17.871, Political Science Lab
Problem set # 1: Using STATA
Spring 2013

Handed out: Feb. 13
Due: Feb. 25, *at the beginning of class. (Please print before coming to class. Unfortunately, the printers in our classroom are ridiculously slow and loud.)*

Write a do-file that responds to all the parts of this problem set. Turn in the do-file and the log-file that shows that the do-file works. Clearly label each do-file. Make a habit of writing comments in the do-file, to help us and you keep track of things. (You can make non-executed comments in a do-file using the front slash and asterisks as follows:

/* THIS IS WHAT A COMMENT LOOKS LIKE IN A DO-FILE */

A word on collaboration: It is fine for you to seek help from or give it to another member of the class. However, I want the do-files you write to be your own. (In other words, no sharing code.)

Part I: Getting data into STATA (five points for the part)

Data comes in many forms. Here's one way to get data into Stata. Using a text editor (such as EMACS), type the text from Exhibit 1 in the handout "How to Use the Stata infile and infix Commands" into Athena and save it in a file named scores.dat on your home directory.

Write the code that will create a Stata data set from this raw data and save it as a file called "scores.dta". Use the list command to view your data.

Part II. Collapsing a data set (five points for the part)

In the Examples folder you will find a dataset named spaesubset2012. This a subset of a survey I have recently conducted, of the experience of voters in the 2012 presidential election. (The larger dataset is what was used to produce the New York Times infographic you can find at http://www.nytimes.com/interactive/2013/02/05/us/politics/how-long-it-took-groups-to-vote.html?smid=tw-share.) There are 200 observations from each state.

One thing to notice about the dataset as you begin is that although all the variables are numerical, I have assigned codes to all the numerical values, except for the "weight" variable, to make most types of analysis easier to understand. For instance, if you type "tab q4" into Stata, this is what you get back:

```
. tab q4

                        Mode of voting |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
In person on Election Day (at polling p |      5,979       63.40       63.40
  In person before Election Day (early) |      1,654       17.54       80.94
            Voted by mail (or absentee) |      1,770       18.77       99.71
                             I don't know |         27        0.29      100.00
---------------------------------------+-----------------------------------
                                 Total |      9,430      100.00
```

However, the variable q4 is actually stored as numerical code, where the value 1 means "In person on Election Day…," 2 means "In person before Election Day (early)," etc. (Refer to Kohler and Kreuter, pp. 15–16 for a discussion of variable and value labels.) To see the numerical codes, you need to use the "nolabel" subcommand, which just prints the table using the numerical values, not the labels. So, if you type "tab q4,nolabel" into Stata, this is what you get back:

```
. tab q4, nol

    Mode of |
     voting |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      5,979       63.40       63.40
          2 |      1,654       17.54       80.94
          3 |      1,770       18.77       99.71
          4 |         27        0.29      100.00
------------+-----------------------------------
      Total |      9,430      100.00
```

(Note that I abbreviated the "nolabel" subcommand "nol".

Here is the assignment:

1. Create a dummy (binary) variable that is equal to 1 if the respondent reported waiting 30 minutes or longer to vote, 0 otherwise.
2. Use the "collapse" command to create a new dataset that contains (1) the percentage of respondents who waited 30 minutes or longer to vote in each state and (2) the number of respondents who were included in the calculation of this percentage for every state. (Note that this will not be 200, despite the fact that this is the sample size for each state, because not everyone in each state actually voted, or voted in person.) When you do the collapsing, you will need to "weight" each observation, to correct for the fact that the sample did not perfectly match the demographic profile of each state. (See Kohler and Kreuter, pp. 60–65 for a discussion of weights in Stata.) You will use the command [aw=weight] in the collapse command. Type "help collapse" to see how to use weights in the collapse command.
3. List the first ten observations of this dataset in your log-file.
4. Save this file with the name wait_by_state_2012.dta.

Part III. Merging two data sets (eight points for the part)

Also in the Examples folder is another file named spaesubset2008, which is a subset of the dataset from the survey I conducted in 2008 about voting in the United States. This file has the same variables as spaesubset2012, except that the names of the variables are different.

1. Create a dataset that is parallel to the dataset you created in Part II, only this time with data from 2008.
2. Save this file by the name wait_by_state2008.dta
3. Merge the 2012 and 2008 datasets together.
4. List the names of the states in which a greater fraction of people waited longer than 30 minutes to vote in 2012 than in 2008. (This will require you to use the "if" statement in the "list" command.) I only want the state names listed, not the whole dataset.

Part IV. Reshaping a dataset (ten points for the part)

Also of interest in research on the time people waited to vote in 2012 is whether people waited longer in line when they vote early, compared to when they vote on Election Day. Both of the datasets you have just used have a variable that records the mode in which the respondents voted — in the traditional way, on Election Day; early, before Election Day; or by mail or absentee ballot.

1. Using the "reshape" command,[1] start with the spaesubset2012 dataset, and create a new dataset that reports (1) the percentage of respondents who voted *on Election Day* who spent 30 minutes or longer voting, (2) the percentage of respondents who voted *early* who spent 30 minutes or longer voting, (3) the number of people involved in the calculation of the percentage of Election Day voters, and (4) the number of people involved in the calculation of the percentage of early voters.

   Hint: When you use the "collapse" command, you will need to include two variables in the "by" subcommand, regstate and q4. Also, don't forget to weight your results.

2. List out the values of the dataset.

[1] The "reshape" command is covered in pp. 238–241 of Kohler and Kreuter.