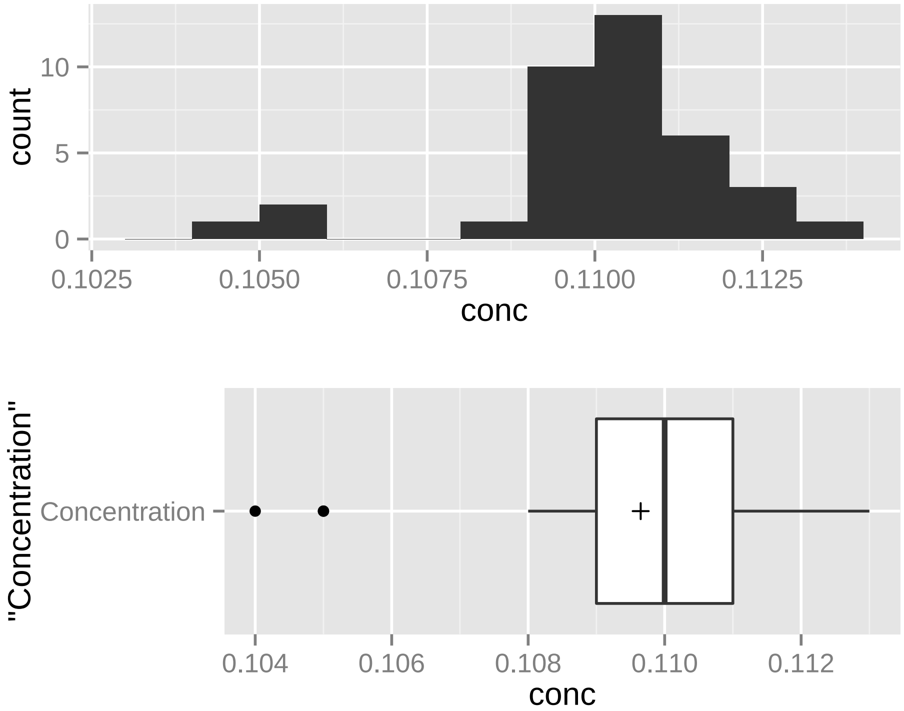


Figure 27: Histogram and Boxplot of Acid 2 Concentration.

```
# qq plot for Acid 2 concentration data
qqPlot(acid2$conc, las = 1, id.n = 0, id.cex = 1, lwd = 1,
       ylab = "Acid 2 Concentration")
```

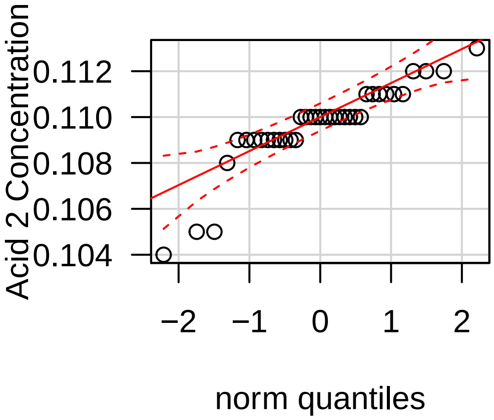


Figure 28: QQ Plot of Acid 2 Concentration Data Distribution.

- Acid 2 Data Distribtion: The Acid 2 data distribution is unimodal, skewed right, and contains outliers to the left. The mean of the data is inside of the IQR but the tails are not equal length. The QQ plot of the data shows the are nearly normally distributed, as 8% of the data are outside of the limits.

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

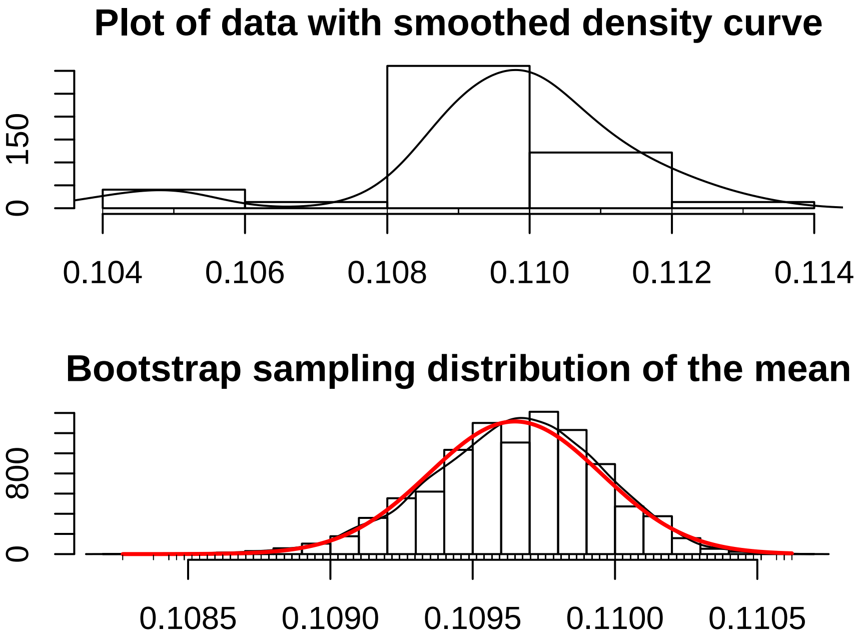


Figure 29: Bootstrap Sample Distribution of Acid 2.

```
acid2.boot.df <- data.frame((acid2.boot))
acid2.boot.df$Acid <- rep("Acid.boot",length(acid2.boot))

# histogram of acid2
acid2.hist.boot <- ggplot(acid2.boot.df, aes(x = X.acid2.boot.))
acid2.hist.boot <- acid2.hist.boot + geom_histogram(binwidth = .0001)

# boxplot of acid2
acid2.box.boot <- ggplot(acid2.boot.df, aes(x = "Concentration", y = X.acid2.boot.))
acid2.box.boot <- acid2.box.boot + geom_boxplot()
acid2.box.boot <- acid2.box.boot + coord_flip()
acid2.box.boot <- acid2.box.boot + stat_summary(fun.y = mean, geom = "point",
                                                shape = 3, size = 2)
```

```
# plot histogram and boxplot
grid.arrange(acid2.hist.boot, acid2.box.boot, nrow = 2)
```

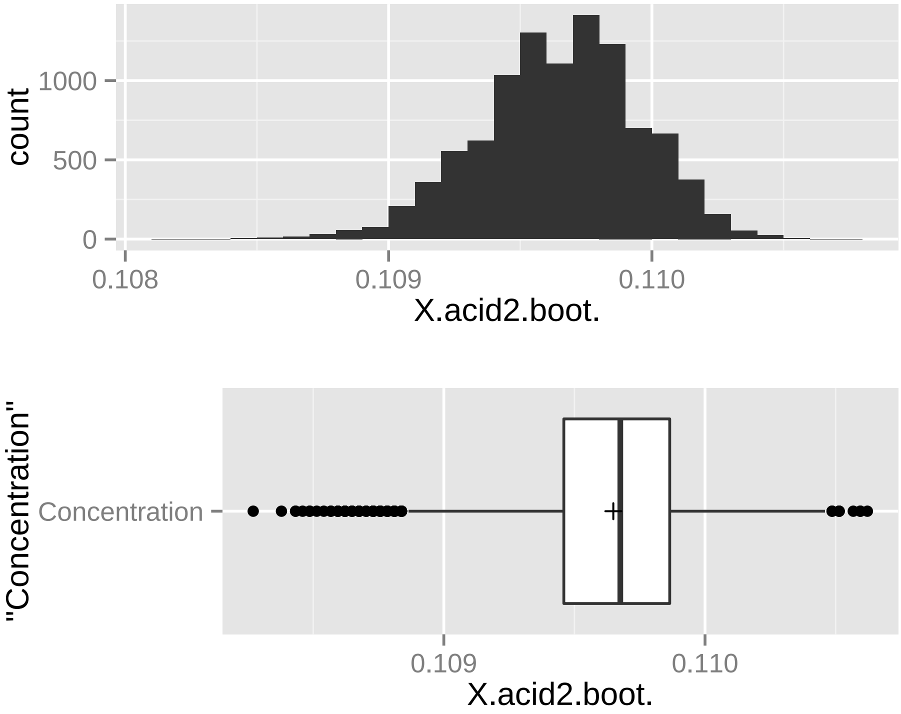


Figure 30: Histogram and Boxplot of Acid 2 Bootstrap Sample Distribution.

```
# qq plot for Acid 2 concentration data
qqPlot(acid2.boot, las = 1, id.n = 0, id.cex = 1, lwd = 1)
```

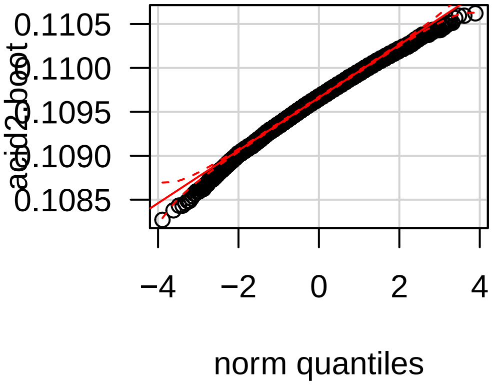


Figure 31: QQ Plot of Acid 2 Concentration Bootstrap Sample Distribution.

- Acid 2 Bootstrap Sample Distribtion: The Acid 2 bootstrap sample distribution is unimodal ish, normal, and contains outliers. The mean of the data is inside of the IQR and the tails are nearly equal length. The QQ plot of the data shows the are normally distributed, as they follow the line representing normality very closely.

2(b) (20 pts)Formally compare the experiments using two-sample t-procedures.

- Definition of Population Parameters: $\mu = \mu_1 - \mu_2 =$ The population parameter is the difference in the population mean acid concentration between Acid 1 and Acid 2.

- Hypothesis: Is it plausible that the difference in the population mean acid concentration is different from zero. In notation: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$

```
# summary of statistics
m1 <- mean(acid1$conc)
s1 <- sd(acid1$conc)
n1 <- length(acid1$conc)
m2 <- mean(acid2$conc)
s2 <- sd(acid2$conc)
n2 <- length(acid2$conc)

c(m1, s1, n1) #Acid 1 statistics

## [1] 1.101e-01 4.544e-03 1.240e+02

c(m2, s2, n2) #Acid 2 Statistics

## [1]  0.109649  0.001844 37.000000

# Two sample T Test test with pooled variance
d2.t <- t.test(acid1$conc, acid2$conc, var.equal = TRUE)
d2.t

##
##  Two Sample t-test
##
## data:  acid1$conc and acid2$conc
## t = 0.5426, df = 159, p-value = 0.5882
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001098  0.001930
## sample estimates:
## mean of x mean of y
##    0.1101    0.1096
```

- Summary: The pooled analysis suggests that the difference in the mean precipitation volume is zero. The t-statistic was 0.5 and two-sided p-value was 0.6 therefore, because the p-value is greater than 0.05, I fail to reject the Null hypothesis ($H_0 : \mu_1 - \mu_2 = 0$) in favor of the Alternative hypothesis ($H_A : \mu_1 - \mu_2 \neq 0$). The Difference in the population mean acid concentration between Acid 1 and Acid 2 are not different.

  With 95% confidence, the difference in the poulation mean acid concentration ($\mu_1 - \mu_2$) is between -1.1×10^{-3} and 2.0×10^{-3} That is, I am 95% confident that the population mean acid concentration for Acid 1 μ_1 does not exceed the population mean acid concentration for Acid 2 μ_2.

```
#### Visual comparison of whether sampling distribution is close to Normal via Bootstrap
acid1.num <- acid1[,1]
acid2.num <- acid2[,1]
bs.two.samp.diff.dist(acid1.num, acid2.num)
```

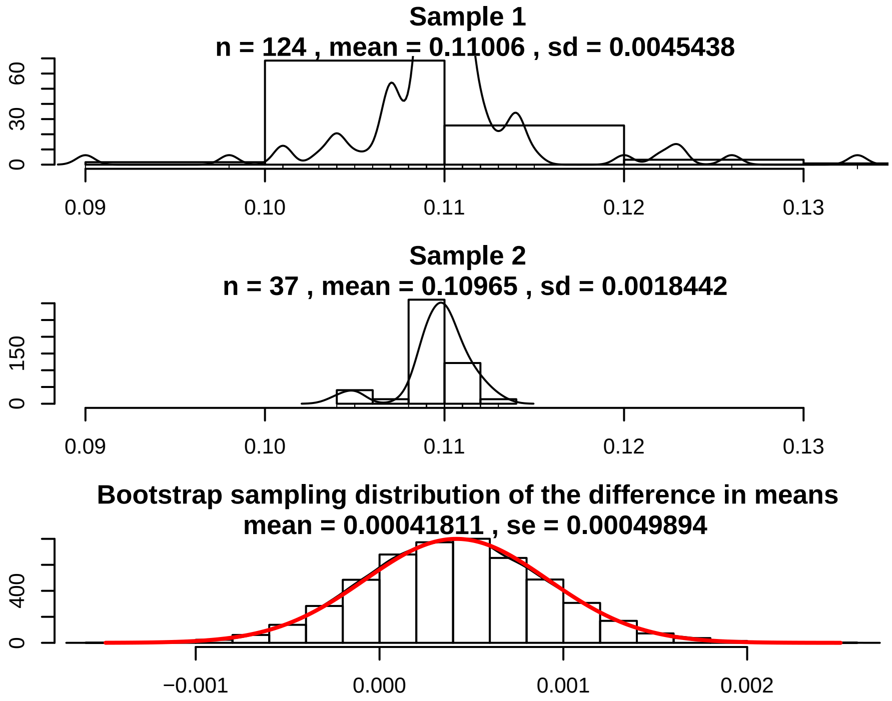


Figure 32: Bootstrap Sampling Distribution of Pooled Acid Data.

- Assumptions: The pooled variance method assumes the populations have normal frequency curves, with similar population standard deviations. As shown above, the distribution of difference in means is very close to normal.

3 cAMP

```
d3 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_03_F14-3.csv")
Dif <- d3$Control - d3$Progesterone
frog <- data.frame(cbind(d3,Dif))
```

The population parameter is the average difference between the cAMP levels for the control and progesterone samples.

3(a) (10 pts) Make a histogram and box plot of the differences between the cAMP levels for the control and progesterone samples.

```
# histogram of frog
frog.hist <- ggplot(frog, aes(x = Dif))
frog.hist <- frog.hist + geom_histogram(binwidth = .1)

# boxplot of frog
frog.box <- ggplot(frog, aes(x = "Difference in cAMP", y = Dif)) # boxplot of frog
frog.box <- frog.box + geom_boxplot()
frog.box <- frog.box + coord_flip()
frog.box <- frog.box + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
```

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

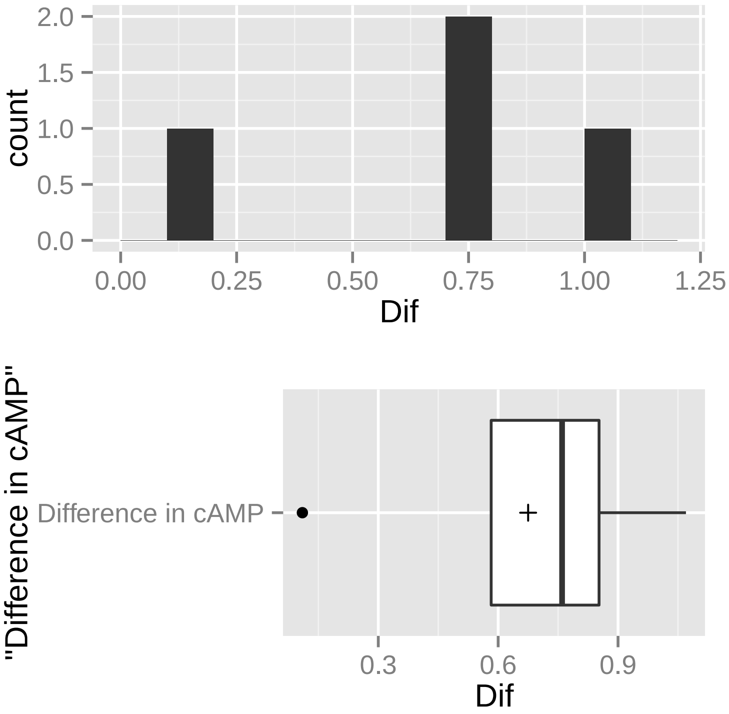


Figure 33: Histogram and Boxplot of Difference.

- Differences Between the cAMP Levels: The distribution is uniform, contains outliers, and is skewed right. The mean of the data is inside of the IQR.

3(b) **(20 pts) Test at the 10% level whether there is any difference in the population mean cAMP levels for batches of oocytes that are untreated versus those treated with progesterone.**

```r
frog.boot <- bs.one.samp.dist(frog$Dif)      # plot histogram and frequency density curve
```

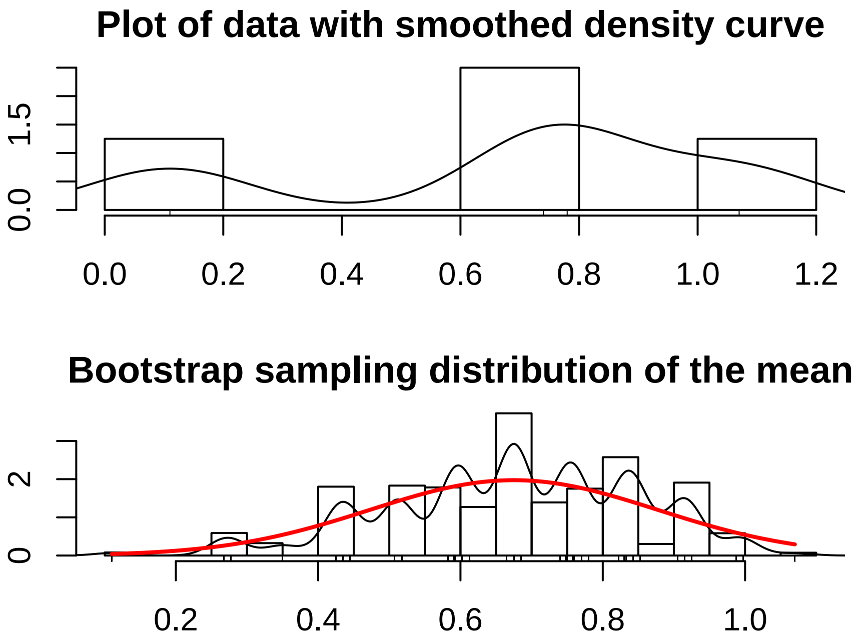


Figure 34: Bootstrap Sample Distribution of the cAMP difference.

```r
frog.boot.df <- data.frame((frog.boot))

# histogram of frog
frog.hist.boot <- ggplot(frog.boot.df, aes(x = X.frog.boot.))
frog.hist.boot <- frog.hist.boot + geom_histogram(binwidth = .05)

# boxplot of frog
frog.box.boot <- ggplot(frog.boot.df, aes(x = "Concentration", y = X.frog.boot.)) # boxplot of frog
frog.box.boot <- frog.box.boot + geom_boxplot()
frog.box.boot <- frog.box.boot + coord_flip()
frog.box.boot <- frog.box.boot + stat_summary(fun.y = mean, geom = "point", shape = 3, size = 2)
```

```
# plot histogram and boxplot
grid.arrange(frog.hist.boot, frog.box.boot, nrow = 2)
```

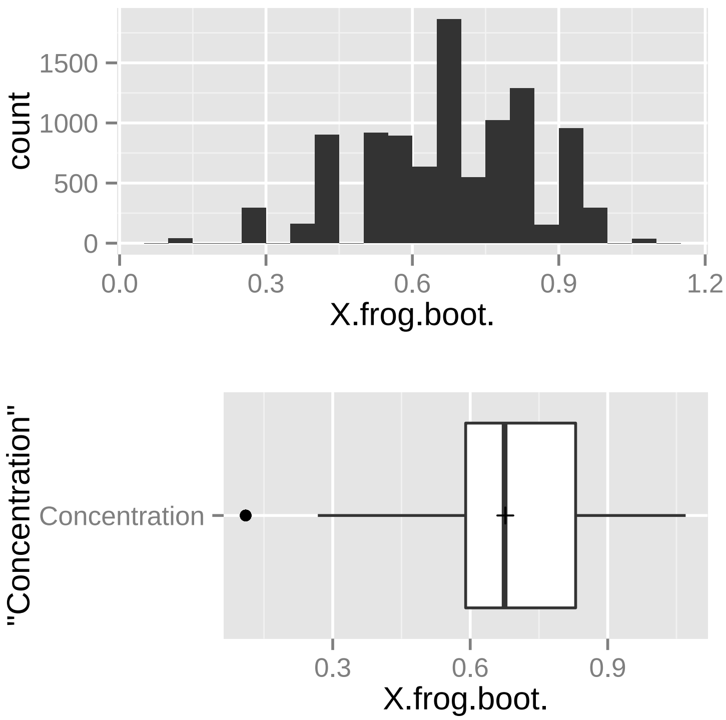


Figure 35: Histogram and Boxplot of cAMP Difference Bootstrap Sample Distribution.

```
# qq plot for acid 1 concentration data
qqPlot(frog.boot, las = 1, id.n = 0, id.cex = 1, lwd = 1)
```

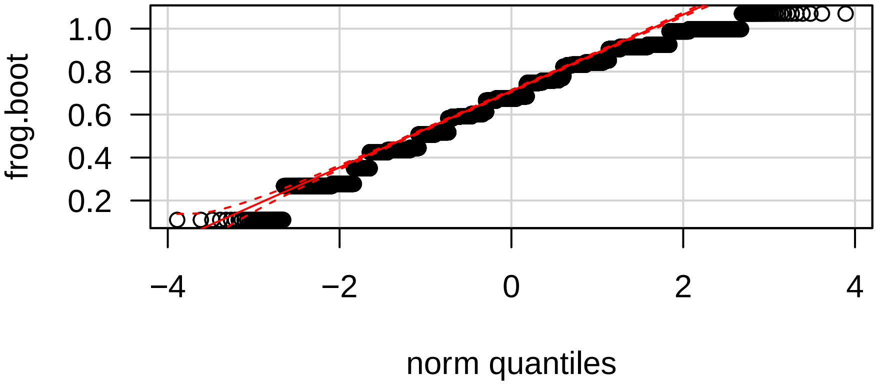


Figure 36: QQ Plot of cAMP Difference Bootstrap Sample Distribution.

- cAMP Bootstrap Sample Distribtion: The bootstrap sample distribution is unimodal, not normal, and contains outliers. The mean of the data is inside of the IQR and the left tail is longer than the right. The QQ plot of the data shows the are not normally distributed. Because the sample differences are not from a normal population, a one sample technique should not be used.

- Definition of Population Parameters: $\mu = \mu_1 - \mu_2 =$ The population parameter is the difference in the population mean cAMP levels between the oocytes not exposed to progesterone (control) and oocytes exposed to progesterone.

- Hypothesis: Is it plausible that the difference in the population mean cAMP Level is different from zero. In notation: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$

- Results: The paired analysis suggests that the difference in the mean precipitation volume is zero. The t-statistic was 3.3 and two-sided p-value was 0.04 therefore, because the p-value is less than 0.1, I reject the Null hypothesis ($H_0 : \mu_1 - \mu_2 = 0$) in favor of the Alternative hypothesis ($H_A : \mu_1 - \mu_2 \neq 0$). The Difference in the population mean cAMP level between untreated oocytes (control) and those treated with progesterone are different.

```
# summary of statistics
m1 <- mean(frog$Control)
s1 <- sd(frog$Control)
n1 <- length(frog$Control)
m2 <- mean(frog$Progesterone)
s2 <- sd(frog$Progesterone)
n2 <- length(frog$Progesterone)

c(m1, s1, n1) #Control
## [1] 2.980 2.047 4.000
c(m2, s2, n2) #Progesterone
## [1] 2.305 1.952 4.000
# Paired Two sample T Test with CI at 90%
frog.t <- t.test(frog$Control, frog$Progesterone, paired = TRUE, conf.level = 0.9)
frog.t
##
##  Paired t-test
##
## data:  frog$Control and frog$Progesterone
## t = 3.339, df = 3, p-value = 0.04443
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  0.1992 1.1508
## sample estimates:
## mean of the differences
##                   0.675
```

3(c) (10 pts) Compute and interpret a 90% CI for the difference in population mean cAMP levels for batches of oocytes that are untreated versus those treated with progesterone.

With 90% confidence, the difference in the poulation mean cAMP level ($\mu_1 - \mu_2$) is between 0.2 and 1.2. That is, I am 90% confident that the population mean cAMP level for untreated oocytes (control) μ_1 exceeds the population mean cAMP level for treated oocytes μ_2 by between 0.2 and 1.2 pmol/oocyte.

3(d) (10 pts) Discuss any statistical assumptions that you have made in carrying out the analysis, and whether the assumptions seem reasonable.

The paired variance method assumes the paired differences have normal frequency curves, with similar population standard deviations. As shown above, the distribution of difference in means is not normal. We also assume the data were randomly sampled.

3(e) (10 pts) Write a short summary to the problem.

The paired analysis suggests that the difference in the mean precipitation volume is zero. The t-statistic was 3.3 and two-sided p-value was 0.04 therefore, because the p-value is less than 0.1, I reject the Null hypothesis ($H_0 : \mu_1 - \mu_2 = 0$) in favor of the Alternative hypothesis ($H_A : \mu_1 - \mu_2 \neq 0$). The Difference in the population mean cAMP level between untreated oocytes (control) and those treated with progesterone are different. That is, I am 90% confident that the population mean cAMP level for untreated oocytes (control) μ_1 exceeds the population mean cAMP level for treated oocytes μ_2 by between 0.2 and 1.2 pmol/oocyte.