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

The population of interest is all the poeple in the US who are eligible to vote.

Part 2 What is the sampling frame?

The sampling is people who own a phone number. This could be a personal cell phone, buisness phone, landline etc. This can also include those who are not eligible to vote because having a phone and being eligible to vote are independent events.

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.

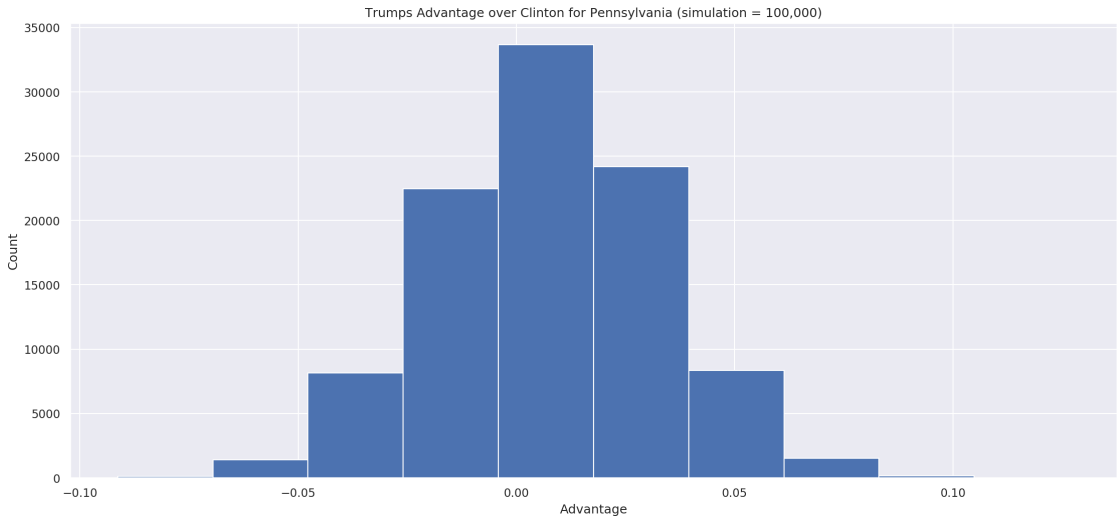
It is very hard to asses the impact of the other two biases since we cannot conduct a simple random sample (SRS). A SRS assumes that the information providede by those in the sample cannot be accounted for unless it was something we were aware of before. Also voters could change their mind the day before the election and their new preference would not be accounted for. They were already recorded for the survey thus their new answer cannot be accounted for in the sample thus providing inaccurate data.

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 [49]: plt.hist(simulations)
         plt.title('Trumps Advantage over Clinton for Pennsylvania (simulation = 100,000)')
         plt.xlabel('Advantage')
         plt.ylabel('Count')
```

```
Out[49]: Text(0, 0.5, 'Count')
```

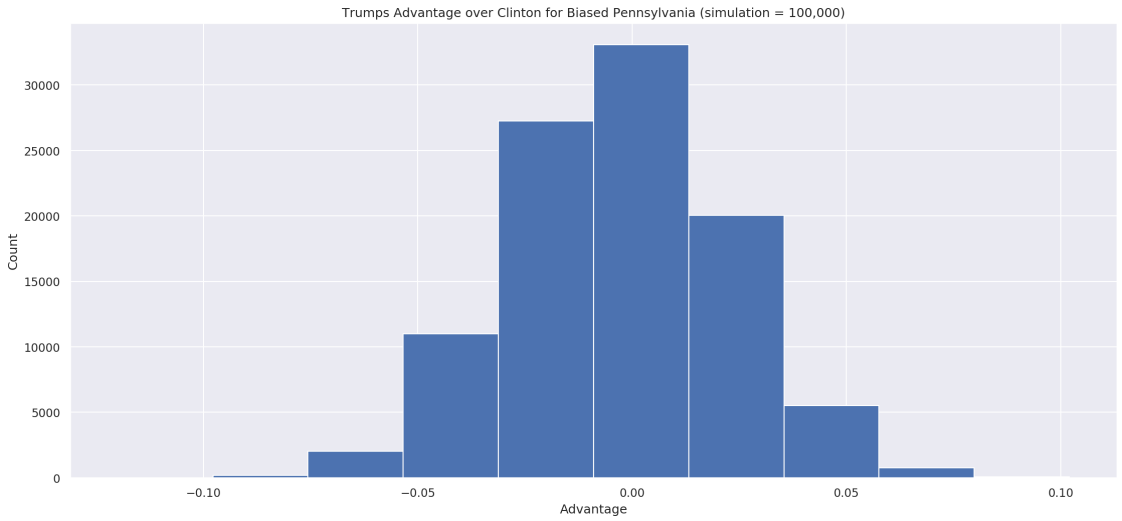


Part 2 Make a histogram of the new sampling distribution of Trump's proportion advantage now using these biased samples. That is, your histogram should be the same as in Q6.4, but now using the biased samples.

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

```
In [56]: plt.hist(biased_simulations)
         plt.title('Trumps Advantage over Clinton for Biased Pennsylvania (simulation = 100,000)')
         plt.xlabel('Advantage')
         plt.ylabel('Count')
```

```
Out[56]: Text(0, 0.5, 'Count')
```



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

Q7.2 shows the graph shifted more to the right of 0.00, Q6.4 shows the graph shifted more to the left of 0.00

Write your answer in the cell below.

The increased sample size from 5,000 to 100,000 increased the predictions for the unbiased results for trump to win. The increased sample size also decreased the predictions for the biased results for trump to win. If we use a sample size of 100,000 we can see that the output is roughly 99.99% for trump to win the unbiased vote and the biased result of 41.36%. Therefore we can conclude from these results that increasing the sample size helps to reduce 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?

One thing we have seen (especially the census) is that these types of polls on the grand scale are very costly. In lecture professor mentioned that going out and finding ever single person in the US is very costly which is why we do not choose that method and only track the census every 5-10 years. Therefore polling agencies would not be able to handle the scale of increasing the sample size significantly probably due to a money constraint.