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

Population of interest are the US citizens who have the ability to vote and are above the age of 18 before the 2016 elections and will be voting in the election.

Part 2 What is the sampling frame?

Sampling frame consists of the individuals and households possesing a telephone number vote and planning on voting in 2016 elections.

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.

Voters hiding their preference or choosing not to participate in the survey contributes to the non-response bias while those changing their preference contribute to the response bias. However, the vote is only taken once before the Election day and it is not possible to survey everyone once they have voted because they would be less likely to participate in another survey.

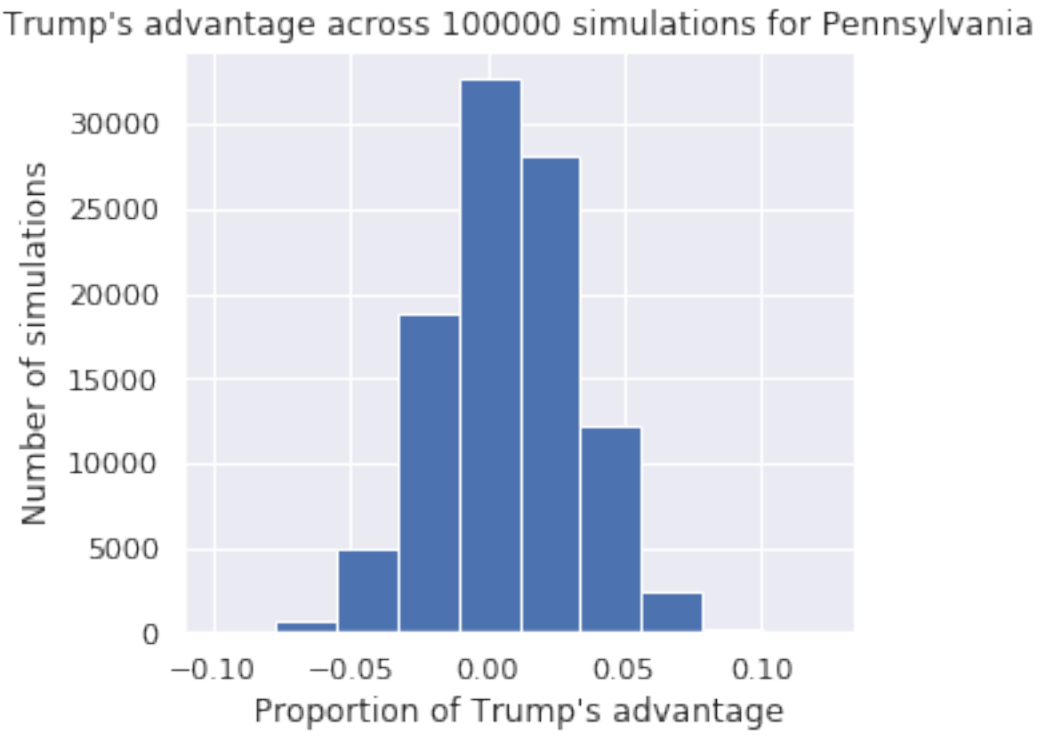
Part 4 Make a histogram of the sampling distribution of Trump's proportion advantage in Pennsylvania. Make sure to give your plot a title and add labels where appropriate. Hint: You should use the `plt.hist` function in your code.

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]: %matplotlib inline
```

```
In [58]: plt.figure(figsize=(4,4))
         plt.title("Trump's advantage across 100000 simulations for Pennsylvania")
         plt.xlabel("Proportion of Trump's advantage")
         plt.ylabel("Number of simulations")
         plt.hist(simulations)
```

```
Out[58]: (array([4.8000e+01, 6.4000e+02, 4.9070e+03, 1.8810e+04, 3.2666e+04,
                 2.8130e+04, 1.2200e+04, 2.3590e+03, 2.2500e+02, 1.5000e+01]),
          array([-0.09933333, -0.07713333, -0.05493333, -0.03273333, -0.01053333,
                  0.01166667,  0.03386667,  0.05606667,  0.07826667,  0.10046667,
                  0.12266667]),
          <a list of 10 Patch objects>)
```

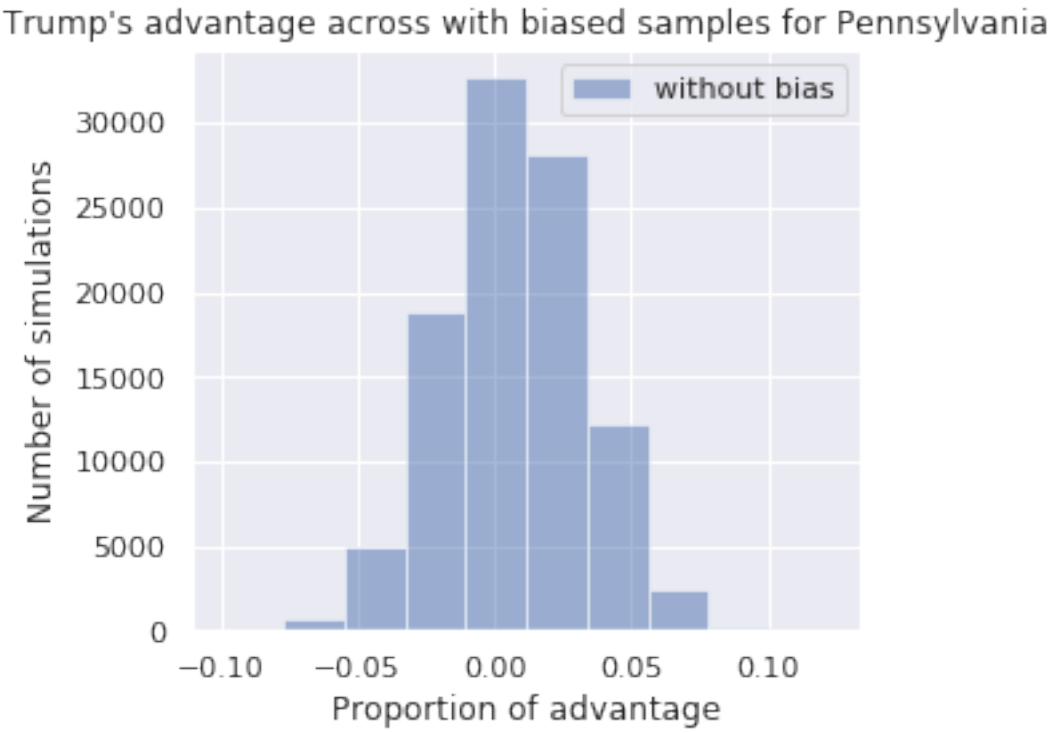


Part 2 Make a histogram of the new sampling distribution of Trump's proportion advantage now using these biased samples. That is, your histogram should be the same as in Q6.4, but now using the biased samples.

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

```
In [46]: plt.figure(figsize=(4,4))
         plt.title("Trump's advantage across with biased samples for Pennsylvania")
         plt.xlabel("Proportion of advantage")
         plt.ylabel("Number of simulations")
         plt.hist(simulations, alpha=0.5, label="without bias")
         plt.legend(loc="upper right")
```

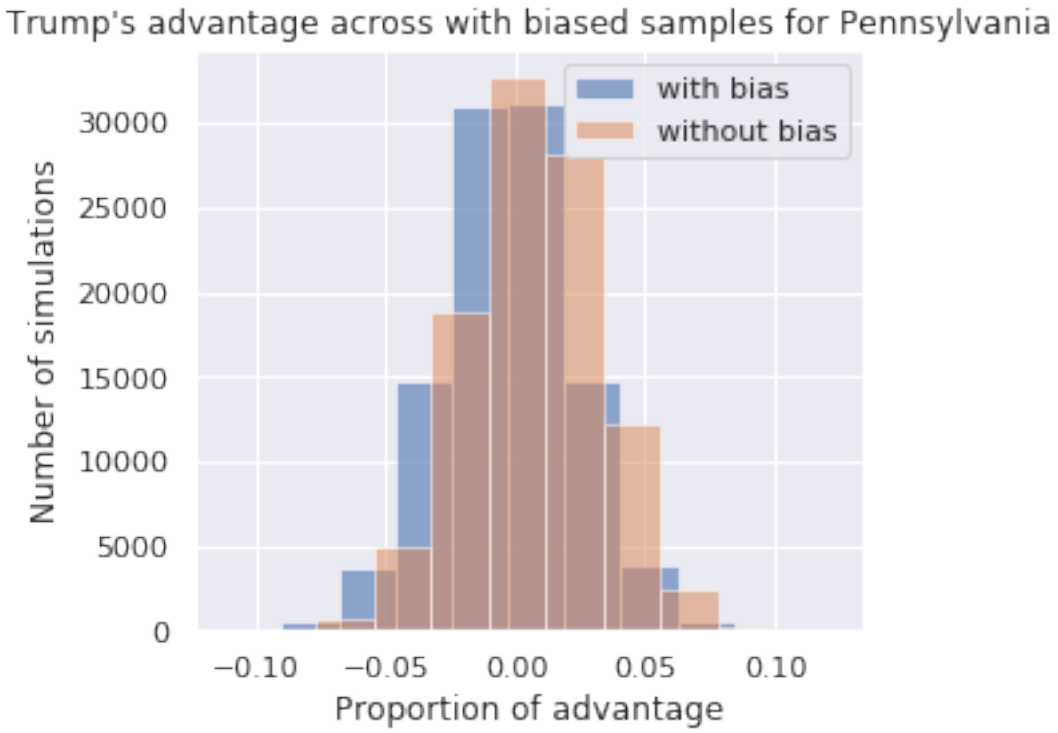
```
Out[46]: <matplotlib.legend.Legend at 0x7f72b55d89d0>
```




```
In [47]: #just for comparison :)
         plt.figure(figsize=(4,4))
         plt.title("Trump's advantage across with biased samples for Pennsylvania")
```

```
plt.xlabel("Proportion of advantage")
plt.ylabel("Number of simulations")
plt.hist(biased_simulations, alpha=0.6, label="with bias")
plt.hist(simulations, alpha=0.5, label="without bias")
plt.legend(loc="upper right")
```

Out[47]: <matplotlib.legend.Legend at 0x7f72b52d3d00>



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

The histogram for the biased samples is a smoother curve compared to the one for unbiased samples. It is also shifted to the left compared to the histogram for unbiased samples. This shows that the histogram is not well-balanced and underestimates the parameter, compared to the histogram for the unbiased data.

Write your answer in the cell below.

As the sample size increased, the gap between the proportions for the biased and unbiased samples increased thus the bias is magnified as the sample size increases. Increasing the size of the sample also reduces the sampling error because a larger sample would make the sample statistic less variable since a larger sample would better represent the population.

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?

While the bias is made more obvious by increasing the sample size, it cannot be corrected by increasing the sample size because it is influenced by other factors such as undercoverage and non-response from the individuals being surveyed. This error lies in the method of collection and perhaps polling agencies should have tried a different survey method.