17.871, Political Science Lab
Spring 2015
Problem set # 2

Handed out: March 2
Due: March 11

Submit

1. A paper copy of your answers to each question, your graphs, and calculations.
2. A log file that shows the answers for each part.

You may work on these together, but you must write them up separately.

Part I: Distributions (2 points for each variable, 6 points total)

Consider the following four variables. Complete the following exercises with respect to each:

1. Create a dataset of each of the variables described below. (That is, create three datasets, one for each variable.) The datasets should include the variable mentioned along with the case identifier(s) mentioned (congressional district, state, etc.) Find the mean, standard deviation, skewness, and kurtosis of the variables. Also produce a histogram, or other appropriate graph, that shows the distribution of the variables. (Produce a do-file that reads in the data, saves a Stata file in .dta format, and then performs the statistical procedures requested.)
2. In order to find the answer to step 1, find the original, authoritative sources for the data you are describing. (By *original, authoritative source,* I mean a source that produces the data in reasonably raw form. Wikipedia is not an original, authoritative source. I also mean you must avoid pointing out a dataset in a spreadsheet that someone has posted on the web, although you may find that someone has already done the data entry for you for this exercise.) Give the citations to how to locate the data, either a traditional bibliographic citation (author, title, etc.) of a book or a URL of an electronic data source.

Here are the variables:

1. The percentage of votes received by Thom Tillis, Republican nominee for Senate, in each county in North Carolina in the 2014 senatorial election. (Hint: Election returns in North Carolina are reported by the North Carolina State Board of Elections)
2. Per capita emissions of CO_2 in 2010 by country.
3. Infant mortality rates by country in 2014.
4. The percentage of women in the U.S. aged 15 to 19 who gave birth in the past 12 months, by county.

Part II. Public opinion (2 points for the first two steps, 6 points for the third step, 10 points total)

You will find a dataset named cces12_common_subset in the Examples folder of the 17.871 course locker. This is a 50% random sample of the dataset from the 2012 Cooperative Congressional Election Study, which MIT had a major hand in originating. You will use this dataset to answer the following questions about perceptions voters hold about the political parties.

It is a huge dataset. The following are the variables you should focus on:

- CC308a: Institutional approval of President Obama
- CC308b: Institutional approval of Congress
- CC308c: Institutional approval of the Supreme Court
- CC334A: The respondent's own placement of him/herself on a 7-point ideological scale, where 1 is the most liberal and 7 is the most conservative.
- pid3: a "three point" party measure, that records whether the respondent is a Democrat, Republican, or Independent (plus "other" and "not sure").
- CC424: a 5-point scale measuring whether the respondent has a positive or negative view of the Tea Party movement.

1. Generate a system of histograms *that you would be proud to put in a widely distributed paper* that shows the distribution of institutional approval for President Obama, Congress, and the Supreme Court. Treat as missing anyone who gives a "not sure" answer.
2. Generate histograms that allow us to compare approval of the Supreme Court among Democrats, independents, and Republicans.
3. Generate histograms that allow us to compare the institutional approval of the Supreme Court among Republicans who have a positive view of the Tea Party movement *vs.* those who have a negative view.
4. Report descriptive univariate statistics (such as mean, standard deviation, skewness, etc.) that would allow you to compare how approval of President Obama, Congress, and the Supreme Court differs among supporters of different political parties.
5. In addition to the do-file, write a paragraph or two in which you describe how this data helps describe how members of the political party view the three branches of the federal government. (In other words, write a couple of paragraphs in which you describe the histograms and statistics you generated here.)
    a. Do the same, this time comparing Republicans who have a positive view of the Tea Party movement with those who has a negative view.

Part III. Finding and analyzing data about military spending (5 points for first step, 5 points for second step, 10 points total)

Navigate to the Correlates of War website (http://www.correlatesofwar.org/) and locate and download the National Material Capabilities dataset, along with the codebook.

1. Find the mean, standard deviation, minimum, maximum, and the number of observations for *milper* separately for the following years: 1900, 1915, 1945, 1980, 2005. Create a table in your write-up that records these values.
2. Propose a (historical) explanation for the changes you observe in these different time periods, addressing the mean, standard deviation, *and* number of observations.

Part IV. Finding and merging data (4 points for the first and last step, 2 points for the other steps, 14 points total)

In this problem, you will, first, calculate the fraction of adults (18 and older) in each state who hold a driver's license. Second, you will explore the data and identify states that stand out as outliers — states with either unusually low or unusually high values of a measure.

You may use the datasets that were downloaded as part of the exercise you did with Kate McNeil, the MIT data librarian.

1. Create a new dataset that contains one observation for each state and variables as follows:
    a. A series of variables that records the number of people who hold driver's licenses in the following age categories: 18-19, 20-24, 25-29, 30-34 . . . 80-84, and 85 and older.
    b. A series of variables that records the number of residents in the state in the same age categories as in the point above.
2. Draw a graph that shows the percentage of each age group that holds a driver's license, nationwide. (Hint: the *x*-axis should be the age categories and the *y*-axis should be the percentage.)
3. Draw a graph that summarizes the overall rate of driver's license possession in each state.
4. Draw a graph that summarizes the overall rate of driver's license possession among those who are 18 and 19 years old in each state.
5. Write a couple of paragraphs in which you discuss whether any states appear to be outliers, in terms of driver's license possession — either for the entire population (step 3) or the 18-19 year-olds (step 4). What do you think causes these states to be outliers?

Part V. Describe the data for your intended final project (6 points)

Write a couple of paragraphs in which you state what you plan to write your final paper about.

1. Describe the general question you intend to pursue.
2. Identify one dependent variable you will pursue and one independent variable you will use to explain your dependent variable.
3. Identify the data source that will supply values of your dependent and independent variables. Be specific. Provide a precise bibliographic citation or a URL link.

This does not commit you to a final project. However, you will be graded on how well the dependent and independent variables match the general question, as you state it, and whether the dataset(s) you identify will allow you to generate the variables you need.