Statistics 411/511
Homework 3
Due Tuesday, October 19 by midnight

- **Instructions:** Please see the end of the syllabus for guidelines. Upload your homework to Gradescope via Canvas (access specific homework assignments from the "Assignments" link at the left of the Canvas course page). Your file must be a pdf document. **There will be a one-point deduction if you don't assign pages** (see this Gradescope help video).
- Do the computational part of the homework shortly after completing the week's lab activity.
- The problems are assigned from the **third edition** of the textbook. If you have another edition, consult the copy on one-hour reserve at the library website for the homework problems.
- **Academic Integrity** You are encouraged to *discuss* the homework with other students, but what you turn in must be your own work in your own words. **DO NOT** copy someone else's homework. You may share ideas and R code, but do not share R output or written language. The syllabus contains details and links to OSU's Student Conduct Code and procedure for reporting suspected academic misconduct.

1. This exercise is intended to give you practice log-transforming data and reporting result of an analysis on log-transformed data. You will work with the skin cancer data of exercise 23 on page 80. The data frame is `ex0323`.

   (a) Examine the structure of the data frame using `head()`. Turn in your R code and output.

   (b) Obtain "summaries" of the skin cancer rate for each of the two sunspot activity groups using `summary()`. See item 11 of Lab 1 for example code. Turn in your R code and output.

   (c) Use the example code on page 5 of Outline 3 to produce histograms for the sunspot activity groups. Turn in your R code and graph. [The argument `xlim=` is the minimum and maximum values for the horizontal axis. If you omit this argument, R will try to pick sensible values.]

   (d) Log-transform the cancer rates. Obtain summaries as in (b) and histograms as in (c) using the logged data. Turn in your R code, summary output, and graph. [Remember: "log" means natural log in R and in Statistics.]

   (e) The appearance of histograms depends greatly on the number and width of the bins. Boxplots are more standard. The bottom and top of the box represent the first and third quartiles, respectively. The line in the middle of the box represents the median. Produce side-by-side boxplots of the logged cancer rate data for the two sunspot groups. Compare the boxes to the summaries from part (d), and note whether the box geometry agrees with the summaries. [If you are interested in the details of boxplots, type `help(geom_boxplot)` in the Console window.]

   (f) State the three assumptions needed to use the t-tools. For each assumption, state your opinion whether it is reasonable for the untransformed data, then for the transformed

(Problem 1 continued on next page)

data. Give a reason for your opinion. You may be very brief here. One or two sentences per assumption will be enough.

(g) The t-tools are robust to many departures from the assumptions, so even if you don't believe the assumptions are met, perform a two-sample t-test using R on the logged data to answer the research question, "is the skin cancer rate higher when there is more sunspot activity?" Submit your R code but not output.

(h) Give a "statistical conclusion" reporting the results of your hypothesis test in part (g). A statistical conclusion should be on the original scale, not the log scale.

(i) Obtain a two-sided 95% confidence interval for the difference in population means, using the logged data, then back-transform the endpoints of the interval. Submit your R code and the resulting back-transformed interval.

(j) Give a "statistical conclusion" reporting your back-transformed confidence interval from (i)