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 eligible voters who will vote in the election.

Part 2 What is the sampling frame?

It depends on the sampling methodology. For instance, using random digit dialling would result in a sampling frame of all people with phone numbers (which is a sampling frame that *overlaps* the population of interest but is not exclusively a subset of the population of interest).

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.

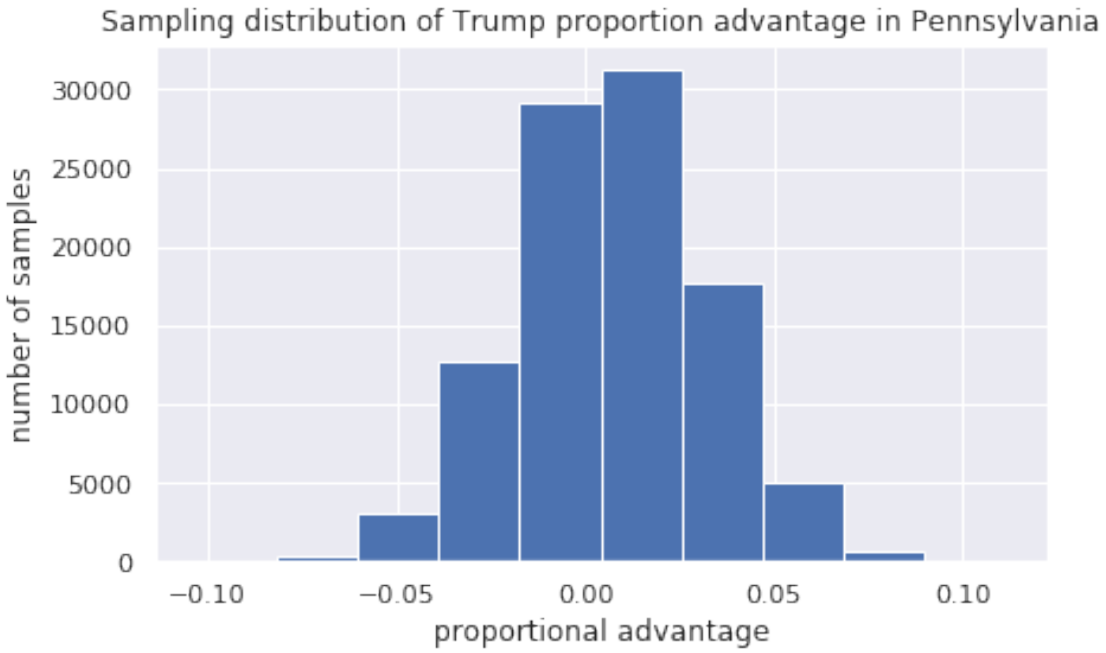
There's really no nice way to *quantify* the number of voters who changed their preferences without additional data. As a result, we can't do post-mortem analyses to control for this parameter with the information that we currently have. The same reasoning holds for the voters who hid their preferences; in addition, voters who hid their initial preferences are not liable to admit to such.

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("Sampling distribution of Trump proportion advantage in Pennsylvania")
         plt.xlabel("proportional advantage")
         plt.ylabel("number of samples")
```

```
Out[38]: Text(0, 0.5, 'number of samples')
```

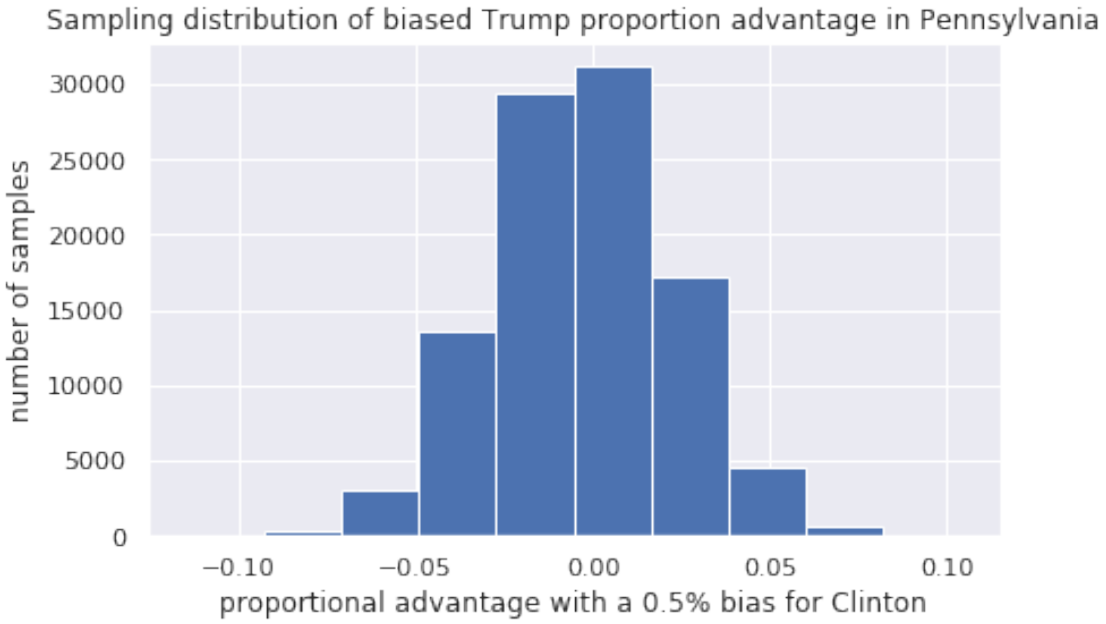


Part 2 Make a histogram of the new sampling distribution of Trump's proportion advantage now using these biased samples. That is, your histogram should be the same as in Q6.4, but now using the biased samples.

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

```
In [45]: plt.hist(biased_simulations)
         plt.title("Sampling distribution of biased Trump proportion advantage in Pennsylvania")
         plt.xlabel("proportional advantage with a 0.5% bias for Clinton")
         plt.ylabel("number of samples")
```

```
Out[45]: Text(0, 0.5, 'number of samples')
```



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

Purely looking at both histograms, it's clear that the biased dataset has a slight edge for Clinton over that of Trump. The cell below verifies this with the actual data:

```
In [46]: print("Mean:\t\t{}\nMedian:\t\t{}\nBiased mean:\t{}\nBiased median:\t{}"
               .format(np.mean(simulations), np.median(simulations),
                       np.mean(biased_simulations), np.median(biased_simulations)))
```

```
Mean:                0.007334473333333334
Median:                0.007333333333333333
Biased mean:        -0.0028152466666666663
Biased median:         -0.0026666666666666666
```

Write your answer in the cell below.

It's clear that simply increasing sample size does not actually "fix" biases inherent in the surveying methodology. The graph above demonstrates that on an unbiased dataset, the accuracy across sampling distributions strengthens as sample size increases. However, while increasing sample size of the biased sampling distribution did appear to shift the distribution, the distributions actually got *increasingly wrong* as sample size increased.

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?

There are a number of reasons preventing the usage of larger sample sizes.

1. A larger sample size would have been subject to the same response and non-response biases as the original sample. As a result, biases inherent in the survey and surveying method would manifest even in a far greater sample size (for instance, the 0.5% Clinton bias we introduced would exist in *any* SRS taken from a sampling frame, which was inherently biased).
2. A significantly larger sample size introduces cost overhead. To greatly expand the reach of a poll requires investing more in the poll, with diminishing returns.
3. Pollsters may not have realised that the margin of victory was as small as it was. If they had not realised this, there would have been no reason to readminster a survey to a larger sample size.