Advanced Data Analysis 1, Stat 427/527 Homework 2 Name: Answer Key
Fall 2014 Prof. Erik B. Erhardt

Part I. (120 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. Unseeded vs seeded precipitation:** The data were collected in southern Florida between 1968 and 1972 to test a hypothesis that massive injection of silver iodide into cumulus clouds can lead to increased rainfall. On each of 52 days that were deemed suitable for cloud seeding, a random mechanism was used to decide whether to seed the target cloud that day or to leave it unseeded as a control. An airplane flew through the cloud in both cases, since the experimenters and the pilot were themselves unaware of whether on any particular day the seeding mechanism in the plane was loaded or not (that is, they were blind to the treatment). Precipitation was measured as the total rain volume falling from the cloud base following the airplane seeding run, as measured by radar.

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

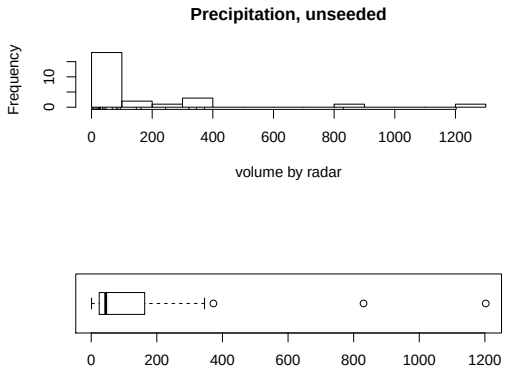
Read the data from the website with:

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

(a) (10 pts) Obtain histograms and boxplots of unseeded days. Describe the distributions. Why is this shape almost certain to occur here?

Solution:

```
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)
```



The distribution is unimodal, skewed right, with a few outliers. This shape occurs because rain volume is non-negative and a volume over an area is proportional to a squared value. Thus, if the amount of rain in a small area is normally distributed over the days, the volumes over a larger area will be distributed as the square of a normal distribution. Thus, most rain falls in smaller volumes, while extreme volumes are rare and large.

(b) (10 pts) Obtain a 95% confidence interval for the mean precipitation amount of the unseeded group. Do the assumptions for the method appear to be appropriate? Discuss.

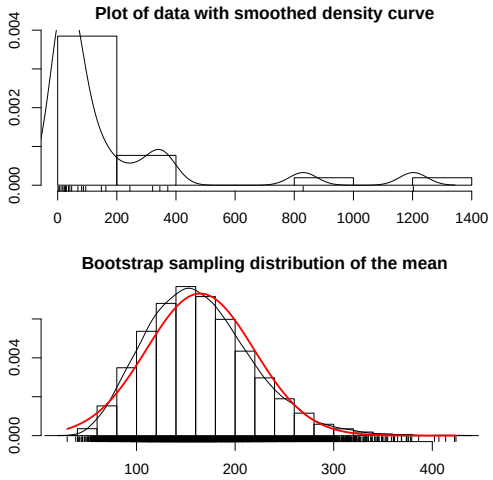
Solution: The distribution is skewed, so the normality assumption is not met. Since the assumptions are not met for the CI, I would not use this CI for inference for these population mean.

```
t.summary.unseeded <- t.test(d1$unseeded)
t.summary.unseeded$conf.int
```

```
## [1]  52.13 277.05
## attr(,"conf.level")
## [1] 0.95
```

Assessing assumptions using the bootstrap function from Chapter 2, there is enough deviation in the sampling distrubiton from normality to consider the CI inappropriate for this data.

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

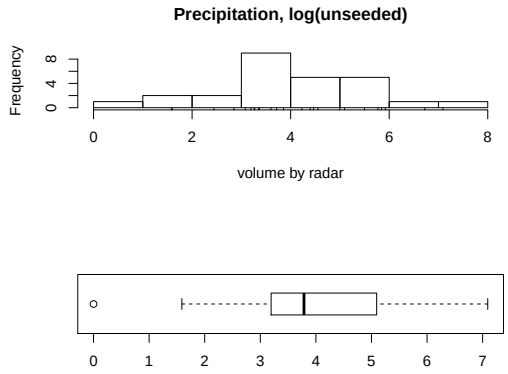


(c) (10 pts) Transform the data by taking the `log()` of each value of the unseeded data. Make a histogram and boxplot of the transformed data. Describe the distribution of the transformed data.

Solution:

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

```
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)
```



The log distribution is unimodal, symmetric, with one or no outliers. These data are consistent with being normally distributed.

(d) (10 pts) Obtain a 95% confidence interval for the mean `log()` precipitation amounts of the unseeded data. Do the assumptions for the method appear to be appropriate? Discuss.

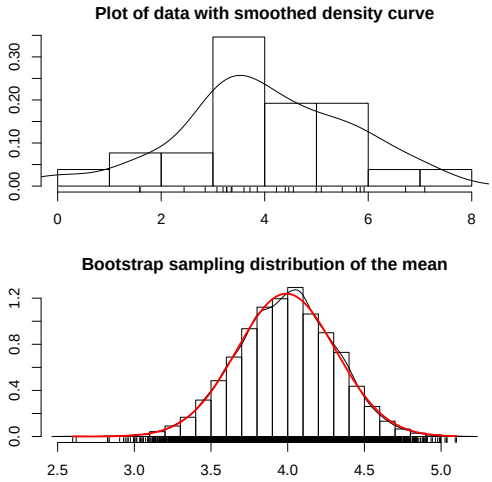
Solution: The normality assumption is reasonable, so the CI below is appropriate for the log volumes.

```
t.summary.logunseeded <- t.test(d1$logunseeded)
t.summary.logunseeded$conf.int
```

```
## [1] 3.327 4.654
## attr(,"conf.level")
## [1] 0.95
```

Assessing assumptions using the bootstrap function from Chapter 2, the sampling distrubiton is consistent with normality to consider the CI appropriate for this data.

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



(30pts) **2. TCL:** The following data are the total cholesterol levels (TCL) for a sample of 14 young adult males (aged 25 years or less) on the Kaiser Health plan in California:

TCL
227
239
221
213
218
246
218
224
210
204
197
229
220
197

(a) (25 pts) Suppose it is believed that the mean TCL of all adult males in the United States is 210. Is it plausible the (population) mean TCL of all young adult males on the Kaiser plan is the same as the U.S. male population mean TCL? Test at the 5% level. As with any hypothesis test, assure that your solution includes the following (and label the parts in this assignment): (A) define the population parameter in context, (B) clearly state the hypotheses in notation and in words, (C) state assumptions and how assumptions will be assessed, (D) evaluate assumptions based on graphical summaries, and (E) discuss the test, and the decision made in context.

Solution:

(A) Let μ = average TCL level for all adult males on the Kaiser health plan (in California).

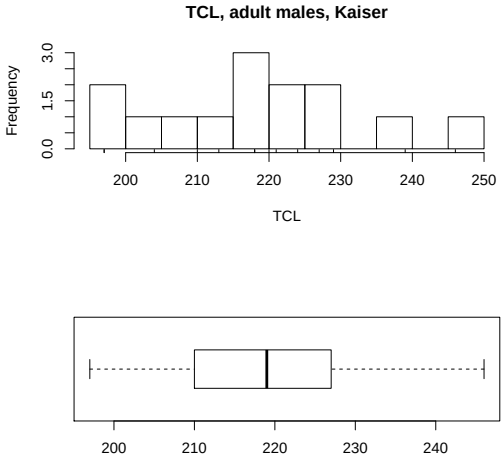
(B) We are interested in whether it is plausible that this population mean agrees with the mean TCL level of all adult males in the US, which is claimed to be 210. That is, we wish to test $H_0 : \mu = 210$ against $H_A : \mu \neq 210$.
This is a two-sided test because there is no specified direction of difference that we wish to detect.

(C) The one-sample t-test that we will perform assumes that the data are a random sample from a population with a normal frequency curve. Without information on the sampling design, we make the simple random sample assumption. Boxplot and histograms (see below) of the sample were generated to assess the normality assumption.

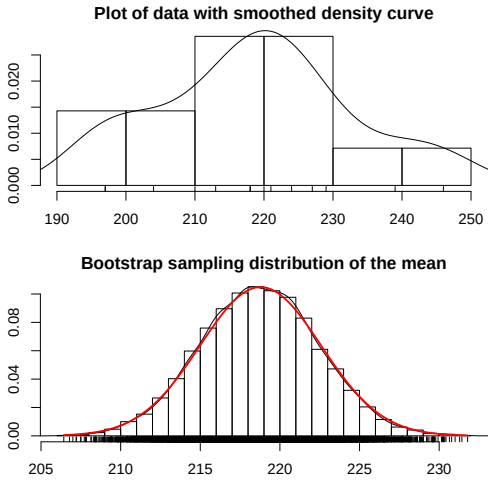
(D) Looking at the graphical summaries, we see the distribution is fairly symmetric, unimodal, and contains no outliers. An assumption of normality is a reasonable operational assumption — there is no strong evidence of non-normality.

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

par(mfrow=c(2,1))
hist(d2$TCL, breaks = 10, main = "TCL, adult males, Kaiser", xlab = "TCL")
rug(d2$TCL)
boxplot(d2$TCL, horizontal=TRUE)
```



```
bs.one.samp.dist(d2$TCL)
```



(E) R output for a two-sided test and the corresponding two-sided CI for μ are given below. The p-value for the test is p= 0.038, which is less than 0.05, the desired size of the test (i.e., a 5% test). Thus we have sufficient evidence to reject the null hypothesis in favor of the alternative that the population mean TCL for young adult males on the Kaiser plan differs from the US population mean of 210.

```
  # defaults include: alternative = "two.sided", conf.level = 0.95
t.summary <- t.test(d2$TCL, mu = 210)
t.summary
```

```
##
##  One Sample t-test
##
## data:  d2$TCL
## 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
```

(b) (5 pts) Also, construct and interpret a 95% CI for the Kaiser population mean. (F) Discuss the CI and interpret the result in context.
REMARK: Note that this problem involves two populations, one of which is a subset of the other. Furthermore, we are assuming that the mean for the larger population is known. This is a one-sample problem because only 1 sample was taken.

Solution: The 95% CI leads to the same conclusion since 210 is not contained within the interval. In particular, we are 95% confident that the population mean TCL for young adult males on the Kaiser plan is between 210.57 and 227.00.

```
# from above
t.summary$conf.int
```

```
## [1] 210.6 227.0
## attr(,"conf.level")
## [1] 0.95
```

(50 pts) **3. Acid:** Rows labelled "Acid1" are the results of a titration to determine the acidity of a solution in a chemistry class. Rows labelled "Acid2" are the results from a second experiment.

Acid1				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

The instructor knew in both cases that the correct value for this solution was 0.110. Use a test of hypothesis and corresponding CIs to see if the class is "biased" — that is, to see if the class is systematically too high or too low. Be sure to state and check all assumptions. (Note that the goal is not to compare the results of the two experiments to each other.)

(a) (25 pts) Do this for experiment 1 (Acid1).

Solution:

(A) Let μ = average acidity for the chemistry class.

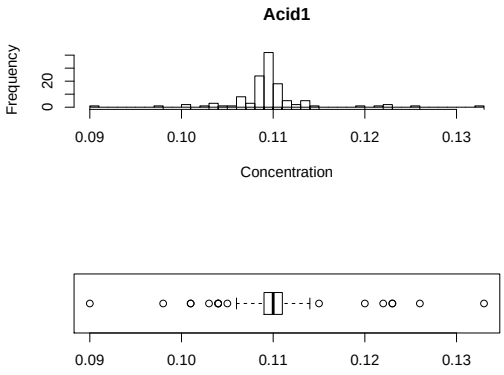
(B) We are interested in whether it is plausible that this population mean agrees with the true mean acidity, which is claimed to be 0.110. That is, we wish to test
$H_0 : \mu = 0.110$ against $H_A : \mu \neq 0.110$.
This is a two-sided test because there is no specified direction of difference that we wish to detect.

(C) The one-sample t-test that we will perform assumes that the data are a random sample from a population with a normal frequency curve. Without information on the sampling design, we make the simple random sample assumption. Boxplot and histogram (see below) of the sample were generated to assess the normality assumption.

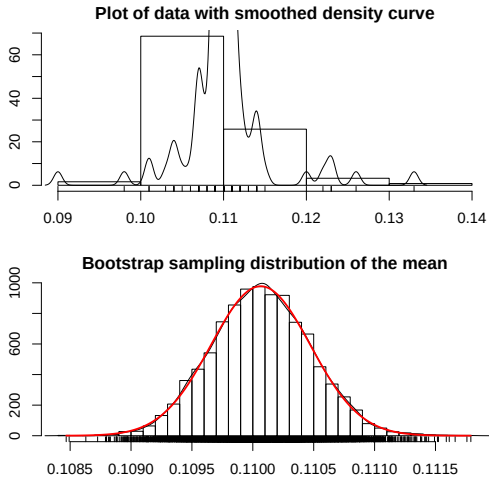
(D) Looking at the graphical summaries, we see the distribution is fairly symmetric, unimodal, but contains outliers because the distribution is more "peaky" than a normal distribution. Though, the sample size is large enough for the sampling distribution to be close to normal.

```
d3 <- read.csv("http://statacumen.com/teach/ADA1/ADA1_HW_02_F14-3.csv")
Acid1 <- subset(d3, exper == "Acid1")[,1]
Acid2 <- subset(d3, exper == "Acid2")[,1]

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




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



(E) R output for a two-sided test and the corresponding two-sided CI for μ are given below. The p-value for the test is p= 0.875, which much larger than 0.05, the desired size of the test (i.e., a 5% test). Thus, the experiment is on target with the true acidity value.

```
  # defaults include: alternative = "two.sided", conf.level = 0.95
t.summary <- t.test(Acid1, mu = 0.110)
t.summary
```

```
##
##  One Sample t-test
##
## data:  Acid1
## 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
```

(F) The CI above includes 0.110.

(b) (25 pts) Do this for experiment 2 (Acid2).

Solution: Similar to part (a), abbreviated solution:

(A) Let μ = average acidity for the chemistry class.

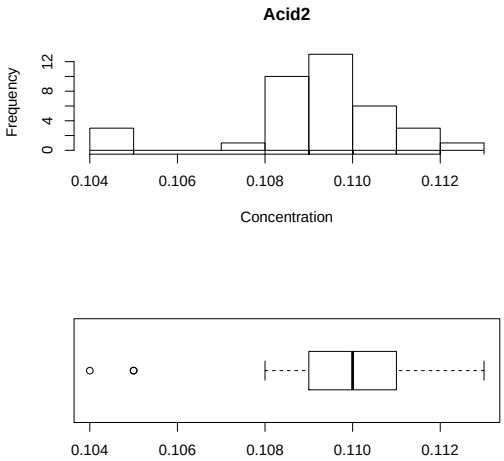
(B) $H_0 : \mu = 0.110$ against $H_A : \mu \neq 0.110$.

Two-sided test.

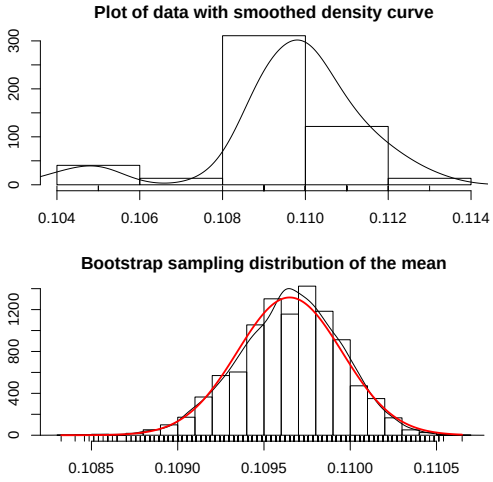
(C) The one-sample t-test, same assumptions.

(D) Looking at the graphical summaries, we see the distribution is fairly symmetric, unimodal, but contains a substantial outlier. This outlier should be investigated for whether extenuating circumstances explain why this value is so extreme. If it can be explained, it can be removed and the analysis can go on without it.

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




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



(E) With or without the 3 outliers, p> 0.05, though another statistical test (that we'll learn later) should be used with this outlier.

```
  # defaults include: alternative = "two.sided", conf.level = 0.95
t.summary <- t.test(Acid2, mu = 0.110)
t.summary

##
##  One Sample t-test
##
## data:  Acid2
```

```
## 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
```

```
# remove 3 outliers to see the affect on the result
t.summary <- t.test(sort(Acid2)[-c(1:3)], mu = 0.110)
t.summary
```

```
##
##  One Sample t-test
##
## data:  sort(Acid2)[-c(1:3)]
## t = 0.4631, df = 33, p-value = 0.6463
## alternative hypothesis: true mean is not equal to 0.11
## 95 percent confidence interval:
##  0.1097 0.1105
## sample estimates:
## mean of x
##    0.1101
```

(F) The CI above includes 0.110.