Homework 1

- ### Release Date: Sunday, February 2
- ### Due Date: Saturday, February 15 at 12:00PM

Introduction

We want to get a better understanding of surveys for collecting data. Polling agencies use surveys extensively in forecasting elections. Since the outcome of the 2016 US presidential election surprised many people, we want to determine possible bias in the data collection employed by polling agencies. Why did the predictions differ to such an extent from the outcome? Using randomization, we will attempt to understand these differences. By completing Homework 1, you should get...

- Practice with simple random samples to collect data from a population. How can the **volume** of data help with the predictions?
- Awareness of selection bias in surveys arising from collection of **nonrepresentative** datasets.
- Experience with plotting **histograms** to visualize occurences of numbers in ranges.
- Intuition about estimating **random numbers**. We want to enable you to study problems subject to uncertainty in rigorous and reproducible ways. So we will try to develop this intuition throughout this class for your future work in data science.

We will guide you through the problems step by step. However, we encourage you to discuss with us in Office Hours and on Piazza so that we can work together through these steps.

Submission Instructions

Submission of homework requires two steps. See **Homework 0** for more information.

Step 1

You are required to **submit your notebook on JupyterHub**. Please navigate to the `Assignments` tab to

- fetch
- modify
- validate
- submit

your notebook. Consult the instructional video (https://nbgrader.readthedocs.io/en/stable/user_guide/highlights.html#student-assignment-list-extension-for-jupyter-notebooks) for more information about JupyterHub.

Step 2

You are required to **submit a copy of your notebook to Gradescope**. Follow these steps

Formatting Instructions

1. Download as HTML (`File->Download As->HTML(.html)`).
2. Open the HTML in the browser. Print to .pdf
3. Upload to Gradescope. Consult the instructional video (https://www.gradescope.com/get_started#student-submission) for more information about Gradescope.
4. Map your responses on Gradescope. i.e. Tag your answer's page numbers to the appropriate question on Gradescope. See instructional video for more information.

Note that

- You should break long lines of code into multiple lines. Otherwise your code will extend out of view from the cell. Consider using `\` followed by a new line.

- For each textual response, please include relevant code that informed your response.
- For each plotting question, please include the code used to generate the plot. If your plot does not appear in the HTML / pdf output, then use `Image('name_of_file')` to embed it.
- You should not display large output cells such as all rows of a table. Instead convert the input cell from Code to Markdown back to Code to remove the output cell.

Important: Gradescope points will be awarded if and only if all the formatting instructions are followed.

Collaboration Policy

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you **write your solutions individually**. If you do discuss the assignments with others please **include their names** at the top of your solution.

Collaborators: *list names here*

Rubric

Question	Points
1.1	1
1.2	1
1.3	1
2.1	1
2.2	1
2.3	1
3.1	2
3.2	2
3.3	2
4	0 (Optional)
5.1	1
5.2	1
6	1
7.1	2
7.2	1
7.3	1
7.4	1
7.5	2
7.6	1
8.1	2
8.2	1
8.3	2
8.4	1
9.1	1
9.2	2

In [ ]:

```
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
plt.rcParams['figure.figsize'] = (16,8)
plt.rcParams['figure.dpi'] = 150
sns.set()

from IPython.display import display, Latex, Markdown, Image
```

Background

In Pennsylvania, 6,165,478 people voted in the election. Trump received 48.18% of the vote and Clinton recieved 47.46%. This doesn't add up to 100% because other candidates received votes. All together these other candidates received 100% - 48.18% - 47.46% = 4.36% of the vote.

Suppose we could select one person at random from the 6+ million voters in Pennsylvania . We are interested in the chance that we'd choose a Trump, Clinton, or Other voter.

Below is a probability table for the choice in Pennsylvania:

Voted for	Trump	Clinton	Other
Probability	0.4818	0.4746	0.0436
Number of people	2,970,733	2,926,441	268,304

Question 1

Suppose we take a simple random sample of $n = 1500$ voters from the 6+ million voters in Pennsylvania.

- What is the expected number of Trump voters?
- What is the expected number of Clinton voters?

To answer these questions, let T_1 be 1 if the first voter chosen for the sample voted for Trump and 0 if they voted for Clinton or another candidate. Let T_2 be 1 if the second voter chosen for the sample voted for Trump and 0 if they voted for Clinton or another candidate, and so on.

Let's start by finding:

$$P(T_{1000} = 1)$$
$$P(T_{1000} = 0)$$
$$\mathbb{E}(T_{17})$$

Part 1

$P(T_{1000} = 1)$:

In [ ]:

```
q1a_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q1a_answer <= 1
```

In [ ]:

```

```

Part 2

$P(T_{1000} = 0)$:

In [ ]:

```
q1b_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q1b_answer <= 1
```

In [ ]:

```

```

Part 3

$E(T_{17})$:

In [ ]:

```
q1c_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q1c_answer <= 1
```

In [ ]:

```

```

Question 2

For each voter chosen for the sample, let's keep track of whether they voted for Trump, Clinton, or another candidate.

We can do this with the triple (T_i, C_i, O_i), $i = 1, 2, \ldots, 1500$, where

- $T_i = 1$ if the ith voter sampled voted for Trump and 0 otherwise (this is the same random variable as in Question 1),
- $C_i = 1$ if the ith voter sampled voted for Clinton and 0 otherwise, and
- $O_i = 1$ if the ith voter sampled voted for another candidate and 0 otherwise.

Part 1

Find

$$T_{17} + C_{17} + O_{17} = ?$$

In [ ]:

```
q2a_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q2a_answer <= 1
```

In [ ]:

```

```

Part 2

Define $N_T = \sum_{i=1}^{1500} T_i$ and $N_C = \sum_{i=1}^{1500} C_i$, and $N_O = \sum_{i=1}^{1500} O_i$

Notice that because they are sums of random variables, N_T and N_C and N_O are random variables, too.

Find the expected value of N_T, i.e.,

$$\mathbb{E}(N_T)$$

In other words, find the expected number of Trump voters in our simple random sample of 1500 voters from Pennsylvania.

In [ ]:

```
q2b_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q2b_answer <= 4500
```

In [ ]:

```

```

Part 3

Find

$$N_T + N_C + N_O = ?$$

In [ ]:

```
q2c_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q2c_answer <= 4500
```

In [ ]:

```

```

Question 3

Given our population of Pennsylvania voters, every possible SRS has a certain well defined probability of occurring. For example, if we collected a SRS of only 2 voters without replacement (instead of 1500), the chance of each SRS is given in the probability distribution table below:

N_T	N_C	N_O	p
0	0	2	0.00189373515
0	1	1	0.04131080160
1	0	1	0.04193604504
0	2	0	0.22529213625
1	1	0	0.45740422268
2	0	0	0.23216338697

As an exercise in probability, we will have you compute similar probabilities for a simple random sample of 4 people without replacement.

Part 1

Find the following probability.

$$P(N_T = 4, N_C = 0, N_O = 0)$$

Hint: It is a product of four fractions, each of which has a distinct numerator and denominator.

In [ ]:

```
q3a_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q3a_answer <= 1
```

In [ ]:

```

```

Part 2

The answer from part 1 was a bit unwieldy.

We can simplify the problem by assuming that the draws are *with replacement*. That is, once a draw is picked as a sample, it can be picked again in the future. In this case, what is the following probability:

$$P(N_T = 4, N_C = 0, N_O = 0)$$

In [ ]:

```
q3b_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q3b_answer <= 1
```

In [ ]:

```

```

Part 3

Under this same simplfying assumption (that we sample with replacement), find the following probabilities:

$$P(N_T = 2, N_C = 2, N_O = 0)$$

In [ ]:

```
q3c1_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q3c1_answer <= 1
```

In [ ]:

```

```

$$P(N_T = 2, N_C = 1, N_O = 1)$$

In [ ]:

```
q3c2_answer = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 <= q3c2_answer <= 1
```

In [ ]:

```

```

Question 4 (Optional)

Can you generalize the above probability calculation and express the probability in terms of a probability mass function?

To set up the problem, let random variables N_1, N_2, N_3 be the number of Trump, Clinton, and Other voters selected, respectively.

Let p_1, p_2, p_3 be the chance of a Trump, Clinton, Other voter being chosen, respectively, and let n be the size of the sample drawn (with replacement).

In general, what is

$$P(N_1 = k_1, N_2 = k_2, N_3 = k_3) = ?$$

Hint: The answer involves $n!, k_1!, k_2!, k_3!$ and p_1, p_2, p_3 raised to various powers. Also note this may be a particularly tough problem depending on your math background. Take your time, and please discuss with fellow students and instructors. You don't need to get this problem right to complete the questions later in this homework.

YOUR ANSWER HERE

Election Polling

Political polling is a type of public opinion polling that can at best represent a snapshot of public opinion at the particular moment in time. Voter opinion shifts from week to week, even day to day, as candidates battle it out on the campaign field.

Polls usually start with a "horse-race" question, where respondents are asked whom they would vote for in a head-to-head race if the election were tomorrow: Candidate A or Candidate B. The survey begins with this question so that the respondent is not influenced by any of the other questions asked in the survey. Some of these other questions are asked to help assess how likely is it that the respondent will vote. Other questions are asked about age, education, and sex in order to adjust the findings if one group appears overly represented in the sample.

Pollsters typically use [random digit dialing (https://en.wikipedia.org/wiki/Random_digit_dialing)](https://en.wikipedia.org/wiki/Random_digit_dialing) to contact people.

Question 5

Part 1

If we're trying to predict the results of the Clinton vs. Trump presidential race, what is the population of interest?

YOUR ANSWER HERE

Part 2

What is the sampling frame?

YOUR ANSWER HERE

How might the sampling frame differ from the population?

After the fact, many experts have studied the 2016 election results. For example, according to the American Association for Public Opinion Research (AAPOR), predictions made before the election were flawed for three key reasons:

1. voters changed their preferences a few days before the election
2. those sampled were not representative of the voting population, e.g., some said that there was an overrepresentation of college graduates in some poll samples
3. voters kept their support for Trump to themselves (hidden from the pollsters)

In the next two problems on this homework, we will do two things:

- HW Question 7: We will carry out a study of the sampling error when there is no bias. In other words, we will try to compute the chance that we get the election result wrong even if we collect our sample in a manner that is completely correct. In this case, any failure of our prediction is due entirely to random chance.
- HW Question 8: We will carry out a study of the sampling error when there is bias of the second type from the list above. In other words, we will try to compute the chance that we get the election result wrong if we have a small systematic bias. In this case, any failure of our prediction is due to a combination of random chance and our bias.

Question 6

Why can't we assess the impact of the other two biases (voters changing preference and voters hiding their preference)?

Note: You might find it easier to complete this question after you've completed the rest of the homework including the simulation study.

YOUR ANSWER HERE

How large was the sampling error?

In some states the race was very close, and it may have been simply sampling error, i.e., random chance that the majority of the voters chosen for the sample voted for Clinton.

One year after the 2016 election, Nate Silver wrote in *The Media Has A Probability Problem* that the "media's demand for certainty -- and its lack of statistical rigor -- is a bad match for our complex world." FiveThirtyEight forecasted that Clinton had about a 70 percent chance of winning.

A 2- or 3-point polling error in Trump's favor (typical error historically) would likely be enough to tip the Electoral College to him.

We will first carry out a simulation study to assess the impact of the sampling error on the predictions.

The Electoral College

The US president is chosen by the Electoral College, not by the popular vote. Each state is alotted a certan number of electoral college votes, as a function of their population. Whomever wins in the state gets all of the electoral college votes for that state.

There are 538 electoral college votes (hence the name of the Nate Silver's site, FiveThirtyEight).

Pollsters correctly predicted the election outcome in 46 of the 50 states. For these 46 states Trump received 231 and Clinton received 232 electoral college votes.

The remaining 4 states accounted for a total of 75 votes, and whichever candidate received the majority of the electoral college votes in these states would win the election.

These states were Florida, Michigan, Pennsylvania, and Wisconsin.

State	Electoral College Votes
Florida	29
Michigan	16
Pennsylvania	20
Wisconsin	10

For Donald Trump to win the election, he had to win either:

- Florida + one (or more) other states
- Michigan, Pennsylvania, and Wisconsin

The electoral margins were very narrow in these four states, as seen below:

State	Trump	Clinton	Total Voters
Florida	49.02	47.82	9,419,886
Michigan	47.50	47.27	4,799,284
Pennsylvania	48.18	47.46	6,165,478
Wisconsin	47.22	46.45	2,976,150

Those narrow electoral margins can make it hard to predict the outcome given the sample sizes that the polls used.

Simulation Study of the Sampling Error

Now that we know how people actually voted, we can carry out a simulation study that imitates the polling.

Our ultimate goal in this problem is to understand the chance that we will incorrectly call the election for Hillary Clinton even if our sample was collected with absolutely no bias.

Question 7

Part 1

For your convenience, the results of the vote in the four pivotal states is repeated below:

State	Trump	Clinton	Total Voters
Florida	49.02	47.82	9,419,886
Michigan	47.50	47.27	4,799,284
Pennsylvania	48.18	47.46	6,165,478
Wisconsin	47.22	46.45	2,976,150

Using the table above, write a function `draw_state_sample(N, state)` that returns a sample with replacement of N voters from the given state. Your result should be returned as a list, where the first element is the number of Trump votes, the second element is the number of Clinton votes, and the third is the number of Other votes. For example, `draw_state_sample(1500, "florida")` could return `[727, 692, 81]`. You may assume that the state name is given in all lower case: `florida`, `pennsylvania`, `michigan`, `wisconsin`.

You might find `np.random.multinomial` useful.

In [ ]:

```
def draw_state_sample(N, state):
    # YOUR CODE HERE
    raise NotImplementedError()
```

In [ ]:

```
# TEST
len(draw_state_sample(1500, "florida")) == 3
```

In [ ]:

```
# TEST
sum(draw_state_sample(1500, "michigan")) == 1500
```

In [ ]:

```

```

Part 2

Now, create a function `trump_advantage` that takes in a sample of votes (like the one returned by `draw_state_sample`) and returns the difference in the proportion of votes between Trump and Clinton. For example `trump_advantage([100, 60, 40])` would return `0.2`, since Trump had 50% of the votes in this sample and Clinton had 30%.

In [ ]:

```
def trump_advantage(voter_sample):
    # YOUR CODE HERE
    raise NotImplementedError()
```

In [ ]:

```
# TEST
-1 < trump_advantage(draw_state_sample(1500, "wisconsin")) < 1
```

In [ ]:

```
# TEST
np.isclose(trump_advantage([100, 60, 40]), 0.2)
```

In [ ]:

```

```

Part 3

Simulate Trump's advantage across 100,000 simple random samples of 1500 voters for the state of Pennsylvania and store the results of each simulation in a list called `pa_simulations`.

That is, `pa_simulations[i]` should be Trump's percentage advantage for the `i` th simple random sample.

In [ ]:

```
pa_simulations = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
len(pa_simulations) == 100000
```

In [ ]:

```
# TEST
sum([-1 < x < 1 for x in pa_simulations]) == len(pa_simulations)
```

In [ ]:

```

```

Part 4

Make a histogram of the sampling distribution of Trump's percentage advantage in Pennsylvania. Make sure to give your plot a title and add labels where appropriate. Hint: You should use the `plt.hist` (https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.hist.html) function in your code.

Make sure to include a title as well as axis labels. You can do this using `plt.title`, `plt.xlabel`, and `plt.ylabel`.

In [ ]:

```
Image('q7d.png',width=750)
```

In [ ]:

```
# YOUR CODE HERE
raise NotImplementedError()
```

Part 5

Now write a function `trump_wins(N)` that creates a sample of N voters for each of the four crucial states (Florida, Michigan, Pennsylvania, and Wisconsin) and returns 1 if Trump is predicted to win based on these samples and 0 if Trump is predicted to lose.

Recall that for Trump to win the election, he must either:

- Win the state of Florida and 1 or more other states
- Win Michigan, Pennsylvania, and Wisconsin

The output of `trump_advantage` needs to be greater than 0.

In [ ]:

```
def trump_wins(N):
    # YOUR CODE HERE
    raise NotImplementedError()
```

In [ ]:

```

```

Part 6

If we repeat 100,000 simulations of the election, i.e. we call `trump_wins(1500)` 100,000 times, what proportion of these simulations predict a Trump victory? Give your answer as `percent_trump`.

This number represents the percent chance that a given sample will correctly predict Trump's victory *even if the sample was collected with absoutely no bias*. Note that many people incorrectly assume that this number should be 1.

In [ ]:

```
percent_trump = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 < percent_trump < 1
```

In [ ]:

```

```

We have just studied the sampling error, and found how our predictions might look if there was no bias in our sampling process. Essentially, we assumed that the people surveyed didn't change their minds, didn't hide who they voted for, and were representative of those who voted on election day.

Simulation Study of Selection Bias

According to an article by Grotenhuis, Subramanian, Nieuwenhuis, Pelzer and Eisinga (https://blogs.lse.ac.uk/usappblog/2018/02/01/better-poll-sampling-would-have-cast-more-doubt-on-the-potential-for-hillary-clinton-to-win-the-2016-election/#Author (https://blogs.lse.ac.uk/usappblog/2018/02/01/better-poll-sampling-would-have-cast-more-doubt-on-the-potential-for-hillary-clinton-to-win-the-2016-election/#Author)):

"In a perfect world, polls sample from the population of voters, who would state their political preference perfectly clearly and then vote accordingly."

That's the simulation study that we just performed.

It's difficult to control for every source of selection bias. And, it's not possible to control for some of the other sources of bias.

Next we investigate the effect of small sampling bias on the polling results in these four battleground states.

Throughout this problem, we'll examine the impacts of a 0.5 percent bias in favor of Clinton in each state. Such a bias has been suggested because highly educated voters tend to be more willing to participate in polls.

Question 8

Throughout this problem, adjust the selection of voters so that there is a 0.5% bias in favor of Clinton in each of these states.

For example, in Pennsylvania Clinton received 47.46 percent of the votes and Trump 48.18 percent. Increase the population of Clinton voters to 47.46 + 0.5 percent and correspondingly decrease the percent of Trump voters.

Part 1

Simulate Trump's advantage across 100,000 simple random samples of 1500 voters for the state of Pennsylvania and store the results of each simulation in a list called `biased_pa_simulations` .

That is, `biased_pa_simulation[i]` should hold the result of the `i` th simulation.

That is, your answer to this problem should be just like your answer from Question 7 part 3, but now using samples that are biased as described above.

In [ ]:

```
def draw_biased_state_sample(N, state):
    # YOUR CODE HERE
    raise NotImplementedError()
```

In [ ]:

```
# TEST
len(draw_biased_state_sample(1000, "wisconsin")) == 3
```

In [ ]:

```
# TEST
sum(draw_biased_state_sample(1000, "michigan")) == 1000
```

In [ ]:

```

```

In [ ]:

```
biased_pa_simulations = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
len(biased_pa_simulations) == 100000
```

In [ ]:

```
# TEST
sum([-1 < x < 1 for x in biased_pa_simulations]) == len(biased_pa_simulations)
```

In [ ]:

```

```

Part 2

Make a histogram of the new sampling distribution of Trump's advantage now using these biased samples. That is, your histogram should be the same as in Q8.4, but now using the biased samples.

Make sure to give your plot a title and add labels where appropriate.

In [ ]:

```
Image('q8b.png', width=750)
```

In [ ]:

```
# YOUR CODE HERE
raise NotImplementedError()
```

Part 3

Compare the histogram you created in Q8.2 to that in Q7.4.

YOUR ANSWER HERE

Part 4

Now perform 100,000 simulations of all four states and return the proportion of these simulations that result in a Trump victory. This is the same fraction that you computed in Q7.6, but now using your biased samples.

Give your answer as `percent_trump_biased` .

This number represents the chance that a sample biased 0.5% in Hillary Clinton's favor will correctly predict Trump's victory. You should observe that the chance is signficantly lower than with an unbiased sample, i.e. your answer in Q7.6.

In [ ]:

```
def trump_wins_biased(N):
    # YOUR CODE HERE
    raise NotImplementedError()
```

In [ ]:

```
percent_trump_biased = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
0 < percent_trump_biased < 1
```

In [ ]:

```

```

Question 9

Would increasing the sample size have helped?

Part 1

Try a sample size of 5,000 and run 100,000 simulations of a sample with replacement. What proportion of the 100,000 times is Trump predicted to win the election in the unbiased setting? In the biased setting?

Give your answers as `high_sample_size_unbiased_percent_trump` and `high_sample_size_biased_percent_trump`.

In [ ]:

```
high_sample_size_unbiased_percent_trump = ...
high_sample_size_biased_percent_trump = ...

# YOUR CODE HERE
raise NotImplementedError()
```

In [ ]:

```
# TEST
np.abs(high_sample_size_unbiased_percent_trump - 0.829) <= 0.02
```

In [ ]:

```

```

Part 2

What do your observations from part 1 say about the impact of sample size on the sampling error and on the bias?

Extra question for those who are curious: Just for fun, you might find it interesting to see what happens with even larger sample sizes (> 5000 voters) for both the unbiased and biased cases. Can you get them up to 99% success with sufficient large samples? How many? Why or why not? If you do this, include your observations in your answer.

In [ ]:

```
# Feel free to use this cell for any scratch work (creating visualizations, examining d
ata, etc.)
```

YOUR ANSWER HERE

According to FiveThirtyEight: "... Polls of the November 2016 presidential election were about as accurate as polls of presidential elections have been on average since 1972." When the margin of victory may be relatively small as it was in 2016, why don't polling agencies simply gather significantly larger samples to bring this error close to zero?

Well, we have seen that the predictions are driven by the sample size and the size of the bias. A larger sample size has reduced the sampling error. Unfortunately, if there is bias, then the predictions are close to the biased estimate. If the bias pushes the prediction from one candidate (Trump) to another (Clinton), then we have a "surprise" upset.