PUBPOL 639: ASSIGNMENT 1 - Solutions
Winter 2011

In this assignment you will examine the relationship between school resources and educational outcomes. The data is a sample of 420 school districts in California. It has been posted on the CTools site as well as the online archive. Assume you can treat this data as a random sample of all US school districts. It includes information on average student test scores, expenditure per student, total enrollment, number of computers per district, the average student-teacher ratio, the identity and location of the district (e.g. urban/rural, county), and some socioeconomic measures of students and families in the district (average family income, fraction eligible for free lunch, fraction potentially eligible for TANF).

1. Write down two important causal questions that one could potentially use this data to explore?

Some possible answers:
- *What is the effect of school district spending on students' test scores?*
- *Would greater family income increase school district spending?*
- *What is the effect of computer access on students' test scores?*
- *What is the effect of district size on students' test scores?*

2. Write down two important non-causal questions that one could potentially use this data to explore?

Some possible answers:
- *Do poorer districts have fewer school resources (spending, computers)?*
- *What is the relationship between the student-teacher ratio and test scores?*
- *How does this relationship differ in urban vs. rural districts?*
- *How highly correlated are different types of school resources (spending, computers per student, student-teacher ratio) across districts?*
- *Do larger districts tend to be more urban?*

3. Pick one of your causal questions from above. What is an ideal experiment that you could use to answer it?

What is the effect of school district spending on students' test scores? To answer this, the Federal government could randomly select some districts (the treatment group) to receive large grants that would enable them to increase spending. The spending in other districts (the control group) would be held fixed at its level prior to the experiment. Students' test scores in each district would be assessed prior to randomization and also several times afterwards. The effect of district spending on test scores can be estimated by comparing the test score difference between the two groups sometime after the grants were made. If the randomization was good, there should be no difference in test scores (or any other variable) between the treatment and control groups before the experiment. This set-up would provide an estimate of the effect of district spending in general. To determine whether it is teacher salaries, classroom size, computers, infrastructure, etc that is behind the relationship, one would need to specify that the grants be used for that specific purpose, rather than spent however districts wanted to.

```
sum testscr comp_stu

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     testscr |       420    654.1565    19.05335     605.55     706.75
    comp_stu |       420    .1359266    .0649558          0   .4208333
```

4. Above is a table of summary statistics for two key variables: test scores (`testscr`) and computers per student (`comp_stu`). Use this table of summary statistics to report or calculate the following:

a. Average number of computers per student across all districts

$mean(comp\_stu) = 0.136$

b. Standard deviation and variance of computers per student

sd(comp_stu) = 0.0650 var(comp_stu) = 0.00421

c. Standard error of average computers per student

$$se(mean(comp\_stu)) = \frac{sd(comp\_stu)}{\sqrt{n}} = \frac{0.065}{\sqrt{420}} = 0.00317$$

d. Standard error of average test scores

$$se(mean(testscr)) = \frac{sd(testscr)}{\sqrt{n}} = \frac{19.05}{\sqrt{420}} = 0.930$$

e. 95% Confidence Interval for mean number of computers per student. Show your work.

$$95\%\ \ CI = mean(comp\_stu) \pm 1.96 * se(mean(comp\_stu)) = 0.136 \pm 1.96 * (.00317) = (0.130,\ 0.142)$$

```
by comp_group: sum testscr

-------------------------------------------------------------------
-> comp_group = Low

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     testscr |       210    649.3207    17.95616        609      700.3

----------------------------------------------------------------------
-> comp_group = High

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     testscr |       210    658.9924     18.9309     605.55     706.75
```

5. The table above summarizes test score data for districts with high levels of computer availability and those with low levels of computer availability. "High" is defined as having computers per student greater than the median.

a. Calculate the difference in mean test scores between high- and low-computer use districts

$$\overline{H} - \overline{L} = 658.99 - 649.32 = 9.67$$

b. Calculate the standard error of this difference. Assume that high and low computer-use districts were sampled independently.

$$SE(\overline{H} - \overline{L}) = \sqrt{\frac{(18.93)^2}{210} + \frac{(17.95)^2}{210}} = 1.80$$

c. Test the null hypothesis that the mean test scores for the two types of districts are the same. State your null hypothesis, your alternative hypothesis, your test statistic, and your conclusion.

H_0: *The difference in mean test scores between high and low computer districts in the population is zero*
H_A: *The difference in mean test scores between high and low computer districts in the population is not zero*

$$t = \frac{(\overline{H} - \overline{L})}{SE(\overline{H} - \overline{L})} = \frac{9.67}{1.80} = 5.37$$

Since the test statistic is much greater than the critical value (1.96) we reject the null hypothesis that average test scores have the same mean in districts with high and low levels of computer availability.

6. In this question, you will conduct analysis similar to question 5, but will generate the summary statistics yourself. Create a Stata "do-file" that will enable you to answer the following using the dataset assignment1.dta. Include your do-file and log file as an appendix to your assignment.

a. How many total districts are in the dataset?

There are 420 districts in the dataset

b. How many districts span K-6th grade and how many span K-8th grade?

61 districts span K-6th grade and 359 span K-8th grade.

c. Construct a 95% confidence interval for the mean of district income. Show all work.

First we need to calculate the standard error of the mean of avg income.
se(mean(`avginc`)) = sd(`avginc`) / sqrt(n)
= 7.22589/sqrt(420)
= 0.3526

95% CI = mean(`avginc`) ± 1.96 se(mean(`avginc`))*
= 15.3166 ± 1.96(.3526)*
= (14.63, 16.01)

d. Test the null hypothesis that the mean of district income in the population of all districts is equal to $16,000, using an alpha level of .05. Specify null and alternative hypotheses, test statistic used, and interpret your result.

H_0: The mean average district income in the population is $16,000
H_A: The mean average district income in the population is not $16,000

$$t = \frac{\overline{avginc} - 16.000}{se(\overline{avginc})} = \frac{15.3166 - 16.000}{0.3526} = \frac{-0.6834}{0.3526} = -1.94$$

Since the test statistic is (barely!) less than the critical value (1.96) we fail to reject the null hypothesis that the mean of district income is equal to $16,000 .This result can also be seen by the fact that the 95% CI calculated in 6c includes $16,000. Note that if we had used a 10% significance level (critical value = 1.64) we would have rejected the null.

e. **Test the null hypothesis that the mean of district income is the same in high and low computer use districts, where high is defined as having computers per student above the sample median. To do this, you will need to construct a measure of high/low computer use districts as was used in question 5. State your null hypothesis, your alternative hypothesis, your test statistic, and your conclusion.**

H_0: *The difference in mean average income between high and low computer districts in the population is 0*
H_A: *The difference in mean average income between high and low computer districts in the population is not 0*

$$t = \frac{(\overline{H} - \overline{L})}{SE(\overline{H} - \overline{L})} = \frac{(16.49838 - 14.1348)}{\sqrt{\frac{(8.54515)^2}{210} + \frac{(5.371275)^2}{210}}} = \frac{2.364}{0.696} = 3.394$$

Since the test statistic is much greater than the critical value (1.96) we reject the null hypothesis that average district income has the same mean in districts with high and low levels of computer availability.

7. **In light of your answers to questions 5 and 6, can you conclude that investing in more computers would raise students' test scores? Explain.**

Not necessarily. Indeed we have found that test scores are greater in districts with more computers per student. However, districts with more computers per student may be different than those with fewer computers in other ways that influence student achievement. For instance, in question 6 we found that districts with more computers also have families with higher incomes. Since family income also is likely to influence student achievement, we should not conclude from this data that computers causally increase student test scores. The positive correlation between computer density and test scores may reflect family income or other factors.

---------------------------- Do file --------------------

```
# delimit;
clear all;
set mem 100m;
capture log close;

log using assignment1.log, text replace;

* ============================================================
* Public Policy 639
* Assignment 1
* ============================================================;

* Q1: By hand ;

* Q2: By hand ;

* Q3: By hand ;

* Q4: By hand ;

* Q5: By hand ;

* Q6;
* ------------ ;

* First, load the dataset;
       use assignment1.dta, clear;

* To answer Q6a, describe the dataset. This will tell you how many observations
* are in the dataset. Each observation corresponds to a district;
       describe;

* To answer Q6b, we want to tabulate the variable gr_span;
       tab gr_span, missing;

* To answer Q6c and Q6d, you should print summary statistics for the variable avginc ;
       sum avginc;

* To answer Q6e you need to calculate summary statistics for avginc separately by
* computer use group. This computer use group is actually already constructed for you;

       bysort comp_group: sum avginc;

* Stata also has a built-in function that performs this test automatically. Try this: ;
       ttest avginc, by(comp_group);

* Q7: By hand;
log close;
```

---------------------- Log file ---------------------

```
       log:  assignment1.log
  log type:  text


. * =============================================================
> * Public Policy 639
> * Assignment 1
. * =============================================================;
. * Q1: By hand ;
. * Q2: By hand ;
. * Q3: By hand ;
. * Q4: By hand ;
. * Q5: By hand ;
. * Q6;
. * ------------ ;
. * First, load the dataset;
.         use assignment1.dta, clear;

. * To answer Q6a, describe the dataset. This will tell you how many observation
> s
> * are in the dataset. Each observation corresponds to a district;
.         describe;

Contains data from assignment1.dta
  obs:           420
 vars:            11                          13 Jan 2010 10:43
 size:        48,300 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
county          str18  %18s                   County name
district        str53  %53s                   District name
gr_span         str8   %8s                    Grade span
enrl_tot        float  %9.0g                  Total enrollment
calw_pct        float  %9.0g                  Percent qualifying for CalWORKS
meal_pct        float  %9.0g                  Percent qualifying for reduced
                                                price lunch
testscr         float  %9.0g                  Average test score
comp_stu        float  %9.0g                  Computers per student
expn_stu        float  %9.0g                  Expenditures per student
avginc          float  %9.0g                  Average income in district
comp_group      float  %9.0g       comp_groupl
                                              Computers per student - High
-------------------------------------------------------------------------------
Sorted by:  comp_group

.                 * To answer Q6b, we want to tabulate the variable gr_span;
.         tab gr_span, missing;

 Grade span |      Freq.     Percent        Cum.
------------+-----------------------------------
      KK-06 |         61       14.52       14.52
      KK-08 |        359       85.48      100.00
------------+-----------------------------------
      Total |        420      100.00

.         * To answer Q6c and Q6d, you should print summary statistics for the v
> ariable avginc ;
.         sum avginc;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      avginc |       420    15.31659     7.22589      5.335     55.328

.         * To answer Q6e you need to calculate summary statistics for avginc se
> parately by
> * computer use group. This computer use group is actually already constructed
> for you;
.         bysort comp_group: sum avginc;
```

```
--------------------------------------------------------------------------------
-> comp_group = Low

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      avginc |       210     14.1348    5.371275      5.699      43.23

--------------------------------------------------------------------------------
-> comp_group = High

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      avginc |       210    16.49838     8.54515      5.335     55.328


. * Stata also has a built-in function that performs this test automatically. Tr
> y this: ;
.         ttest avginc, by(comp_group);

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     Low |     210     14.1348    .3706532    5.371275     13.4041     14.8655
    High |     210    16.49838    .5896713     8.54515    15.33591    17.66084
---------+--------------------------------------------------------------------
combined |     420    15.31659    .3525873     7.22589    14.62353    16.00965
---------+--------------------------------------------------------------------
    diff |           -2.363578    .6964884               -3.732635   -.9945221
------------------------------------------------------------------------------
    diff = mean(Low) - mean(High)                                 t =  -3.3936
Ho: diff = 0                                     degrees of freedom =      418

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0004         Pr(|T| > |t|) = 0.0008          Pr(T > t) = 0.9996

.         * Q7: By hand;
. log close;
      name:  <unnamed>
       log:  assignment1.log
  log type:  text
--------------------------------------------------------------------------------
```