Advanced Data Analysis 1, Stat 427/527 | Homework 2 | Name: ____________________
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?

(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.

(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.

(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.

(30^{pts}) **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.

(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.

(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).

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