Stat 401 A XM Homework 4
On-Campus Due Date: Wednesday, September 24
Off-Campus Due Date: Wednesday, October 1

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 (b) and confidence interval in (c) 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 earlier this semester in lab. Approximately half of the students (selected at random) were asked if the population of California was greater or less than 7 million. The other students were asked if the population of California was greater or less than 70 million. Then all students 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. 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 drawbacks to analyzing the data on the original scale for this dataset.

   (f) List drawbacks 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. Provide a 95% confidence interval along with your estimate. You may wish to use *sunscreen.sas* to help you formulate your answer. The program *sunscreen.sas* performs two different analyses of the data. Use the best one to help you obtain the estimate and confidence interval.

4. Consider the data in problem 23 from Chapter 3 of your text on the relationship between skin cancer rates and sunspot activity. The program *sunspot.sas* reads the data and produces output that will help you answer the following questions. I encourage you to study the code and output carefully to try to understand what each step of the SAS code is doing.

   (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) Based on what you discovered in part (b), do you think the p-value computed in part (a) is smaller than it should be, about right, or bigger than it should be? Explain. (Hint: Our test depends on a pooled estimate of variability among skin cancer rates within the high and low-sunspot-activity groups. How will the relationship that you described in part (b) affect our estimates of within-group variability among skin cancer rates? If sunspot activity had been constant over time, would we have ended up with smaller or larger estimates of within-group variability?)

   (d) Examine the code used to produce the last plot in the program *sunspot.sas*. Explain what the *plot rate*year=ssact;* command does.

   (e) Examine the last plot produced by the program *sunspot.sas*. Describe any features of this plot that could suggest that sunspot activity is associated with skin cancer rate.