Questions

1. *Suppose that Y1,… Yn are random variables with an unknown probability distribution. You want to calculate the P(Y-bar≤0.1). Would it be reasonable to use the normal approximation if n=5? What about when n=100? Explain.*

   The rule-of-thumb for using the normal approximation is n=30, so n=5 is too small but n=100 is sufficient.

2. *Suppose a standardized test is given to 100 randomly selected third-grade students in New Jersey. The sample average score Y-bar 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?*

      The null hypothesis is "There is no statistical difference in performance on the standardized exam between third-grade students in Iowa and New Jersey." The alternative hypothesis is "There is a statistical difference in performance on the standardized exam between third-grade students in Iowa and New Jersey."

Lab Problems

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

```
-------------------------------------------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\econ485-master\econ485-master\lab2\pciformatting.log
  log type:  text
 opened on:   7 Feb 2013, 11:24:16

. /*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). */
> cd "C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\econ485-master\econ485-master\lab2";
C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\econ485-master\econ485-master\lab2

. insheet using "per capita income 1969 to 2008.csv", names;
(43 vars, 3140 obs)

. /*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. */
> local i = 1969;

. foreach v of varlist v*{;
  2.         rename `v' farmse`i';
  3.         local i = `i'+1;
  4. };

. /*d. Saves the data as "pci 1969 to 2008.dta."*/
> save "pci 1969 to 2008", replace;
file pci 1969 to 2008.dta saved

. log close;
      name:  <unnamed>
       log:  C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\econ485-master\econ485-master\lab2\pciformatting.log
  log type:  text
 closed on:   7 Feb 2013, 11:24:16
```

```
/*        A. Shawn Bandy


          Lab #2
          February 7th, 2013
*/
/* close previous run do-files */
cap log close
set more 1
clear
#delimit ;


/*a. Creates a log file named pciformatting.log on your flash drive that records all output./*

log using pciformatting.log , replace;


/*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). */

cd "C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\econ485-master\econ485-master\lab2";
insheet using "per capita income 1969 to 2008.csv", names;


/*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. */
local i = 1969;
foreach v of varlist v*{;
          rename `v' farmse`i';
          local i = `i'+1;
};


/*d. Saves the data as “pci 1969 to 2008.dta.”*/
save "pci 1969 to 2008", replace;

log close;
```

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

```
-------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\lab2\lab2\appalachian1st.l
> og
  log type:  text
 opened on:  12 Feb 2013, 10:57:40

. /*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. stat
> es that follow
> //   the Appalachian Mountains, including counties within the federally-designate
> d
> //   Appalachian Regional Commission region and those that surround them.
> */
> insheet using "Appalachian Dataset 1.csv", names;
(8 vars, 555 obs)

. //c. Renames the variable emp06 as total_emp06. Note all data variables are from
> 1990 except for emp06.
> rename emp06 total_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
> */
> label variable manu_emp "Manufacturing Employment 1990";

. label variable total_emp "Total Employment 1990";

. label variable total_emp06 "Total Employment 2006";

. //e. Creates a new variable pct_manuemp90 which is manu_emp/total_emp*100
> generate pct_manuemp90 = manu_emp/total_emp*100;

. //f. Creates a new variable which shows the growth rate in total employment betwe
> en 1990 and 2006, pct_empgrowth9006
> generate pct_empgrowth9006 = (total_emp06 - total_emp) / total_emp * 100;

. //g. Saves the data as appalachianupdated.dta and summarizes the data.
> save appalachianupdated, replace;
file appalachianupdated.dta saved

. /*h. Uses STATA commands to manually calculate the t-statistic to test whether th
> e mean of pct_manuemp90 is different from 20.
> Also calculate the p-value.
> */
>
> sum pct_manuemp90;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
pct_manue~90 |       555    21.16078    11.00066   .7445443   53.52263

. scalar tstat = (r(mean) - 20)/(r(sd) / sqrt(r(N)));

. scalar list;
     tstat =  2.4858751

. //i. Uses the STATA command ttest to test whether the mean of pct_manuemp90 is di
> fferent from 20.
>
> ttest pct_manuemp90 = 20;

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
pct_m~90 |     555    21.16078    .4669518    11.00066    20.24357      22.078
------------------------------------------------------------------------------
    mean = mean(pct_manuemp90)                                    t =   2.4859
Ho: mean = 20                                 degrees of freedom =      554

    Ha: mean < 20               Ha: mean != 20                 Ha: mean > 20
 Pr(T < t) = 0.9934         Pr(|T| > |t|) = 0.0132          Pr(T > t) = 0.0066

. /*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.
> */
>
> ttest total_emp= total_emp06, unpaired unequal;
```

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
total_~p |     555    34486.84    3205.713     75521.6       28190    40783.68
total~06 |     555    42619.79    3788.532     89251.9    35178.14    50061.43
---------+--------------------------------------------------------------------
combined |    1110    38553.32    2483.295    82735.09    33680.83     43425.8
---------+--------------------------------------------------------------------
    diff |           -8132.944    4962.819              -17870.82    1604.931
------------------------------------------------------------------------------
    diff = mean(total_emp) - mean(total_emp06)                    t =  -1.6388
Ho: diff = 0                     Satterthwaite's degrees of freedom =  1078.46

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0508         Pr(|T| > |t|) = 0.1016          Pr(T > t) = 0.9492

. //k. Calculates the correlation coefficient between pct_manuemp90 and pct_empgrow
> th9006 using correlate.
> correlate pct_manuemp90 pct_empgrowth9006;
(obs=555)

             | pct_m~90 pct~9006
-------------+------------------
pct_manue~90 |   1.0000
pct_emp~9006 |  -0.1256   1.0000


. /*l. Creates a two-way scatterplot between pct_manuemp90 and pct_empgrowth9006 an
> d include a fitted line.
>    Export the graph as manuempfitted.emf.
> */
> twoway (lfit pct_manuemp90 pct_empgrowth9006)(scatter pct_manuemp90 pct_empgrowth
> 9006);

. graph export manuempfitted.png, replace;
(file manuempfitted.png written in PNG format)

. //m. Calculates the covariance between pct_manuemp90 and pct_empgrowth9006 using
> correlate with the option covariance.
> correlate pct_manuemp90 pct_empgrowth9006, covariance;
(obs=555)

             | pct_m~90 pct~9006
-------------+------------------
pct_manue~90 |  121.014
pct_emp~9006 | -112.571  6634.08


. log close;
      name:  <unnamed>
       log:  C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\lab2\lab2\lab1problemset.l
> og
  log type:  text
 closed on:  12 Feb 2013, 10:57:41
-------------------------------------------------------------------------------

/*        A. Shawn Bandy
          Lab #2
          February 7th, 2013
*/
/* close previous run do-files */
cap log close
set more 1
clear
#delimit ;

cd "C:\Users\cla-spa206.CAMPUS-DOMAIN\Downloads\lab2\lab2";


//a. Creates a log file named appalachian1st.log on your flash drive that records all output.
log using appalachian1st.log  , replace;

/*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.
*/
insheet using "Appalachian Dataset 1.csv", names;


//c. Renames the variable emp06 as total_emp06. Note all data variables are from 1990 except for emp06.
rename emp06 total_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
*/
label variable manu_emp "Manufacturing Employment 1990";
label variable total_emp "Total Employment 1990";
label variable total_emp06 "Total Employment 2006";

//e. Creates a new variable pct_manuemp90 which is manu_emp/total_emp*100
generate pct_manuemp90 = manu_emp/total_emp*100;

//f. Creates a new variable which shows the growth rate in total employment between 1990 and 2006, pct_empgrowth9006
generate pct_empgrowth9006 = (total_emp06 - total_emp) / total_emp * 100;


//g. Saves the data as appalachianupdated.dta and summarizes the data.
save appalachianupdated, replace;

/*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.
*/

sum pct_manuemp90;
scalar tstat = (r(mean) - 20)/(r(sd) / sqrt(r(N)));
scalar list;

//i. Uses the STATA command ttest to test whether the mean of pct_manuemp90 is different from 20.

ttest pct_manuemp90 = 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.
*/

ttest total_emp = total_emp06, unpaired unequal;

//k. Calculates the correlation coefficient between pct_manuemp90 and pct_empgrowth9006 using correlate.
correlate pct_manuemp90 pct_empgrowth9006;

/*l. Creates a two-way scatterplot between pct_manuemp90 and pct_empgrowth9006 and include a fitted line.
   Export the graph as manuempfitted.emf.
*/
twoway (lfit pct_manuemp90 pct_empgrowth9006)(scatter pct_manuemp90 pct_empgrowth9006);
graph export manuempfitted.png, replace;

//m. Calculates the covariance between pct_manuemp90 and pct_empgrowth9006 using correlate with the option covariance.
correlate pct_manuemp90 pct_empgrowth9006, covariance;

log close;
```

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

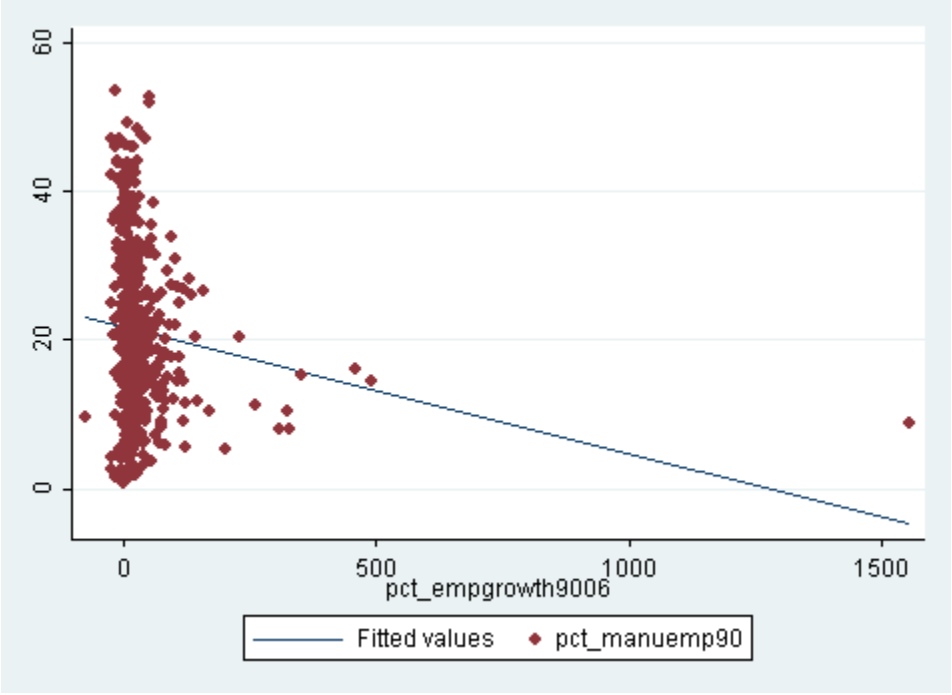
a. *What are the null and alternative hypotheses in L2 parts h and i? Using your results, explain whether or not you would reject the null hypothesis at the 95% confidence level and why.*

The null hypothesis is "The percent of those employed in manufacturing in 1990 was 20 percent." The alternative hypothesis is "The percent of those employed in manufacturing in 1990 was different than 20 percent." I would reject the null hypothesis because 20 is not in the 95% confidence interval (20.24357 to 22.078).

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.*

The null hypothesis is "There is no difference in mean employment in 1990 and 2006." The alternative hypothesis is "There is a difference in mean employment in 1990 and 2006." I would not reject the null hypothesis because zero is in the 95% confidence interval (-17870.82 to 1604.931).

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.*



The correlation coefficient for the percent of total employment in manufacturing in 1990 and the growth rate in total employment from 1990 to 2006 is -0.1256 which suggests a weak relationship between the two. From the graph, I would not necessarily describe this relationship as being linear in nature.

d. *Using the output from L2 parts g and m, calculate the estimate for the coefficient β_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!*

The coefficient β_1 is the covariance of growth rate of total employment and the percent of total employment that was in manufacturing in 1990 divided by the variance of the percent of total employment that was in manufacturing in 1990. $\beta_1 =$ -112.571/121.014 = -0.930231.