

**Stat 401, Section F Homework 4**  
**Due Date:** Wednesday, September 19

1. A researcher wanted to compare the abundance of crabs in two different coastal areas. Each coastal area was divided into several hundred strips. Each strip had approximate dimensions of 10 feet by 300 feet and was oriented roughly perpendicular to the shoreline. A random sample of 20 strips was selected from each of the coastal areas. With the help of a Global Position System (GPS) device, the researcher walked each of the randomly selected strips and recorded the number of crabs observed in each strip. The natural log of the count in each strip was computed. Summary statistics for the coastal areas are provided below.

Coastal Area	Number of Strips	Mean of Log Counts	Standard Deviation of Log Counts
A	20	1.88	1.1
B	20	2.75	1.2

- (a) Is the investigation described in this problem an observational study or an experiment?
- (b) Estimate the difference between the mean log crab count per strip in coastal area B and the mean log crab count per strip in coastal area A.
- (c) Find a 95% confidence interval for the difference in mean log crab counts between the two coastal areas.
- (d) Provide an interpretation of your estimate in (a) and confidence interval in (b) as discussed in Section 3.5.2.
2. The program *california.sas* on the course web site contains a subset of data from an experiment conducted with a group of ISU students. Approximately half of the students (selected at random) in the group were asked if the population of California was greater or less than 7 million; call this collection of students the “7” million treatment group. The remaining students were asked if the population of California was greater or less than 70 million; call this collection of students the “70” million treatment group. Then all students (in both “7” and “70” million treatment groups) were asked to guess the population of California in millions. The purpose of the experiment was to determine if the lead question could influence students’ guesses about the population of California. Such an influence would support a phenomenon in psychology known as *anchoring*. Anchoring is often used by advertisers or salespeople to try to get customers to pay a high price for an item while believing that it is a bargain compared to some other option.
- (a) The first portion of the code (before *data two; set one;*) provides an analysis of the raw data. Run the program and report a test statistic, *p*-value, and a conclusion for testing whether the first question had an effect on the students’ guesses.
- (b) Provide a 95% confidence interval for the difference between the mean guess of the “7 million group” and the mean guess of the “70 million group.”
- (c) The last portion of the code (beginning with *data two; set one;*) analyzes the data on the log scale. Report a relevant test statistic, *p*-value, and conclusion.
- (d) Based on the analysis of the log-transformed data, provide statements that summarize the effect of the first question on the students’ guesses. Include a 95% confidence interval (make sure to use a back-transformation by *e*). Your answer should be very similar to the last paragraph on page 57 regarding the summary of statistical findings in the cloud seeding experiment.

- (e) List a drawback to analyzing the data on the original scale for this dataset.
  - (f) List a drawback to analyzing the data on the log scale for this dataset.
3. Reread Sections 3.6.2 and 3.6.3 on robustness and transformations for paired  $t$ -tools. Read problem 30 on page 109 of Chapter 4. Analyze the data to estimate the sunlight protection factor for the sunscreen used in this experiment with the SAS program *sunscreen.sas*. (The sunlight protection factor can be considered to be the median value of the ratio

$$\frac{\text{tolerance to sunlight after sunblock}}{\text{tolerance to sunlight before sunblock}}$$

for the population of all people who would use the sunblock.) Provide a 95% confidence interval along with your estimate.

4. Researchers examined the amount of a particular soybean protein produced in the root tissue of each of 5 soybean plants. Prior to measuring the amount of the protein, 3 plants - randomly selected from the 5 - were infected with soybean cyst nematodes. A gel-like substance containing the nematodes was spread over the roots of these three plants to induce the infection. The roots of the other 2 plants were treated with a gel-like substance that contained no nematodes. After a period of time, the protein amount of each plant was measured, resulting the following data:

DATA	Protein Amounts		
Infected plants	10	7	4
Uninfected plants	5	3	
Sample average protein of Infected	Sample average protein of Uninfected		Difference
7	4		3

- (a) Compute the one-sided p-value from a randomization test for assessing whether infected plants have a larger mean protein amount than uninfected plants.
- (b) Compute the two-sided p-value from a randomization test assessing whether mean protein amounts differ between infected and uninfected plants.

(To compute the p-values, you need to consider all the possible ways that the 5 plant protein values could be divided into 2 groups: one group with 3 values (“infected”) and one group with 2 values (“uninfected”). It helps to consider all possible groups consisting of 2 values from the 5 protein amount values (the “uninfected” group), placing the remaining 3 values into the other (“infected”) group. Once all the groupings are determined, you can examine each grouping and determine the difference in two sample averages between the “infected” and “uninfected” values. Then compute the p-value as an appropriate proportion. Unlike a t-test, you may find here that the two-sided p-value from a randomization test is not simply twice the one-sided p-value; this can happen in a randomization test when the group sizes differ. For a t-test on the other hand, the two-sided p-value is almost always twice the one-sided p-value.)

5. Consider the data in problem 23 from Chapter 3 of your text on the relationship between skin cancer rates and sunspot activity. Use **R** to answer the following questions:
- (a) Conduct a two-sample  $t$ -test to determine if there is a significant difference between the mean skin cancer rate in years following high sunspot activity and the mean skin cancer rate in years following low sunspot activity. Give a the test statistic,  $p$ -value, and conclusion.

- (b) Examine a scatterplot of skin cancer rate vs. year for the “high” years. Do the same for the “low” years. There appears to be a relationship between skin cancer rate and year. Describe the relationship in a sentence.
- (c) Is there any reason to suspect that using the two sample  $t$ -test to compare skin cancer rates in the two groups is inappropriate? (Are any problems indicated indicated by the plots?)