Homework 6

Statistics 411: Spring 2018

Due: In class Wedneday, February 21

As in HW5, 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. Work together in a team of 2-4 students and submit a single report.

1. Using your ANOVA results from HW#5, 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. Report the p-value for this test. Also report a 95% CI for the ratio in medians. Put the R-code and R-output in the Appendix.

2. Apply a permutation test of the hypothesis that the median income of a worker is associated with the worker's level of education.

   (a) Check that the assumptions of the permutation are satisfied. One assumption that we have not discussed is that a permutation test of a difference of means or medians assumes that the groups have the same shape (not necessarily normal) and hence also the same variance! So, if you transformed these data in HW5 because of non-normality before application of ANOVA, then you do not need to apply that same transform to these data prior to performing a permutation test. However, if you transformed these data in HW5 because of non-constant variance before application of ANOVA, then you do need to apply that same transform to these data prior to performing a permutation test. Indicate whether you transformed the data prior to the permutation test and why.

   (b) As we discussed in class, there are few different test statistics that you can use (SSG, MSG, variance of the means, F). Use the variance of the means as your test statistic. Include a histogram of this test statistic over many permutations of the data and report the test statistic value calculated from the data.

   (c) Report the p-value.

   (d) Compare the results from the permutation test to the ANOVA results from HW5. Are they similar? Did you expect the results to be similar? Why or why not?

3. 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?

4. 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.

```
b = read.csv("http://www.math.montana.edu/parker/courses/STAT411/HW6_biofilmdata.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 $\log_{10}$-transformed densities and uses a 1-sample t-CI with 8 df. Explain why a 1-sample 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 $\log_{10}$-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 $\log_{10}$-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.