Stat 333: Applied Regression Analysis
Spring 2015
Due Date: Friday, January 30 in class

Relevant text chapters: Appendix A

Instructions: You may (and are encouraged to) discuss homework problems with other students, but the solutions that you should provide should be your own and not directly copied from another student. Show your work wherever possible, and write out the formulas you used to arrive at your solutions. If you have any questions or difficulties, you are more than welcome to come see me in my office.

1. Suppose that you have a random sample of size n, $Y_1, Y_2, \ldots, Y_n$, that follows some probability distribution with mean μ and variance σ^2. Given the sample mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Show that $E(\bar{Y}) = \mu$ and that $Var(\bar{Y}) = \frac{\sigma^2}{n}$.

2. In problem 1 we have the conclusion that $E(\bar{Y}) = \mu$, which could be carelessly stated in words that "The mean of the (sample) mean equals the (population) mean". Using your own words, explain as best as you can what are the differences between these three means. You may find it helpful to use a simple numerical example to demonstrate the differences, but this is not required.

3. A group of friends who love eating fresh apples decided to conduct a small study to see if there was a difference in the mean weight of apples from two local apple orchards. The group of friends randomly selected 11 apples from "Fisher's Farm & Orchard" and independently selected at random 10 apples from "Pearson's Produce" and measured the apples' weights in ounces. The data obtained are shown below:

Fisher: 4.7, 5.3, 5.9, 4.8, 5.1, 6.2, 6.1, 6.1, 5.3, 6.1, 4.9
Pearson: 6.3, 5.7, 5.8, 4.9, 6.9, 6.8, 7.2, 6.9, 6.8, 7.3

Produce a side-by-side boxplot of this data (see the document "Boxplots in R" on Learn@UW for help with the R commands) and describe in words any key features that you see.

Next, assuming that the data were drawn from normal populations, test the hypothesis at the 5% significance level that the mean weights of apples from Fisher's orchard are smaller than the mean weight from Pearson's. Use a pooled two-sample t-test for this purpose.

Note: For your test, make sure to show your test statistic, rejection region, and the decision of your test.

4. By extension from the pooled two-sample t-test, in a one-way ANOVA we can estimate σ^2 with

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_k - 1)S_k^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1)}$$

So that S^2 is just a weighted average of the *k* sample variances. Using the data for the wheat yield example in class, verify that S^2 as computed above matches the value for MSE from the R output for this data shown in class.

5. Posted at Learn@UW is the dataset Yield_Data.csv, which presents hypothetical wheat yield data (in bushels per acre) similar to that shown in class. Read this dataset into R and do the following:

a) Give for each treatment summary statistics including the sample mean and variance. Note: the function tapply – as described in the document "Boxplots in R" – will be very useful for this!

b) Compute the ratio of the largest treatment variance to the smallest. Given this, would you feel comfortable with the assumption that the variances of the populations from which these data were drawn were the same?

c) Use the functions lm and anova to generate the ANOVA table for this simple experiment.

Also on Learn@UW is the datatset Yield_Data2.csv. Read this dataset into R and repeat parts a) – c) with this new data. What are the primary differences between the two data sets? Do you believe the two sets of data were drawn from the same population?