STATISTICS 500 – Fall 2009 – HW 3

DUE DATE:

on campus Friday, 18 Sept 2009, in lecture (11 am)
or by e-mailto Chuanlong, dclong@iastate.edu, no later than noon.

off campus Monday, 21 Sept 2009, by 4 pm to Nicole Rembert, email: rembeall@iastate.edu or FAX: 515-294-4040 (please include cover page with Stat 500 / Nicole Rembert).

1. Cloud seeding – Clouding seeding is the practice of using an airplane to spray certain chemicals into a cloud. The hope is to increase the amount of rain that falls from the cloud. It is difficult to evaluate the effectiveness of cloud seeding. The following data are from one of the best studies.

   This study was conducted in southern Florida between 1968 and 1972. The weather was watched daily. If the cloud conditions were considered suitable, then that day was included in the experiment. A total of 52 days were considered suitable. Each suitable day, one target cloud was arbitrarily chosen and randomly assigned to one of two treatments:

   seed: fly an airplane through the cloud and spray the seed chemical

   control: fly an airplane through the cloud but the sprayer was not loaded with the seed chemical.

   The total rain that fell from the target cloud was measured. This experiment was double blind (neither the pilots nor rainfall observers knew which treatment was used). The file rainfall.txt on the class web site contains two variables: treatment and rain for each of the 52 clouds in the study.

   (a) What are the experimental and observational units for this study and data set? Do you expect a problem with the assumption of independence?

   (b) Consider the untransformed rainfall values. Is a normal distribution reasonable? Explain why or why not.

   (c) For the untransformed rainfall values, use a graphical method to assess the assumption of equal variances. Is the assumption reasonable? Explain why or why not.

   (d) Use Levene's test to assess the assumption of equal variances. What is an appropriate conclusion?

   (e) Calculate the mean and standard deviation for each treatment (remember, using the original untransformed rainfall values). Use the log(mean) and log(s.d.) to calculate the parameter that describes the variance-mean relationship (β in my lecture). What transformation does this suggest?

   (f) Now let's check the properties of the log-transformed values. Is a normal distribution reasonable for the log transformed values?

   (g) It is reasonable to assume that the log-transformed values have equal variances? You can use either a graphical approach or a formal test.

   (h) Use the Wilcoxon rank-sum test to test H_o: no effect of seeding on untransformed rainfall. What is the p-value? Repeat using log transformed rainfall. Are the two p-values the same? Explain why you should (or should not) expect the same p-value.

2. Outliers – The following data are metabolic expenditures (amount of energy expended by patients) for 8 patients admitted to a hospital for reasons other than trauma and for 8 patients admitted for trauma (multiple fractures).

Nontrauma:	18.7	46.3	21.7	17.8	21.5	19.4	19.5	21.3
Trauma:	25.1	28.4	23.2	26.4	25.7	20.3	21.4	24.7

(a) Examine the distributions of these scores. Is the assumption of normality reasonable? Since the sample sizes are small, I suggest you pool the residuals from the two groups before assessing normality.

(b) Carry out a two-sample t-test comparing the means of the populations.

(c) Other data about the patient with the value of 46.3 were examined. That patient had a thyroid disease that affects metabolic rate. She was the only patient in this study with that disease. Delete this value from the data set and repeat the analysis. Are the conclusions similar to those in part (2b)?

(d) Explain why it is useful to know that the patient with the value of 46.3 was the only patient with thyroid disease.

3. Outliers 2 – The data in illust.txt are made up to make a point about why I check for outliers. There are two groups of observations, each with 10 values.

   (a) Calculate the means for the two groups and use usual t-based methods to calculate the 95% confidence interval for the difference of means.

   (b) Calculate and plot residuals against predicted values. Do you see any concerns? If so, list your concerns.

   (c) It is reasonable that the decimal point was misplaced in observation 10 of group 1. It is probably 7.99. Recompute the means and the 95% ci for the difference. Does fixing a typo change the conclusions about the difference in groups?

4. Mercury in catfish – Over the last couple of weeks, we have seen a variety of methods for answering questions about data from 2 groups. Lecture and the previous homework questions should have given you a sense of some of the differences between methods. My thoughts are explained at the beginning of the 9/14 lecture, which I recommend you see before starting this problem. This problem asks you to use what you know to choose a method to analyze a particular data set.

   The data in fishHg.txt are from a survey of mercury concentrations in catfish, a common sportfish. Catfish are often eaten by people. High mercury concentrations are health hazard. These data are from a survey of agricultural and urban watersheds in the continental US. Watersheds were randomly chosen (a long and complicated procedure, but very well done in this study), catfish were collected from each watershed, and the mercury concentration in the flesh was measured. There is one value for each watershed. There are a total of 42 agricultural watersheds and 30 urban watersheds in the study. The numbers are unequal for good reasons that are irrelevant to this question. The data are in fishHg.txt on the class web site.

   The investigators want to know:

   (a) Is there a difference in concentration between the ag and urban watersheds? (i.e. a hypothesis test)

   (b) How big is the difference? (i.e. a point estimate)

   (c) How precise is the estimated difference?
   (i.e. a confidence interval for the effect).

   They have brought you the data and asked for help answering these questions.

(a) What method do you believe is the most appropriate to analyze these data? Explain your choice. Your answer should be a short paragraph suitable for a materials and methods section of a scientific paper.

(b) Analyze the data and answer the investigator's questions using the most appropriate method. If possible, provide a hypothesis test, a point estimate, and a confidence interval. If you transform the data, you may report estimates and intervals for the transformed quantities. If you can not construct a confidence interval that you "believe", i.e. that you trust as a reasonable statement about the uncertainty in the estimate, then explain why you aren't providing one. Your answer should a paragraph suitable for the results section of a scientific paper.

    You are allowed to present results from only one analysis. You must decide which analysis is most appropriate. You are not allowed to present results from more than one method because that will confuse the public.

These are real data, so the right answer is unknown. Some analyses are more appropriate than others, but there may be more than one correct answers. These are moderately 'nasty' data, in the sense that most (all?) analyses have some sort of problem. Some analyses have many problems. I haven't found the perfect analysis for these data. I don't expect you to.

Your mark for this question will be based on your explanation of the choice of analysis and whether you provide appropriate answers to the investigator's questions. Points will be deducted if you present results from multiple methods. Grading will be lenient. Thinking about the issues is more important than the grade.