Homework 5  
Statistics 411: Fall 2017  
Due: In class Monday, October 9

As in HW4, the first TWO problems of this HW consider Exercise 25 on page 147 of your text 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 do not need to report your answers to this problem in the format according to the Syllabus and Writing a Statistical Report available on the course website. In an Appendix, include your R-code and R-output and figures and any tables.

As in HW4, load in the data for this problem and re-order the levels of the Education factor. In R:

```r
library(Sleuth3)
d = ex0525
d$Educ = factor(d$Educ, levels = c("<12","12","13-15","16",">16"))
```

1. HW4 explicitly instructed you to NOT transform the data prior to ANOVA. You could be critiqued for this approach because, according to one of our rules of thumb, one of the ANOVA assumptions was violated.

   (a) State the ANOVA assumption that was violated.

   (b) \( \log_{10} \) -transform the incomes in preparation of applying an ANOVA. After a \( \log \)-transform, are the assumptions of the ANOVA satisfied? Explain and use histogram, boxplot, normal probability, and residual vs. fits plots to justify your answer.

   (c) Apply an ANOVA to the \( \log_{10} \)-transformed data to test the hypothesis that the MEDIAN income of an employee is associated with the employee's level of education. Be sure to check the assumptions that the ANOVA is appropriate. Put the R-code and R-output in the Appendix.

   (d) State a proper conclusion and a scope of inference (which, unlike HW4, will refer to the median income).

   (e) Apply a follow-up two-sample t-test of the planned test that the median income level for workers with a high school education is less than the median income of workers with a bachelors degree. By “follow-up”, I mean to use the output of the ANOVA to perform this t-test (see Chapter 5 notes for R-code). Be sure to report the p-value and a 95% CI for the difference in means. Put the R-code and R-output in the Appendix.

2. Apply a Kruskal-Wallis non-parametric ANOVA to test whether the median income is different for the educational levels. First, apply Kruskal-Wallis to the incomes. Apply it a second time to the \( \log_{10} \) transformed incomes. You will see that you get exactly the same test statistic value and exactly the same p-value. Why is this to be expected?

3. Three experiments were at performed at MSU’s Center for Biofilm Engineering in 2017, in January, February and October. In each experiment, a biofilm that represents those found in chronic wounds was grown on three different surfaces for a total of \( n = 9 \) biofilms. The density of bacteria in the biofilm are recorded as number of bacteria per cm\(^2\). It is important to grow up a similar biofilm in multiple experiments so that any results regarding topical antibiotics or antimicrobials are presented with the same bio-challenge. Download and graph the following data from the STAT411 website using the following R-code.

```r
b = read.csv("http://www.math.montana.edu/parker/courses/STAT411/HW6_biofilmd ata.csv")
boxplot(log10(Number)~experiment,data=b)

# This next plot is an 'individual value plot'
dev.new()
require(ggplot2)
qplot(b$experiment,log10(b$Number))
```
(a) For a data set like this, many researchers prefer displaying the data using an individual value plot as opposed to a boxplot. Why?

(b) A microbiologist wants to present a 95% CI of the true mean log\textsubscript{10}-transformed densities and uses a 1-sample \textit{t}-CI with 8 df. Explain why a 1-sample \textit{t}-CI is inappropriate for these data.

(c) Fit a random effects ANOVA to these data with a random effect for experiment. Check the model assumptions using the appropriate diagnostics. Put the R-code and R-output in the Appendix.

(d) What is the correlation of the log\textsubscript{10}-transformed densities that were produced from the same experiment.

(e) What is the proportion of variance due to experiment-to-experiment sources?

(f) Build a 95% CI for the true mean log\textsubscript{10}-transformed densities of the chronic wound biofilms and interpret in terms of the problem.

(g) Report a 95% CI for the true median density of the chronic wound biofilms.