HW2P2

September 21, 2016

0.0.1 Homework 2: More Exploratory Data Analysis

0.1 Gene Expression Data and Election Polls

Due: Thursday, September 29, 2016 11:59 PM

Submission Instructions To submit your homework, create a folder named **lastname_firstinitial_hw#** and place your IPython notebooks, data files, and any other files in this folder. Your IPython Notebooks should be completely executed with the results visible in the notebook. We should not have to run any code. Make sure to share the private repo with my github account (mdog) and submit the repo path through blackboard.

0.2 Introduction

John Tukey wrote in Exploratory Data Analysis, 1977: "The greatest value of a picture is when it forces us to notice what we never expected to see." In this assignment we will continue using our exploratory data analysis tools, but apply it to new sets of data: gene expression and polls from the 2012 Presidental Election and from the 2014 Senate Midterm Elections.

First: You will use exploratory data analysis and apply the singular value decomposition (SVD) to a gene expression data matrix to determine if the the date that the gene expression samples are processed has large effect on the variability seen in the data.

Second: You will use the polls from the 2012 Presidential Elections to determine (1) Is there a pollster bias in presidential election polls? and (2) Is the average of polls better than just one poll?

Finally: You will use the HuffPost Pollster API to extract the polls for the current 2014 Senate Midterm Elections and provide a preliminary prediction of the result of each state.

Data We will use the following data sets:

1. A gene expression data set called `exprs_GSE5859.csv` and sample annotation table called `sampleinfo_GSE5859.csv` which are both available on Github in the 2014_data repository: expression data set and sample annotation table.

2. Polls from the 2012 Presidential Election: Barack Obama vs Mitt Romney. The polls we will use are from the Huffington Post Pollster.

3. Polls from the 2014 Senate Midterm Elections from the HuffPost Pollster API.

0.3 Load Python modules

```
In [1]: # special IPython command to prepare the notebook for matplotlib
        %matplotlib inline

        import requests
        from io import StringIO
        import numpy as np
        import pandas as pd # pandas
        import matplotlib.pyplot as plt # module for plotting
        import datetime as dt # module for manipulating dates and times
        import numpy.linalg as lin # module for performing linear algebra operatio
```

0.4 Problem 2: Is there a pollster bias in presidential election polls?

Problem 2(a) The HuffPost Pollster contains many political polls. You can access these polls from individual races as a CSV but you can also access polls through the HuffPost Pollster API to access the data.

Read in the polls from the 2012 Presidential Election: Barack Obama vs Mitt Romney into a pandas DataFrame called `election`. For this problem, you may read in the polls for this race directly using the CSV file available from the HuffPost Pollster page.

```
In [14]: #your code here
```

Show the head of the `election` DataFrame.

```
In [15]: #your code here
```

How many polls were conducted in November? Define this number as M.

Hint: Subset the `election` DataFrame for only dates in the `Start Date` column that are in November 2012.

```
In [16]: #your code here
```

Answer:

What was the median of the number of observations in the November polls? Define this quantity as N.

```
In [17]: #your code here
```

Answer:

Problem 2(b) Using the median sample size N from Problem 1(a), simulate the results from a single poll: simulate the number of votes for Obama out of a sample size N where $p = 0.53$ is the percent of voters who are voting for Obama.

Hint: Use the binomial distribution with parameters N and $p = 0.53$.

```
In [18]: #your code here
```

Now, perform a Monte Carlo simulation to obtain the estimated percentage of Obama votes with a sample size N where N is the median sample size calculated in Problem 2(a). Let p=0.53 be the percent of voters are voting for Obama.

Hint: You will repeat the simulation above 1,000 times and plot the distribution of the estimated *percent* of Obama votes from a single poll. The results from the single poll you simulate is random variable and will be different every time you sample.

```
In [19]: #your code here
```

Plot the distribution of the estimated percentage of Obama votes from your single poll. What is the distribution of the estimated percentage of Obama votes?

```
In [20]: #your code here
```

Answer:

What is the standard error (SE) of the estimated percentage from the poll.

Hint: Remember the SE is the standard deviation (SD) of the distribution of a random variable.

```
In [21]: #your code here
```

Problem 2(c) Now suppose we run M polls where M is the number of polls that happened in November (calculated in Problem 2(a)). Run 1,000 simulations and compute the mean of the M polls for each simulation.

```
In [22]: #your code here
```

What is the distribution of the average of polls?

Hint: Show a plot.

```
In [23]: #your code here
```

Answer:

What is the standard error (SE) of the average of polls?

```
In [24]: #your code here
```

Answer:

Is the SE of the average of polls larger, the same, or smaller than that the SD of a single poll (calculated in Problem 2(b))? By how much?

Hint: Compute a ratio of the two quantities.

```
In [25]: #your code here
```

Answer:

Problem 2(d) Repeat Problem 2(c) but now record the *across poll* standard deviation in each simulation.

```
In [26]: #your code here
```

What is the distribution of the *across M polls* standard deviation?

Hint: Show a plot.

```
In [27]: #your code here
```

Answer:

Problem 2(e) What is the standard deviation of M polls in our real (not simulated) 2012 presidential election data ?

```
In [28]: #your code here
```

Is this larger, the same, or smaller than what we expeced if polls were not biased.

```
In [29]: #your code here
```

Answer:

Problem 2(f) Learn about the normal approximation for the binomial distribution and derive the results of Problem 2(b) and 2(c) analytically (using this approximation). Compare the results obtained analytically to those obtained from simulations.

```
In [30]: #your code here
```

Answer:

0.5 Discussion for Problem 2

Write a brief discussion of your conclusions to the questions and tasks above in 100 words or less.

---

1 Submission Instructions

To submit your homework, create a folder named **lastname_firstinitial_hw#** and place your IPython notebooks, data files, and any other files in this folder. Your IPython Notebooks should be completely executed with the results visible in the notebook. We should not have to run any code. Make sure to share the private repo with my github account (mdog) and submit the repo path through blackboard.

```
In [ ]:
```