API-201 Section Z

Problem Set #1

Fall 2014

Due 9/11/2013 at 1:00pm (before class)

Instructions: You may provide typed answers or handwritten answers (so long as they are **clearly legible**), but you should carefully write out all code answers when requested, either way. You should print all charts and graphs and attach them at the end. Please start the problem set early, in case you have any trouble with the software.
Remember: Google is your best friend when it comes to STATA and Excel help! Please search for the answer to your problem online before emailing us.

Last Name: _____

First Name: _____

Group members you worked with: [1]

_____

_____

_____

_____

Please use this as cover page and remember to:

- Turn in problem sets to your section only
- Turn them in ***outside*** class ***before*** the class starts
- Use a stapler and not a paper clip

[1] You are reminded that this is a Type II assignment, i.e. you are encouraged to work in a study group, but must submit your own hand- or type-written solutions (please refer to syllabus for details).

QUESTION ONE: CASE STUDY - PINE STREET INN

Please read Part A of the case study on the Pine Street Inn. We have provided you with data on the lengths of stay for guests at the Pine Street Inn in 2007 and 2008, which formed the basis for the second of two consulting studies. (The first suffered from poor data quality). Using the dataset "*Pine Street Inn Length of Stay Data.xls*," please calculate:

1) Mean length of stay at Pine Street Inn
2) Median length of stay at Pine Street Inn
3) The maximum and minimum lengths of stay at Pine Street Inn
4) The 75^{th}, 95^{th}, and 99^{th} percentiles of the distribution of length of stay
5) Total number of bed nights spent at Pine Street Inn
6) Total number of guests who stayed at Pine Street Inn
7) Please fill in the following table (also reproduced in the data worksheet):

Length of Stay	*Number of Guests*	*Number of Bed Nights*	*Fraction of Guests*	*Fraction of Bed Nights*	*Average Bed-Nights per Guest*
3 Days or Less					
4 to 10 Days					
11 to 35 Days					
36 to 150 Days					
151 Days or More					
Total					

Using the data provided with the case, please make the following graphs:

8) A line-graph with two lines, showing for each number of nights the fraction of guests (first line) and fraction of bed nights (second line) accounted for by stays of a given length. The horizontal axis should be number of nights (please use the five intervals provided i.e. 3 days or less, 4 to 10 days, etc), and the vertical axis should run from 0 to 100%.
9) A histogram showing the distribution of lengths of stay at the Pine Street Inn. Please use intervals of the same fixed width (you may choose the number of intervals and the intervals' width). **Do not** use the five intervals from the previous question, as these have variable widths.

Finally, write one short paragraph (no more than 150 words) answering the questions:

10) After you had calculated the mean in (1) above, how did the other statistics you calculated further inform your understanding of the amount of time different men stayed at the Pine Street Inn? How does this affect your understanding of the issues facing the Pine Street Inn?

QUESTION TWO: USING STATA TO ANALYZE AT STATE SPENDING DATA

This question is meant to help introduce you to STATA. Rather than give you a nice, clean dataset, we're going to walk you through how to get data from a public source, import it into STATA, and then turn the raw dataset into a form that we can use very easily. We will tell you each step you need to do to get to a clean dataset—we just want you to see what sorts of things STATA can do. It's fine if you don't totally understand what is going on, just get your hands wet. We'll then ask you to do your own statistical exploration of the data once it's set.

We'll be looking at state and local federal expenditures by state in the USA. The Census of Governments, conducted annually by the Census Bureau, keeps track of the total expenditures within states (including local governments) by category, as well as the source of funds at each level (taxes, fees, or grants from the Federal government). In this exercise, we'll look at how state and local expenditures vary by state—specifically in three categories: health programs (which excludes public hospital operation), basic education, and (criminal) corrections. A challenge, however, is that although the data is all available online, it's kept in a very disorganized manner.

Getting Started:

The main website for the State and Local Government Finances survey is at:

https://www.census.gov/govs/local/

The most recent data is for 2011, which is what we'll use.

Start by investigating the summary report, http://www2.census.gov/govs/local/summary_report.pdf In "Appendix Table A-1," you can find total state/local revenues and expenditures by each category.

A.) The categories we're interested in are components of *Direct* Expenditures. Just to make sure we're on the same page, what was the total *direct expenditure* 2011 according to the summary table? (Note: it is also given on the line "direct expenditure by function") What was the total spent on Elementary and Secondary Education in 2011? Health? Corrections? (*make sure you check the units!)*

Answer:
Total Direct Expenditure: $3,158,332,352,000
Total Elementary and Secondary Education: $564,862,357,000
Total Health: $ 82,369,241,000
Total Corrections: $73,150,911,000

Cleaning Data:

Now, download the data. The documentation is given in the document at
http://www2.census.gov/govs/local/11filelayout.pdf

The data is in the ZIP file
http://www2.census.gov/govs/local/11statetypepu.zip

Download this file, and unzip it (right click on the ZIP file in your browser and select "unzip"). You'll wind up with a text file called "11statetypepu.txt" . Create a folder called "statelocal" on

your desktop and move the text file there. Open it and take a look. Look at the documentation file for a few moments.

B.) Write 2-3 sentences about how the data appear to be organized. How can you tell where columns start and end? How would you find how much Delaware spent on Construction Expenditures for Primary and Secondary Education in this data? (Hint: You might want to go through the next few steps and then come back to this.)

Suggested Answer: The data are in a text file, where the exact spacing of characters allows one to determine where data columns begin and end based on how many spaces in you are. Rather than putting the data in a rectangular form with states as rows and different variables as columns, the raw data is presented as one giant column of dollar amounts—each row contains information about what state, variable, year, and government level the dollar amount corresponds to. One could use the code for Delaware state and local combined (081), the code for Construction-Elementary and Secondary Education (F12), and the right year to find the value.

Now open STATA. Once it is open, go to File->Change Working Directory and select the folder you created called "statelocal."

Reading in "fixed format data": The text data is arranged so that each data cell takes up a fixed number of characters (including spaces) in each row. So, for example, the first two numbers in each row are always a two-digit code that corresponds to a State. We need to tell STATA where to look for each piece of data in the text file. I've set this up for you. Type the following into the command line in STATA:

```
infix state 1-2  govlev 3 str spendcat 5-7 amt 9-20 yr 34-35
                 using "11statetypepu.txt"
```

This tells STATA what columns contain which data. Note that raw data is not always in fixed format. Later, we'll bring in raw data in comma-separated format. It's important that you changed the working directory to the right folder!

Now, let's save this as a STATA data file. Type

```
save statelocal_raw
```

into the command line. You can type **browse** into the command line to explore your data table. You should have five variables in the data set: the year of the survey (*yr)*, the state code (*state*), the level of government code (*govlev)*, the code for spending category (*spendcat*) and the amount expended in that category in DOLLARS *(amt)*.

Turning this into the dataset we want.
We have every category from every level of government in every state in every year, one-by-one. But not only do we not need all of this (we just want one year, four categories of spending, and total state/local spending), we also don't really want these in one long list. We want a table for 2011 that shows has one column for each kind of spending, and a row for each state. Then we can do our usual statistical analyses on those items. Let's get there as follows:

- First, deal with the year variable. Type in **`tabulte yr`** to see all the different years in the data. This doesn't look very helpful, does it? We can drop an entire variable by typing in **`drop yr.`**
- Second, deal with the govlev variable. From the first page of the documentation, we see that "1" means state and local combined expenditure, "2" means state only, and "3" means local only. We are only interested in the combined total.
    - We can limit the data to the items that refer to totals by keeping only the items that are marked with a "1" in the current format. Type in **`keep if govlev==1.`** (Note: we use *two* "=" signs whenever we're interested in whether a statement is true or false. We are keeping observations when it is TRUE that govlev is equal to 1).
    - Now that we don't have any other levels of government, we don't need to keep the column saying what the level of government is. Type in **`drop govlev`**
- Third, type **`replace amt=amt/1000.`** The raw data are in *THOUSANDS of DOLLARS*, but we put the data in terms of *MILLIONS OF DOLLARS* instead here. This replaces the existing value in *amt* with that original value divided by 1000.

Now, we need to make columns for each category. We're going to use a few tricks here, just follow the instructions:

- Type **`generate cat=0`**
    - This creates a new variable called "cat" that has all zeroes to start. (Note, we just used one "=" here. That's because we're telling STATA to set "cat" equal to one, not to test whether it is true that "cat" is already equal to one, like before).
- Type **`replace cat=1 if spendcat=="E32" | spendcat=="F32" |spendcat=="G32"`**
- Type **`replace cat=2 if spendcat=="E12" | spendcat=="F12" |spendcat=="G12"`**
- Type **`replace cat=3 if spendcat=="E04" | spendcat=="F04" |spendcat=="G04" | spendcat=="E05" | spendcat=="F05" |spendcat=="G05"`**
    - "Health," "Primary/Secondary Education," and "Corrections" were subdivided into several subcategories, so we picked out the subcategories that go into each type (according to http://www2.census.gov/govs/local/methodology_for_summary_tabulations.xls), and gave them a "1" if health, a "2" if education, and a "3" if corrections. Since the original category codes are not numbers, we need to treat them as *strings* , which means we need to put all values in quotes. In STATA language " | " means "OR". Finally, "replace" replaces a data cell with the value we tell it to *if* the statement that follows is true. Thus, the first command replaces the zero in *cat* with a 1 if *spendcat* is one of the three specific subcategories corresponding to health.
- Type **`drop spendcat`**
    - With our simplified categories, we don't need the old one.
- Type **`drop if cat==0`**

    - The leftover zeros are not relevant to Health, Primary/Secondary Education, or Corrections
- Type **`collapse (sum) amt, by (state cat)`**
    - The "collapse" command in STATA is very powerful. Right now, we have one column with spending amounts for each *subcategory* in each state. But we want the total of the subcategories *combined* into large categories for each state. We thus "collapse" the total sum of the amounts of each subcategory into amounts for each primary category in each state.
- Type **`reshape wide amt, i(state) j(cat)`**
- Type **`rename amt1 health`**
- Type **`rename amt2 educ`**
- Type **`rename amt3 corrections`**
- Type **`save statelocal2`** to save this dataset
    - Type **`browse`** to explore your data table.
- Type **`sort state`** to order by state number.

C.) Discuss in a few lines what we just did to our data. What happened when you used the "reshape wide" command? Rather, how are the data arranged now?

Suggested Answer: The reshape wide command moved health, education, and corrections spending into separate columns/variables, based on the codes we had created. Now we have a nice, rectangular data set!

D.) Go back to the documentation. Our *health* variable is based on the categories E32, F32, and G32. What does this health expenditure value actually measure?

Each category is the sum of operating expenditures, construction costs, and other capital costs. Thus, total expenditure is operating cost + capital outlay, or operating payments + investment payments. (Corrections were split into "institutions and other", but it doesn't matter if this is not mentioned).

E.) What is one reason it is useful to look at state and local *combined* expenditure, rather than looking separately at state and local. (Hint: Which government plows JFK Street in front of HKS? Who plows the Turnpike? Would all states' split the responsibility the same way? There isn't one right answer.)

Suggested Answer: In the USA, different responsibilities are given to local governments in different states. It's hard to compare spending levels across states without looking at the total of state and local government amounts. The total measure allows us to look at how states differ on the whole in their sub-federal expenditures.

Adding on state names:

DOWNLOAD THE FILE statenames.csv FROM THE COURSE WEBSITE INTO YOUR *STATELOCAL* FOLDER

Right now our state names are written in code. I've created a file that matches each number in the variable *state* to the corresponding name (it also includes 2011 population estimates). I've saved it as a .csv file—each cell is separated by a comma, and a new line starts with a full return. Excel and STATA can import .csv files directly. Download the file (statenames.csv) from the course website into your statelocal folder. We will import these names and merge them to the main data set.

- Type **`clear.`** But make sure you typed **`save statelocal2`** already first!
- Go to File→Import→Text Data (Delimited, .csv…)
    - In the import window, browse and selected the statenames.csv file
    - The first line of the folder has variable names, and the delimiter is a comma, so make sure the "use first row for variable names" option says "always" and the Delimiter option is set to "Automatic"
    - If the preview looks okay, hit OK
- Type **`sort state`**
- Type **`merge 1:1 state using statelocal2`**
    - This tells STATA to match records in this file to the ones in *statelocal2*—in both files, *state* refers to the numeric code for states. Each code appears exactly once in each file—so we call the merge one-to-one ("1:1"). It's important that *state* have the same name in each file and that we sorted each file by *state* first. It's also important that you have your working directory set to your "statelocal" folder.
- Type **`drop _merge`**
- Type **`drop if state==0`** to get rid of the "US TOTAL" line.
- Type **`save spendclean`** to save the cleaned data.
- Type **`browse`** to look at your data.

Looks better, right? *spendclean.dta* is now the clean dataset we were hoping for. Now we can calculate statistics on health, education, and corrections expenditures using STATA's built-in functions.

F.) The raw data were in thousands of dollars. What units are they in now? What did we do to put them in these units?

Answer: The original data was in dollars, but by dividing by 1000 we converted the units to millions of dollars.

G.) The dataset we created includes 2011 population estimates from the Census in the variable *population*. Use the **`generate`** command create three variables, *pc_health*, *pc_educ*, and *pc_corrections* that express the expenditure in ***Dollars Per Capita*** for health, education, and

corrections, respectively. (Hint: Each *?* is something you need to fill in. What units are the original spending variables in?)

Please copy the code you used for your answer:

Answer:

*gen pc_health = (health * 1000000)/population*
*gen pc_educ = (educ * 1000000)/population*
*gen pc_corrections = (corrections * 1000000)/population*

Save your data here! You don't need to change the dataset moving forward, but if you do by accident, all the remaining questions begin with the data you have at this point.

Statistical Analysis in STATA

H.) Calculate and present summary statistics (mean, standard deviation, median, inter-quartile range) for each of the variables *pc_health*, *pc_educ*, and *pc_corrections*. Present these results in a table with a rows for each variable, and a column for each summary statistic—please round everything to one decimal point. (Hint: Type "help summarize" to see how this command works). Include DC, so you should have 51 observations.

Answer:
```
summ(arize) pc_health pc_educ pc_corrections, d(etail)
```

	Mean	Std Dev	Median	IQR
Health	264.1	112.9	235.8	133.4
Education	1857.4	485.9	1786.7	388.3
Corrections	219.0	67.5	198.4	87.2

NOTE: If you did not get a mean of about 264.1 for per capita health expenditures, something has gone awry! Please make sure you followed the instructions correctly. If you continue to struggle, contact a CA or TF, and we can provide you with the properly cleaned data—you'll still be able to get plenty of partial credit.

I.)
(i) Let X_i denote the per capita health spending of state *i*. Write down the formula for the variance of health spending in the data.

Answer:

$$s^2 = \frac{(x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_n - \bar{x})}{n - 1}$$

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

where

$$\bar{X} = \frac{x_1 + x_2 + ... + x_n}{n}$$

$$\bar{X} = \frac{1}{n} \sum x_i$$

(ii) Instead of writing it out as a symbolic formula, write out this formula as a plain-English sentence that tells you how the variance is calculated. (For example, one could write out $\sqrt{a^2 + b^2}$ as "take the square of *a* and *b* and add the resulting amounts together. Then, take the square root of this sum.")

Suggested Answer:

As we do below, the formula says to take the difference of each observed value and the average for all observed values, then take the square each difference, then sum up all of these squared differences, and divide the sum by one less than the total number of observations

(iii.) Now do the following:

- Generate a new variable *x1* for which the value is equal to the mean *pc_health* from question H for all states.
- Generate another new variable *x2* that is equal to *pc_health – x1* for each state.
- Generate another new variable *x3* that is equal to the square of *x2* for each state.

Show how you can calculate the variance and standard deviation of *pc_health* based on *x3*. Display the line(s) of STATA code you would use. What value do you get?

Answer:

All you need to do is sum up all the values of x3 and then divide by 50 (which is n-1) to get the variance. The square root of the variance is the SD. The variance is approximately 12740 (subject to rounding error less than 1), and the SD is 112.9, as before.

Example code:

```
gen x1=264.0836
gen x2 = pc_health - x1
```

```
gen x3 = x2^2
egen x4 = total(x3)
gen var = x4/50
gen sd = sqrt(var)
```

J.) Draw a scatterplot with *total* health expenditure on the *x* axis and *total* primary/secondary educational expenditure on the *y* axis. Present this scatterplot (copied from Stata), carefully labeled with the units specified. Then, draw a scatterplot with *per capita* health expenditure on the *x* axis and *per capita* primary/secondary educational expenditure on the *y* axis. Present this scatterplot (copied from Stata), carefully labeled with the units specified. Comment briefly in 1-2 sentences on what the plots look like, and why you'd expect them to be so different from each other.

Answer:

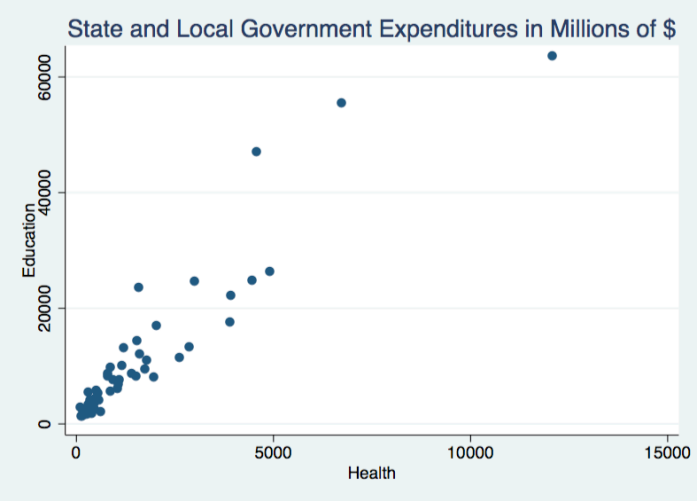
The first graph shows high correlation, but this is because big states spend more on everything in aggregate simply by virtue of their larger population. When we look at spending per person, the correlation turns out to be much weaker.

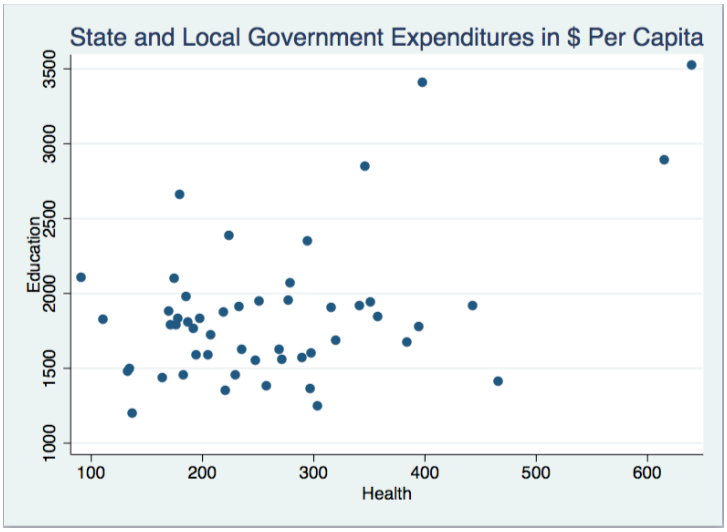
The code would be:

```
twoway (scatter pc_educ pc_health), ytitle(Education)
xtitle(Health) title(State and Local Government Expenditures in
$ Per Capita)
```

and

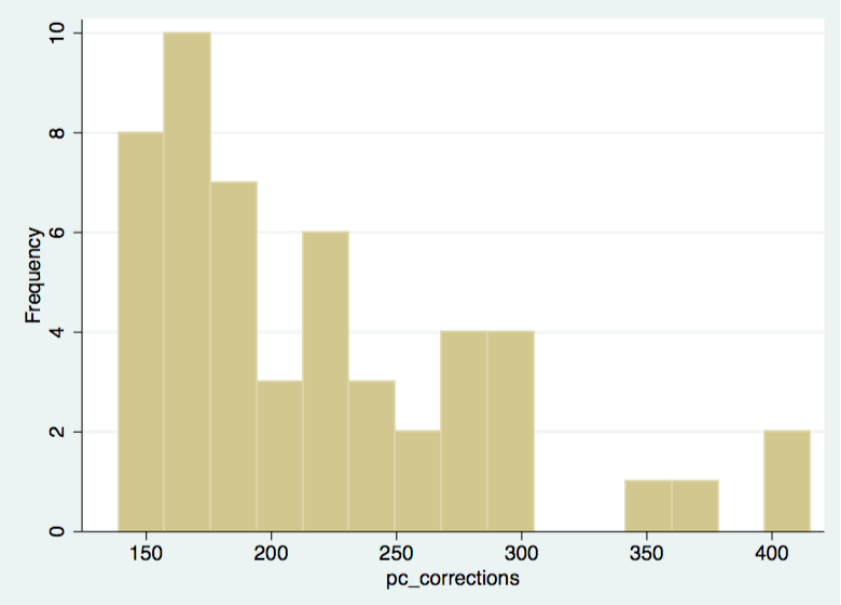
```
twoway (scatter educ health), ytitle(Education) xtitle(Health)
title(State and Local Government Expenditures in Millions of $)
```





K.) Draw a histogram of *pc_corrections*, with 15 bins and frequencies displayed on the Y axis. You can either go to Graphics→Histogram on the drop down menu, or use the **`histogram`** command in the command line (type **`help histogram`** to see how this works). Don't worry about labeling this one.

Answer:
`histogram pc_corrections, bin(15) freq`



L.) Now, let's imagine that Alaska decides its only priority is to improve the correctional facilities. Type the following into STATA:
`replace  pc_corrections=10000 if statecd == "AK"`

Now, calculate the mean, standard deviation, median, and IQR for *pc_corrections*. Explain briefly in 1-2 sentences what is different from what you found in part H, and what this tells you about means, medians, and standard deviations.

Answer:
The median and IQR are unchanged, but the mean and the standard deviation are dramatically affected by this outlier. Alaksa was in the top quartile originally. Making it an outlier does not change percentiles of observations that were originally lower ranked, since their relative ranking is the same as before. The average and SD are quite sensitive to outliers in small datasets, though..

	Mean	Std Dev	Median	IQR
Corrections	407.3	1371.5	198.4	87.2