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 is who we want to study. Our population of interest is voters of the United States.

Part 2 What is the sampling frame?

The sampling frame is who is studied in our poll. We are studying people who have phone numbers.

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.

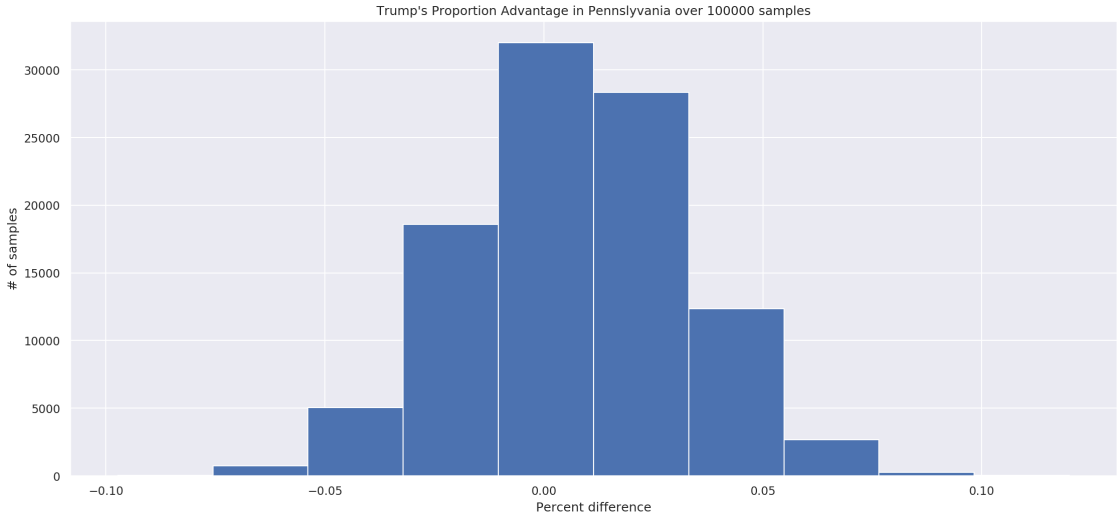
These are cases of response bias, in which our sample has told us some information but has actually voted in some other direction. In these cases, it is very difficult to measure how much of our information is biased. This makes it very difficult to measure and thus assess the impact of the bias. In Q7 we assumed that there was a misrepresentation and changed the data slightly to represent this. To an extenct we can measure how much our population may represent the true population as we are keeping track of name, race, gender and other polling information. But when it comes to response bias, we cannot measure the difference between our polls and how people actually feel.

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 [65]: plt.hist(simulations)
         plt.title("Trump's Proportion Advantage in Pennslyvania over 100000 samples")
         plt.xlabel("Percent difference")
         plt.ylabel("# of samples")
```

```
Out[65]: Text(0, 0.5, '# of samples')
```

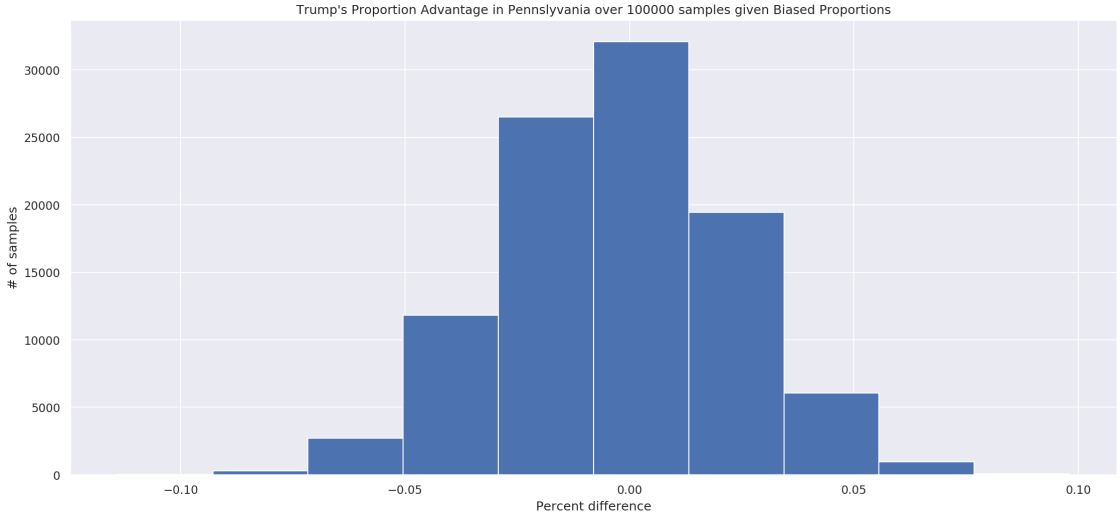


Part 2 Make a histogram of the new sampling distribution of Trump's proportion advantage now using these biased samples. That is, your histogram should be the same as in Q6.4, but now using the biased samples.

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

```
In [72]: plt.hist(biased_simulations)
         plt.title("Trump's Proportion Advantage in Pennslyvania over 100000 samples given Biased Propo
         plt.xlabel("Percent difference")
         plt.ylabel("# of samples")
```

```
Out[72]: Text(0, 0.5, '# of samples')
```



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

The histogram in Q6.4 has a mean centered slightly above 0.0 with a spread from around -0.075 to 0.10. The histogram in Q7.2 is centered slightly below 0.0 with a spread from -0.09 to 0.075. The spreads of both histograms are pretty similar, however the histogram in Q7.2 is slightly shifted lower than the histrogram in Q6.4.

Write your answer in the cell below.

As we increase the size of the sample, the probability of wins start leaning closer to 1 and 0. We start to see less sampling error or bias as we have larger samples. Each simulation starts to vary less, converging to the true value of who will win or lose the election.

In the case of the unbiased trump wins, we went from a win rate of 0.69 to 0.82. For the biased proportions we squeezed it down from 0.46 to 0.44

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?

Primarily we see the win rate be driven by how large the sample is and the size of the bias associated with it. For a bias of size 0.5 we a see a change from 82% win rate to 44% win rate in our date. With larger data, our results wi;; be a lot more spread apart and thus be a reason for an unpredicted win from the other side.

Some other reasons why we may not get larger sizes is due to the cost and difficulty --

For polling it may be difficult to get a large number of samples. In the example we had used above, our polling technique involved generating random digits and then calling those numbers, polling people on the other end. However, the case with this technique is we end up with a lot of non response bias -- people who do not pikc up the calls or pick up the calls and refuse to answer the poll. This makes it difficult to get polls that represent a lot of the population.

Another reason why we dont gather significantly larger samples is the cost associated with getting large samples. In this scenario it may make sense to study more people as we see a significant jump from 0.69 to 0.82 when sampling 3500 more people. But as we sample more and more people, the amount of information we gain from each individual decreases, making it cost a lot of money to gain more information.