Homework #2
***** Due Thursday, February 14, 2013 at the beginning of lecture *****

Please work on **L1, and L2** during lab so you can ask questions of the lab instructor.

Questions

Q1. Suppose that $Y_1,\ldots Y_n$ are random variables with an unknown probability distribution. You want to calculate the $P(\bar{Y} \leq 0.1)$. Would it be reasonable to use the normal approximation if n=5? What about when n=100? Explain.

Q2. Suppose a standardized test is given to 100 randomly selected third-grade students in New Jersey. The sample average score $\bar{Y}$ on the test is 58 points and the sample standard deviation s_Y is 8 points.

a. The State of New Jersey plans to administer the test to all third-grade students in the State. Construct a 95% confidence interval for the mean score of all New Jersey third graders. (Hint: the critical value t-statistic for the 95% confidence interval is 1.96.)
b. Suppose the same test is given to 200 randomly selected third graders from Iowa, resulting in a sample average of 62 points and sample standard deviation of 11 points. Set up a null and alternative hypothesis to test whether the mean score in Iowa is different than New Jersey. Set up and calculate the t-statistic to test the difference in the two means. Would you reject the null hypothesis at the 5% level?

Lab Problems

L1. Write a `do` file named `pciformatting.do` that does the following:

- **a.** Creates a `log` file named `pciformatting.log` on your flash drive that records all output.
- **b.** Imports the file "`per capita income 1969 to 2008.csv`" and includes the variable names. This data set includes information on per capita income for each U.S. county from 1969 to 2008 from the Bureau of Economic Analysis (BEA).
- **c.** Renames the variables for the per capita income data for each of the years that have been given names `v1, v2`, etc. as `pci1969, pci1970`, etc. Use the `local` and `foreach v` commands as shown in lab.
- **d.** Saves the data as "`pci 1969 to 2008.dta`."

L2. Write a do file named `appalachian1st.do` that does the following:

a. Creates a `log` file named `appalachian1st.log` on your flash drive that records all output.
b. Imports the file "`Appalachian Dataset 1.csv`" and includes the variable names. This data set includes information on counties in thirteen eastern U.S. states that follow the Appalachian Mountains, including counties within the federally-designated Appalachian Regional Commission region and those that surround them.
c. Renames the variable `emp06` as `total_emp06.` Note: all data variables are from 1990 except for `emp06`.
d. Changes the labels on the key variables as follows:

Variable Name	New Label
manu_emp	Manufacturing Employment 1990
total_emp	Total Employment 1990
total_emp06	Total Employment 2006

e. Creates a new variable `pct_manuemp90` which is `manu_emp/total_emp*100`
f. Creates a new variable which shows the growth rate in total employment between 1990 and 2006, `pct_empgrowth9006`
g. Saves the data as `appalachianupdated.dta` and summarizes the data.
h. Uses STATA commands to manually calculate the t-statistic to test whether the mean of `pct_manuemp90` is different from 20. Also calculate the p-value.
i. Uses the STATA command `ttest` to test whether the mean of `pct_manuemp90` is different from 20.
j. Uses the STATA command `ttest` to test whether the mean of total employment in 1990 is statistically different from total employment in 2006. Make sure you account for differences in variances as shown in lab.
k. Calculates the correlation coefficient between `pct_manuemp90` and `pct_empgrowth9006` using `correlate`.
l. Creates a two-way scatterplot between `pct_manuemp90` and `pct_empgrowth9006` and include a fitted line. Export the graph as `manuempfitted.emf`.
m. Calculates the covariance between `pct_manuemp90` and `pct_empgrowth9006` using `correlate` with the option `covariance`.

.

L3. Using L2 and your output, answer the following:

a. What are the null and alternative hypotheses in L2 parts h and i? Using your results, explain whether or not you would reject the null hypothesis at the 95% confidence level and why.
b. What are the null and alternative hypotheses in L2 part j? Using your results, explain whether or not you would reject the null hypothesis at the 95% confidence level and why.
c. Based on your results, what is the relationship between the percent of total employment in manufacturing in 1990 and the growth rate in total employment from 1990 to 2006 in this region? Does it appear to be linear? Explain.

d. Using the output from L2 parts g and m, calculate the estimate for the coefficient $\widehat{\beta_1}$ or the estimate of the slope variable of the regression of y on x where y is the growth rate of total employment from 1990 to 2006 and x is the percent of total employment that was in manufacturing in 1990. Hint: Your output has the data you need and the formula is in the notes!

- *Submit your responses to the questions, your do files, log files, your exported scatterplot, and a written response to L3. The responses to L3 should be in complete sentences.*
- *Be sure to follow the "Problem Set Guidelines" distributed at the beginning of the semester.*