Advanced Data Analysis 1, Stat 427/527 Homework 3 Name: Answer Key
Fall 2014 Prof. Erik B. Erhardt

Part I. (130 points) Do all calculations in R. All R code for the assignment should be included with the part of the problem it addresses (for code and output use a fixed-width font, such as Courier). Code is used to calculate result. Text is used to report and interpret results. Do not report or interpret results in the code. Also:

1. Clearly define population parameters in each problem. That is, give a verbal description of what the population mean is in the context of the problem.
2. Clearly specify hypotheses when appropriate (not every problem involves a test of hypothesis).
3. Write a coherent conclusion based on each CI or test.

(40 pts) **1. Cloud seeding:** Return to the cloud seeding problem of Homework 2. This really is a two-sample problem, although we analyzed it in HW 2 as two one-sample problems.

unseeded	seeded
1202.6	2745.6
830.1	1697.8
372.4	1656
345.5	978
321.2	703.4
244.3	489.1
163	430
147.8	334.1
95	302.8
87	274.7
81.2	274.7
68.5	255
47.3	242.5
41.1	200.7
36.6	198.6
29	129.6
28.6	119
26.3	118.3
26.1	115.3
24.4	92.4
21.7	40.6
17.3	32.7
11.5	31.4
4.9	17.5
4.9	7.7
1	4.1

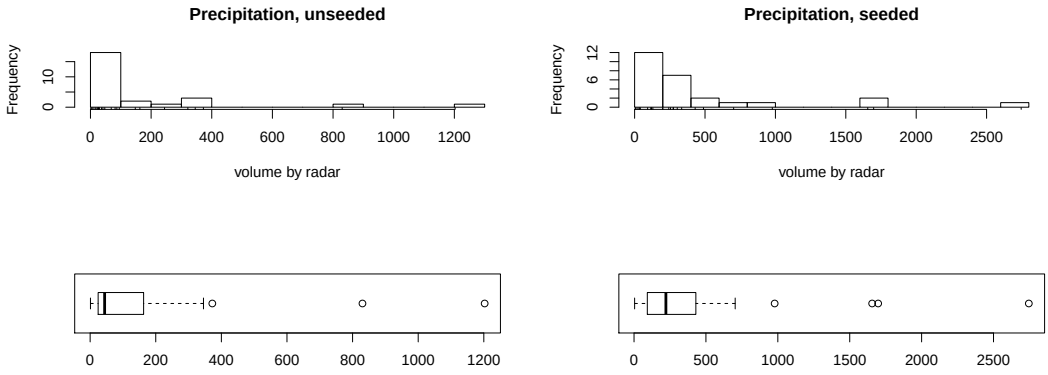
Read the data from the website with:

```
d1 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-1.csv")
```

(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.

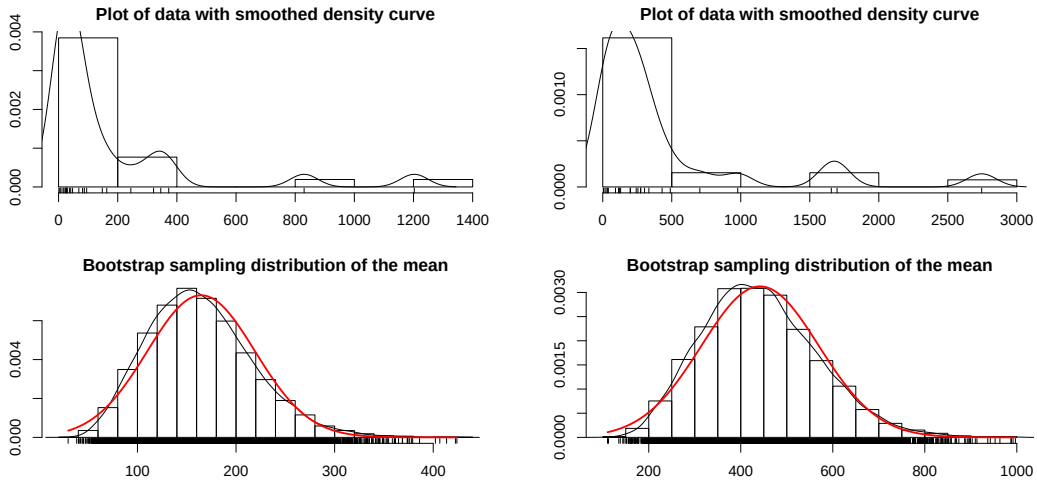
Solution: Both distributions are unimodal, skewed right, with a few outliers.

```
par(mfrow=c(2,1))
hist(d1$unseeded, breaks = 10, main = "Precipitation, unseeded"
    , xlab = "volume by radar")
rug(d1$unseeded)
boxplot(d1$unseeded, horizontal=TRUE)

par(mfrow=c(2,1))
hist(d1$seeded, breaks = 10, main = "Precipitation, seeded"
    , xlab = "volume by radar")
rug(d1$seeded)
boxplot(d1$seeded, horizontal=TRUE)
```



The sampling distributions for the sample mean appear skewed right; the normality assumption does not hold.

```
bs.one.samp.dist(d1$unseeded)
bs.one.samp.dist(d1$seeded)
```

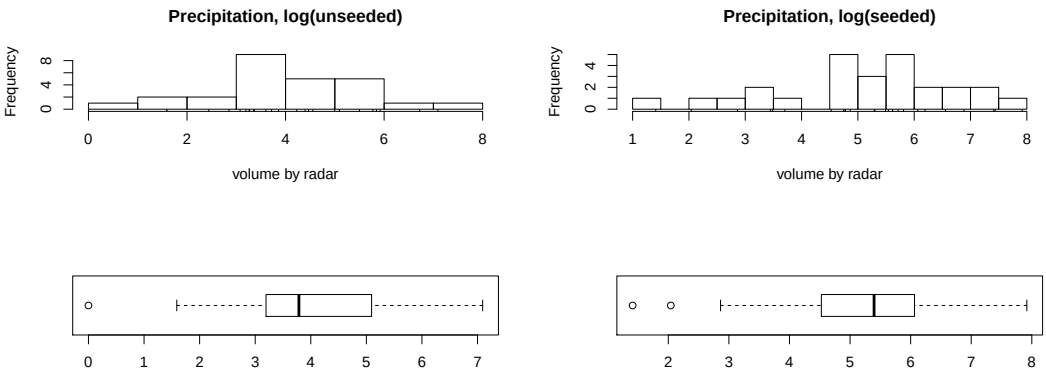


(b) (10 pts) Repeat the previous question for the log-transformed data.

Solution: Both log distributions are unimodal, symmetric, with one or no outliers. These data are consistent with being normally distributed.

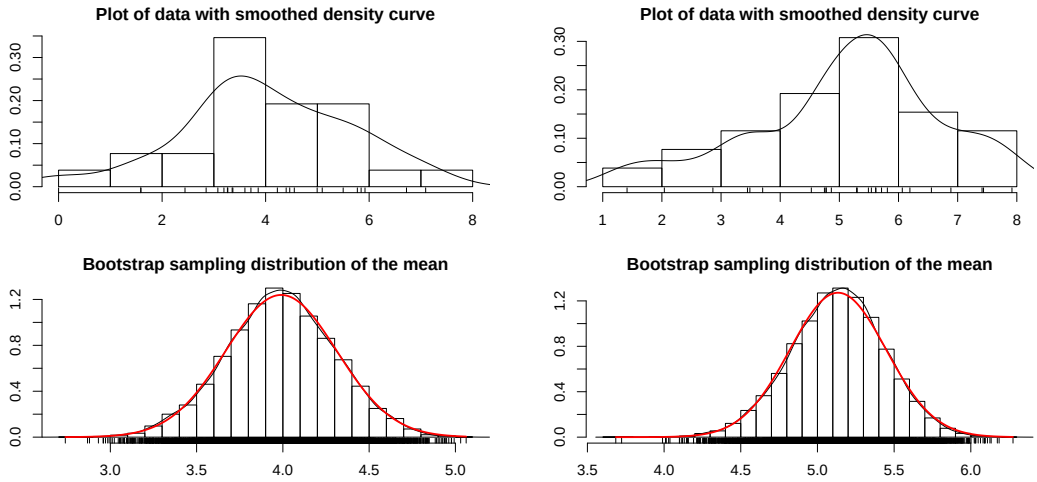
```
# create two new columns on log scale
d1$logunseeded <- log(d1$unseeded)
d1$logseeded <- log(d1$seeded)
```

```
par(mfrow=c(2,1))
hist(d1$logunseeded, breaks = 10, main = "Precipitation, log(unseeded)"
    , xlab = "volume by radar")
rug(d1$logunseeded)
boxplot(d1$logunseeded, horizontal=TRUE)

par(mfrow=c(2,1))
hist(d1$logseeded, breaks = 10, main = "Precipitation, log(seeded)"
    , xlab = "volume by radar")
rug(d1$logseeded)
boxplot(d1$logseeded, horizontal=TRUE)
```



Furthermore, the sampling distributions for the sample mean are nicely normal.

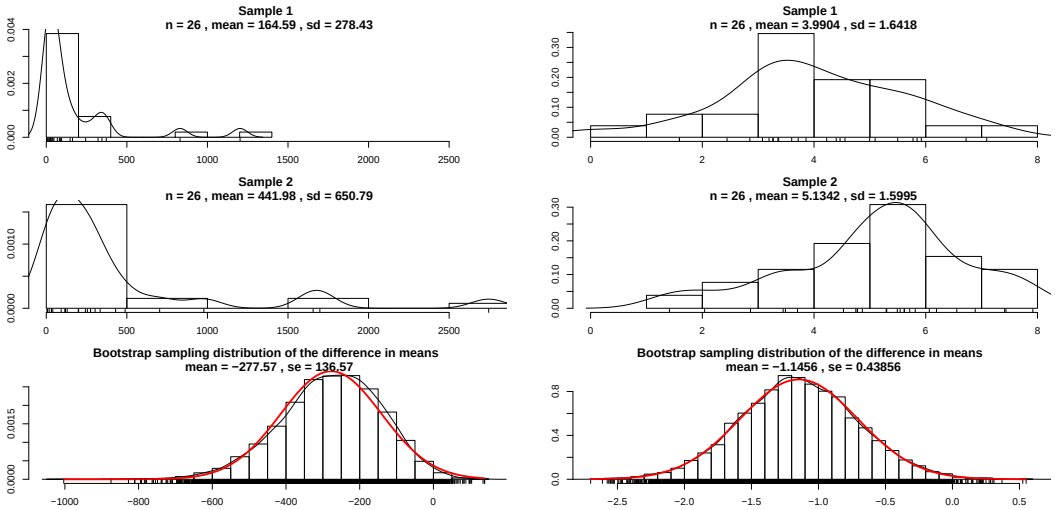
```
bs.one.samp.dist(d1$logunseeded)
bs.one.samp.dist(d1$logseeded)
```



(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.

Solution: First, I check the normality assumption of the sampling distribution of the difference in means using the bootstrap procedure for two samples. Below, the bootstrap sampling distribution of difference in means on the orginal scale has some skewness, but on the log scale it is very consistent with normality.

```
bs.two.samp.diff.dist(d1$unseeded, d1$seeded)
bs.two.samp.diff.dist(d1$logunseeded, d1$logseeded)
```



Thus, analysis is performed on the log scale because normality assumption is met there. Because the standard deviations are very similar (1.60 and 1.64), I used the pooled variance procedure.
Let μ_1 = the population mean volume for seeded, μ_2 = the mean for unseeded. We test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$, therefore, positive differences indicate larger volumes for seeded clouds.
The p-value= $0.014 < 0.05$, therefore we reject H_0 in favor of H_A, and the positive difference of 1.14 implies that seeded clouds have larger rain volumes.

```
# summary for separate vectors
summary(d1$logunseeded)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.21    3.79    3.99    5.07    7.09
summary(d1$logseeded)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.41    4.58    5.40    5.13    6.00    7.92
# comparing spreads, an assumption of equal variances seems reasonable
sd(d1$logunseeded)
## [1] 1.642
sd(d1$logseeded)
## [1] 1.6
IQR(d1$logunseeded)
## [1] 1.858
IQR(d1$logseeded)
## [1] 1.419
## Equal variances
#   var.equal = FALSE is the default

# two-sample t-test specifying two separate vectors
t.summary.eqvar <- t.test(d1$logseeded, d1$logunseeded, var.equal = TRUE)
t.summary.eqvar
##
##  Two Sample t-test
##
## data:  d1$logseeded and d1$logunseeded
## t = 2.544, df = 50, p-value = 0.01408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2409 2.0467
## sample estimates:
## mean of x mean of y
##     5.134     3.990
```

(30pts) **2. Acid:** Use the Acid data (from HW 2).

Acid1	A1-2	A1-3	A1-4	Acid2
0.123	0.110	0.112	0.126	0.109
0.109	0.110	0.123	0.110	0.111
0.110	0.110	0.110	0.109	0.110
0.109	0.090	0.109	0.114	0.110
0.112	0.109	0.110	0.110	0.105
0.109	0.111	0.109	0.110	0.110
0.110	0.098	0.110	0.110	0.111
0.110	0.109	0.109	0.110	0.110
0.110	0.109	0.110	0.110	0.110
0.112	0.109	0.110	0.111	0.111
0.110	0.109	0.111	0.107	0.109
0.101	0.111	0.111	0.110	0.111
0.110	0.109	0.109	0.107	0.109
0.110	0.108	0.107	*	0.112
0.110	0.110	0.120	*	0.109
0.110	0.112	0.133	*	0.109
0.106	0.111	0.107	*	0.111
0.115	0.110	0.103	*	0.110
0.111	0.111	0.111	*	0.112
0.110	0.111	0.110	*	0.112
0.107	0.107	0.122	*	0.109
0.111	0.111	0.109	*	0.110
0.110	0.112	0.108	*	0.110
0.113	0.105	0.109	*	0.109
0.109	0.109	0.109	*	0.113
0.108	0.109	0.114	*	0.108
0.109	0.110	0.107	*	0.105
0.111	0.110	0.104	*	0.110
0.104	0.109	0.110	*	0.109
0.114	0.110	0.114	*	0.109
0.110	0.104	0.107	*	0.110
0.110	0.111	0.101	*	0.110
0.110	0.110	0.111	*	0.110
0.113	0.111	0.109	*	0.104
0.114	0.109	0.110	*	0.109
0.110	0.110	0.111	*	0.110
0.110	0.111	0.110	*	0.111

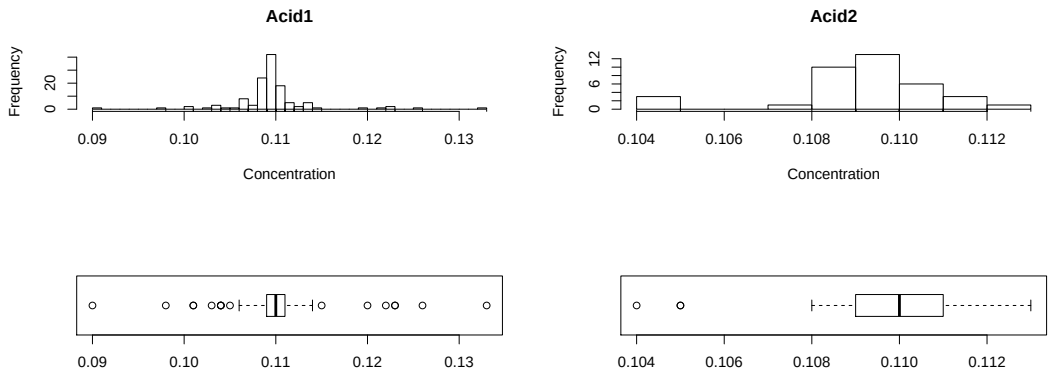
Read the data from the website with:

```
d2 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-3.csv")
```

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

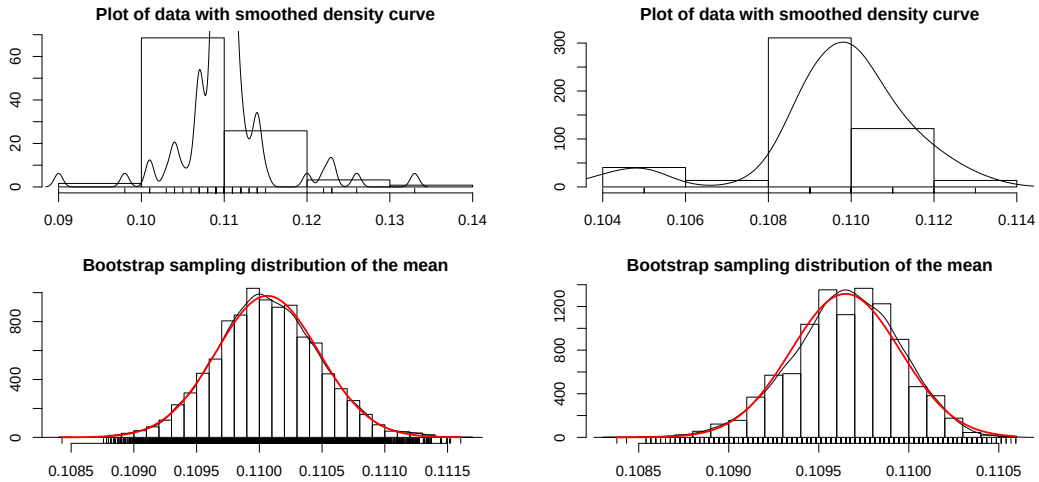
Solution: Acid1; the distribution is fairly symmetric, unimodal, but contains outliers because the distribution is more "peaky" than a normal distribution. An assumption of normality is not a reasonable operational assumption.
Acid2: the distribution is fairly symmetric, unimodal, but contains a couple substantial outliers. With the 3 outliers, normality may not be reasonable.

```
d2 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-3.csv")
Acid1 <- subset(d2, exper == "Acid1")[,1]
Acid2 <- subset(d2, exper == "Acid2")[,1]

par(mfrow=c(2,1))
hist(Acid1, breaks = 40, main = "Acid1", xlab = "Concentration")
rug(Acid1)
boxplot(Acid1, horizontal=TRUE)

par(mfrow=c(2,1))
hist(Acid2, breaks = 10, main = "Acid2", xlab = "Concentration")
rug(Acid2)
boxplot(Acid2, horizontal=TRUE)
```



However, because the distribution of the sample mean appears normal from the bootstrap analysis, there is little to be concerned with for a t-test analysis.

```
bs.one.samp.dist(Acid1)
bs.one.samp.dist(Acid2)
```



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

Solution: Let μ_1 = average acidity for the chemistry class's first experiment (acid1), and μ_2 = average acidity for the chemistry class's second experiment (acid2).
Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_A : \mu_1 - \mu_2 \neq 0$.
While normality assumptions are not met, the larger sample sizes and symmetry remove great cause for concern (as shown by the bootstrap). We use the Satterthwaite two-sample t-procedure since standard deviations are very different (0.00452 vs 0.00184) Consistent with HW 2, because the p-value= $0.415 > 0.05$, we fail to reject H_0, concluding that the means of the two experiments are not statistically different.

```
# summary for separate vectors
summary(Acid1)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.090   0.109   0.110   0.110   0.111   0.133
summary(Acid2)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.104   0.109   0.110   0.110   0.111   0.113
# comparing spreads, an assumption of equal variances seems reasonable
```

```
sd(Acid1)
## [1] 0.004544
sd(Acid2)
## [1] 0.001844
IQR(Acid1)
## [1] 0.002
IQR(Acid2)
## [1] 0.002
## Unequal variances is the default
# two-sample t-test specifying two separate vectors
t.summary.eqvar <- t.test(Acid1, Acid2)
t.summary.eqvar
##
##  Welch Two Sample t-test
##
## data:  Acid1 and Acid2
## t = 0.8181, df = 145.2, p-value = 0.4147
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0005889  0.0014206
## sample estimates:
## mean of x mean of y
##    0.1101    0.1096
```

(60pts) **3. cAMP:** Cyclic adenosine monophosphate (cAMP) is a substance that can mediate cellular response to hormones. In a study of maturation of egg cells in the frog Xenopts laevis, oocytes from each of four females were divided into two batches: one batch was exposed to progesterone and the other was not. After two minutes, each batch was assayed for its cAMP content, with the results given in the table below.

	cAMP (pmol/oocyte)	
Frog	Control	Progesterone
1	6.01	5.23
2	2.28	1.21
3	1.51	1.40
4	2.12	1.38

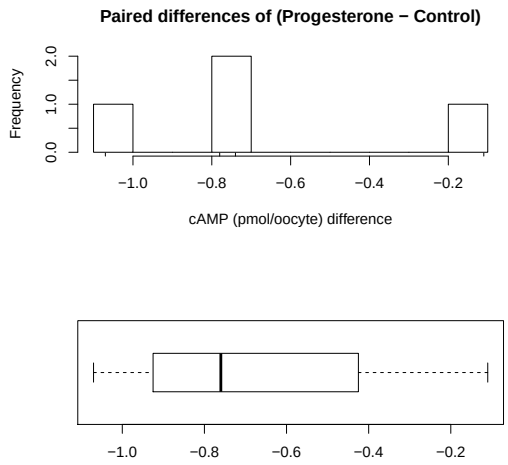
Read the data from the website with:

```
d3 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_03_F14-3.csv")
```

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

Solution: Differences are shown below with a plot. Keep in mind that the box plot shows a five-number summary and we only have 4 values! It's hard to have evidence against the normality assumption with only 4 values.

```
d3$diff <- d3$Progesterone - d3$Control;
d3$diff
## [1] -0.78 -1.07 -0.11 -0.74
par(mfrow=c(2,1))
hist(d3$diff, breaks = 10, main = "Paired differences of (Progesterone - Control)", xlab = "cAMP (pmol/oocyte) differen
rug(d3$diff)
boxplot(d3$diff, horizontal=TRUE)
```



(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.

Solution: Let μ_d = the mean cAMP content difference between the Progesterone minus the Control.
Test $H_0 : \mu_d = 0$ against $H_A : \mu_d \neq 0$.
The p-value= $0.044 < 0.10$, therefore we reject H_0 in favor of H_A, concluding that, because of the negative mean difference, that progesterone results in a lower cAMP content than the control treatment.

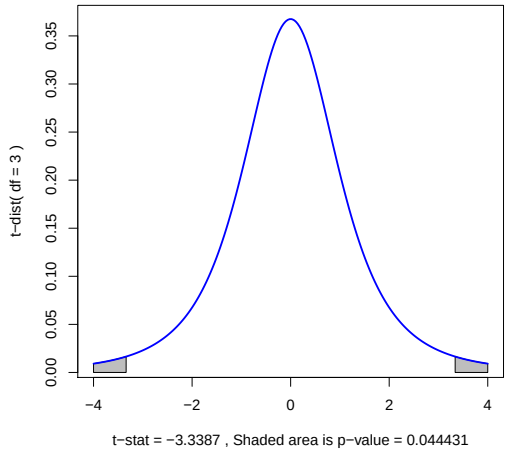
```
# paired t-test
t.summary <- t.test(d3$diff, conf.level = 0.90)
t.summary
```

```
##
##  One Sample t-test
##
## data:  d3$diff
## t = -3.339, df = 3, p-value = 0.04443
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  -1.1508 -0.1992
## sample estimates:
## mean of x
##    -0.675
```

```
t.summary <- t.test(d3$Progesterone, d3$Control, paired = TRUE, conf.level = 0.90)
t.summary
```

```
##
##  Paired t-test
##
## data:  d3$Progesterone and d3$Control
## t = -3.339, df = 3, p-value = 0.04443
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -1.1508 -0.1992
## sample estimates:
## mean of the differences
##                  -0.675
```

```
t.dist.pval(t.summary)
```



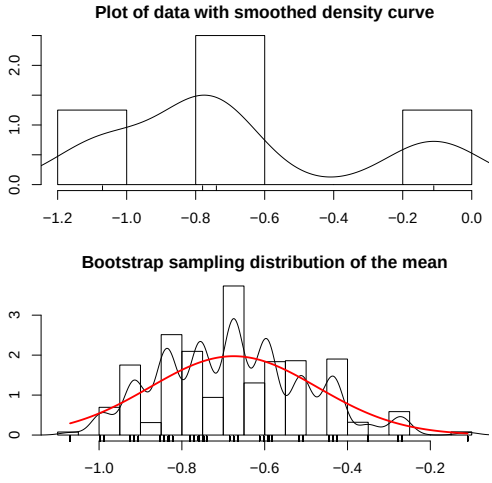
(c) (10 pts) Compute and interpret a 90% CI for the difference in population mean cAMP levels for batches of oocytes that are untreated versus those treated with progesterone.

Solution: We are 90% confident that the true mean difference between cAMP contents from progesterone and control treatments is in the interval $(-1.15, -0.20)$.

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

Solution: We have assumed that the distribution of differences is normal. Given that we only have 4 values, it is hard to make or refute any distributional claims. That being said, we do not have evidence to refute the normality assumption.
Furthermore, the bootstrap sampling distribution of the mean (which is only informed by 4 values) follows a normal curve pretty well (no strong skewness). So the normality assumption seems reasonable.

```
bs.one.samp.dist(d3$diff)
```



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

Solution: Given paired treatments of 4 eggs, we have substantial evidence to conclude that proges-

terone treated eggs have lower cAMP content than control eggs.