Part 1 If we're trying to predict the results of the Clinton vs. Trump presidential race, what is the population of interest?

Individuals who are eligible to vote in the U.S.

Part 2 What is the sampling frame?

Individuals who have a phone (landline or cellular), and thus a phone number that can be generated through rnadom digit dialing. This includes both people who can and cannot vote.

0.0.1 Question 5

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.

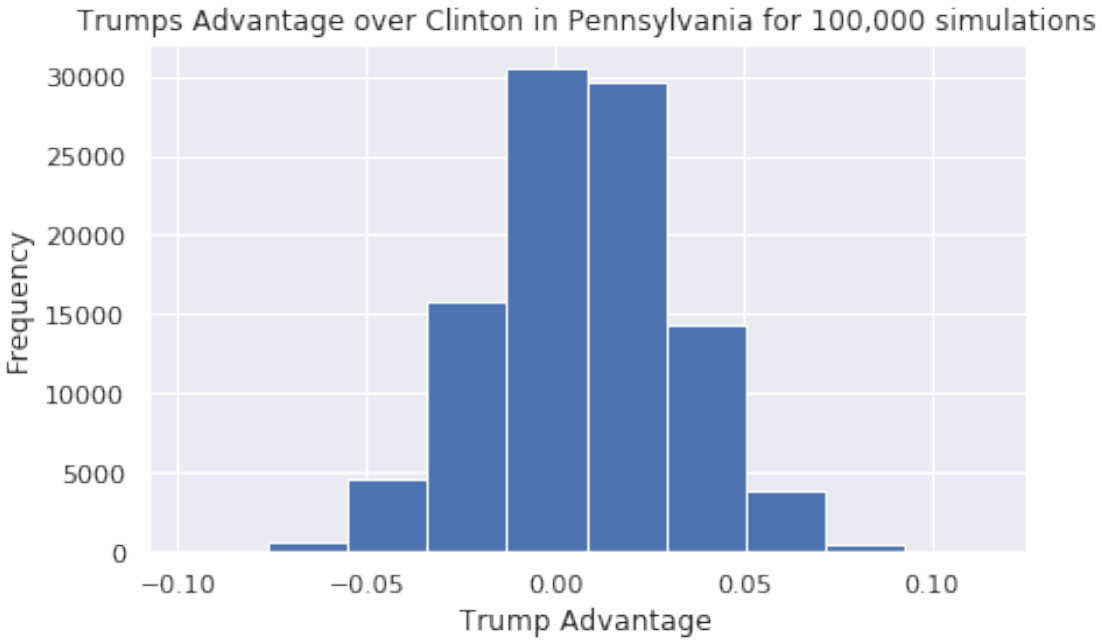
It is impossible to account for these biases in a simple random sample because a simple random samples only has information about the sample at the time the sample was taken and it assumes that the information collected is true. With the first case, the information collected from the sample is assumed to be true so if a participant were to hide their preference, we cannot account for that misinformation unless we were somehow made aware that this participant was hiding their preference. In the second case, the sample collected information on voters at a given time and we cannot control whether or not a specific voter decides to change their preference at the time of the elections – what is provided in the sample (and what is used for analysis) only represents the information provided at the time of the sample collection and we cannot account for the changes in preferences made after.

Part 4 Make a histogram of the sampling distribution of Trump's proportion advantage in Pennsylvania. Make sure to give your plot a title and add labels where appropriate. Hint: You should use the `plt.hist` 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 [38]: plt.hist(simulations)
         plt.title('Trumps Advantage over Clinton in Pennsylvania for 100,000 simulations')
         plt.xlabel('Trump Advantage')
         plt.ylabel('Frequency')
```

```
Out[38]: Text(0, 0.5, 'Frequency')
```

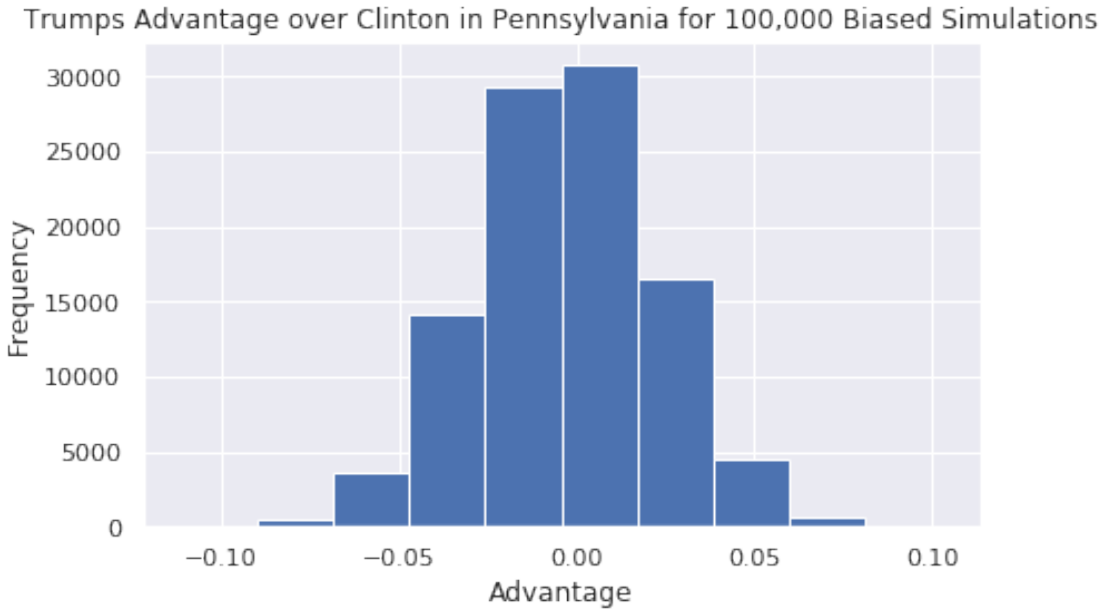


Part 2 Make a histogram of the new sampling distribution of Trump's proportion advantage now using these biased samples. That is, your histogram should be the same as in Q6.4, but now using the biased samples.

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

```
In [45]: plt.hist(biased_simulations)
         plt.title('Trumps Advantage over Clinton in Pennsylvania for 100,000 Biased Simulations')
         plt.xlabel('Advantage')
         plt.ylabel('Frequency')
```

```
Out[45]: Text(0, 0.5, 'Frequency')
```



Part 3 Compare the histogram you created in Q7.2 to that in Q6.4.

This graph that I created for Q7.2 looks like its data shifted to the left. For this graph, the weight of the graph seems to be balanced on a value below 0 whereas the graph for Q6.4 seems to be balanced on a value that is greater than 0. Just by looking at the tallest bins, one can tell that the weight of the graph for Q7.2 shifted to the left.

Write your answer in the cell below.

Increasing the sample size inceases the proportion of times Trump is predicted to win the election in the unbiased setting and decreases the proportion in the biased setting. I was able to get Trump's unbianced chance of winning up to 99% and his biased chance of winning as low as 27% when increasing the sample size to 100,000. With this information, it's fair to say that increasing the sample size increases the success proportion in the unbiased setting and decreases the proportion in the biased setting, thus reducing the sampling error.

0.0.2 Question 9

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?

Increasing the sample size in any case requires more money and more time. It is not only time consuming reaching out to more individuals using the random digit dialing method, but it also requires more money because more staff, who would be working longer hours, would be needed to contact more individuals. Both time and money is limited, especially when certain biases, such as changing voter preference, is affected by time.