Homework 5

Statistics 411/511: Spring 2018

Due: In class February 12

Turn in your solutions in a typeset report. You do need to report your answers to #2 on this homework in the format according to the Syllabus and *Writing a Statistical Report*. You must work with at least one other classmate on this homework and turn in a single solution. The maximum number of students who can work together is FOUR. List all students on the front page of your solution set.

1. Do problem #22 on p. 80 (this is not a data set in the `Sleuth3` package, you'll need to enter it "by hand"). In addition to problems (a)-(d) specified by the text, do parts (e)-(g). Include your R-code and R-output in an Appendix including the Box Cox plot.

(a)-(c) You may use R code for the "by hand" calculations.

(d) You may not use `t.test()` for the "by hand" calculation. Instead, use R to code the formula for the CI. You must interpret the final (back-transformed) CI in terms of the problem and in terms of the parameters being estimated.

(e) Use a scatterplot to graphically display these data.

(f) When attempting to make statements about the effect of voltage on breakdown times, is it important that the data are approximately normal? Explain.

(g) Find the Box Cox transform for the these two samples of breakdown times. Which transform does Box Cox suggest? Does the Box-Cox transform agree with your book's decision to use a log-transform? Explain.

2. Do #25 on page 147 of your text (data set `ex0525` in the `Sleuth3` package) where a random sample of $n = 2584$ Americans with paying jobs were asked about their income level and educational level in 2006. Educational level was simplified into 5 distinct categories. You must report your answers to this problem in the format according to the Syllabus and *Writing a Statistical Report* available on the course website. The report, not including the Appendix that contains your R-code and R-output and figures and any tables, should not exceed two pages. Your grade will be determined by how well you answer the questions and by the organization and clarity of your write-up. Address each of the following issues in your write up.

(a) Indicate the sampling plan and the study design in the *Introduction*. In the *Scope of inference* indicate how the sampling plan and the study design affects the applicability (or lack of applicability) of your conclusion.

(b) Plot side-by-side boxplots of the income of the educational groups in R. Also plot side-by-side boxplots of the $\log_{10}$-transformed income of the educational groups in R. The education categories are initially displayed alpha-numerically, so re-order the categories before you plot.

```
d$Educ = factor(as.character(d$Educ),levels = c("<12","12","13-15","16",">16"))
boxplot(Income2005 ~ Educ,data=d)
boxplot(log10(Income2005) ~ Educ,data=d)
```

Which of these two plots do you prefer for presenting these data to a reader? Why?

(c) Update the R-code in #2b to include the `boxplot()` options `main="Informative Title"`, `xlab="Informative x-axis label"` `ylab="Informative y-axis label"` to label plot appropriately.

(d) Fit the ANOVA model to the (untransformed) incomes and check the assumptions using residual plots. Indicate which, if any assumptions are violated, and why you think an assumptions may be violated. If necessary, transform the data, re-fit an ANOVA to the transformed data, and re-check asssumptions.

(e) Test the hypothesis that the mean or median income of an employee is associated with the employee's level of education. Report the F statistic and the p-value in the *Summary of Statistical Findings* section of your report. Put the R-code and R-output in the Appendix.