SOC245: Visualizing Data
Lecture 8: Uncertainity

Chris Felton and Vikram Ramaswamy

Freshman Scholars Institute
Princeton University

Aug 1, 2022

- On Friday, we analyzed the fraction of respondents from the ACS dataset that were below the poverty line.

- On Friday, we analyzed the fraction of respondents from the ACS dataset that were below the poverty line.
- But, we don't really care about *just* the respondents of this survey:

- On Friday, we analyzed the fraction of respondents from the ACS dataset that were below the poverty line.
- But, we don't really care about *just* the respondents of this survey:
- What can we say about the general population of the United States from this?

Where We've Been and Where We're Going

- Understanding central tendency and spread of a sample

Where We've Been and Where We're Going

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Where We've Been and Where We're Going

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Where We've Been and Where We're Going

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Going from the sample to the population: How confident are we about our estimates?

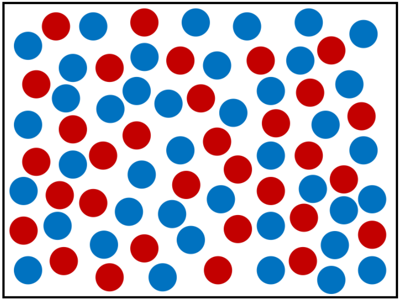
Outline

Outline

Motivating Question

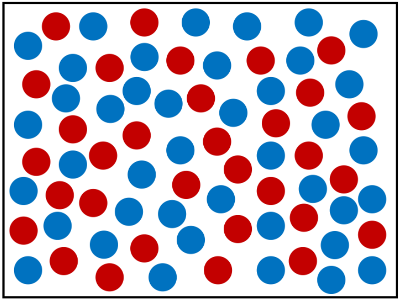
How certain can we be that the Democratic party candidate will win the popular vote?

Population distribution



Suppose this is our population, with blue representing a Democrat and red representing that a Republican.

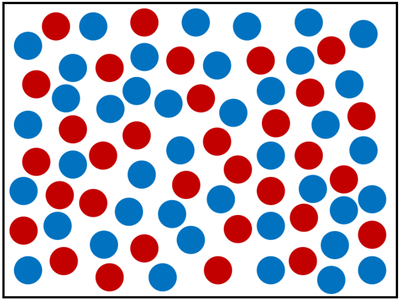
Population distribution



Suppose this is our population, with blue representing a Democrat and red representing that a Republican.

We can ask every single citizen (the total **population**) how they are going to vote ...

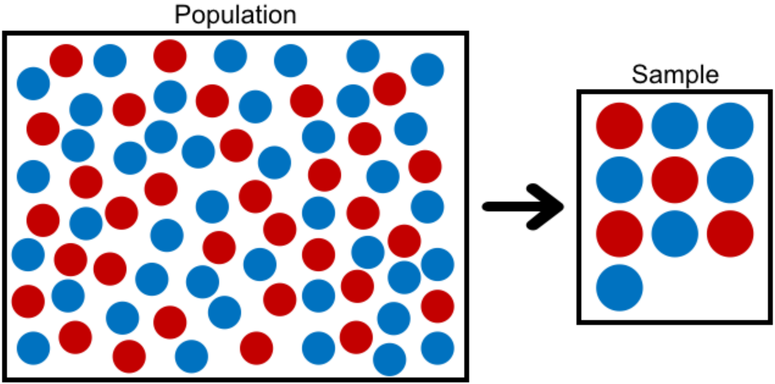
Population distribution



Suppose this is our population, with blue representing a Democrat and red representing that a Republican.
We can ask every single citizen (the total **population**) how they are going to vote . . .
But this is very expensive and time-consuming.

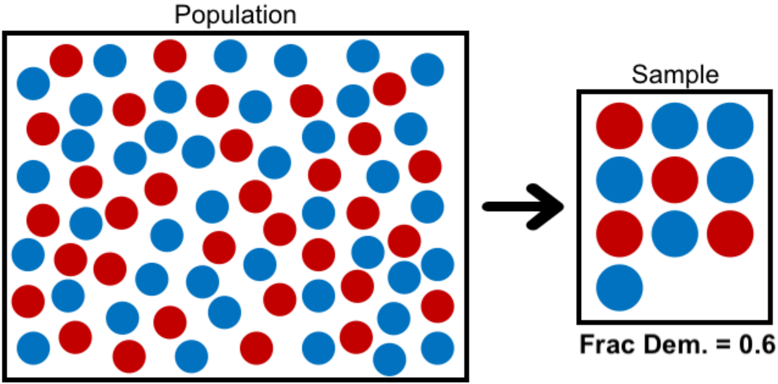
Getting an estimate for the population

We can ask a subset of the total population.



Getting an estimate for the population

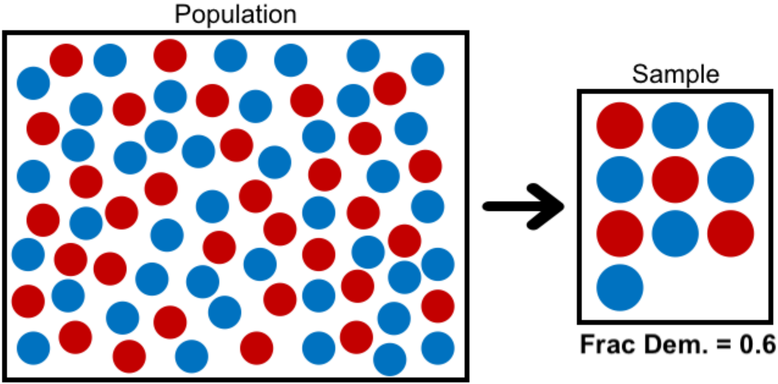
We can ask a subset of the total population.



Now, we can compute the fraction of Democrat votes within the sample.

Getting an estimate for the population

We can ask a subset of the total population.



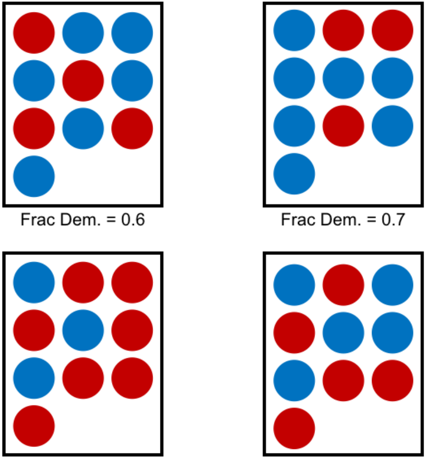
What can we say about the population from the sample?

Estimates depend on the sample

Note that our estimate of the vote might be different based on our sample:

Estimates depend on the sample

Note that our estimate of the vote might be different based on our sample:



Frac Dem. = 0.6

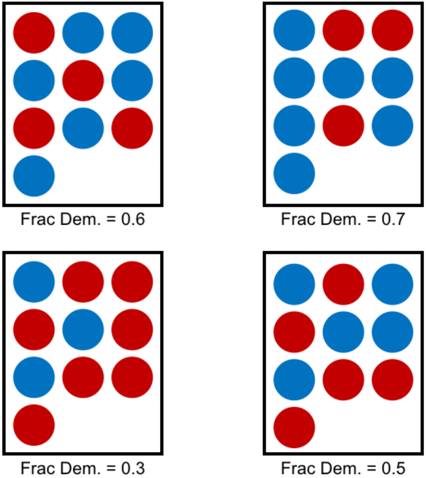
Frac Dem. = 0.7

Frac Dem. = 0.3

Frac Dem. = 0.5

Estimates depend on the sample

Note that our estimate of the vote might be different based on our sample:



But once we pick a sample, the estimate is fixed.

Sampling Distributions

- Suppose we repeat this sampling procedure 10000 times

Sampling Distributions

- Suppose we repeat this sampling procedure 10000 times
- We pick a sample of 10, compute and write down the fraction of votes that are Democrat and repeat.

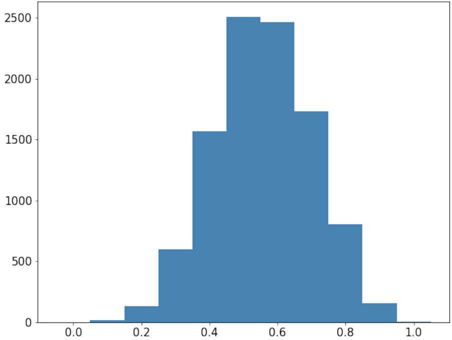
Sampling Distributions

- Suppose we repeat this sampling procedure 10000 times
- We pick a sample of 10, compute and write down the fraction of votes that are Democrat and repeat.
- We now have this set of values, and we can analyze the distribution.

Sampling Distributions

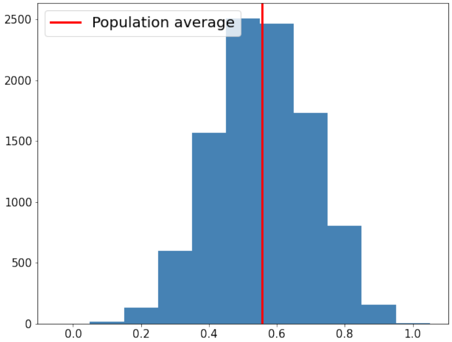
- Suppose we repeat this sampling procedure 10000 times
- We pick a sample of 10, compute and write down the fraction of votes that are Democrat and repeat.
- We now have this set of values, and we can analyze the distribution.
- This is called the **sampling distribution**

Visualizing the sampling distribution



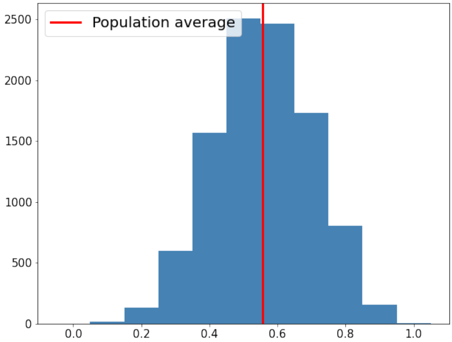
Visualizing the sampling distribution

Here is the true value for the fraction of Democrat votes (if we were able to compute this for the whole population).



Visualizing the sampling distribution

Here is the true value for the fraction of Democrat votes (if we were able to compute this for the whole population).



How can we reason about how close our estimate is to the actual?

Ideal goal

Given a statistic (i.e, a sample), what is the probability of (a parameter) of the population?

Ideal goal

Given a statistic (i.e, a sample), what is the probability of (a parameter) of the population?

In our case, given the fraction of Democrats within our sample, what is the probability the fraction of Democrats in the population is close to the sample fraction?

Ideal goal

Given a statistic (i.e, a sample), what is the probability of (a parameter) of the population?

In our case, given the fraction of Democrats within our sample, what is the probability the fraction of Democrats in the population is close to the sample fraction?

This is called **statistical inference**: using a sample to learn about the underlying population.

Problem

We don't know the full population and we only have access to a single sample (so we can't compute the sampling distribution.)

Outline

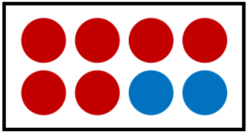
Estimating the sampling distribution

- Suppose we treat the sample (that we have access to) like the population distribution.

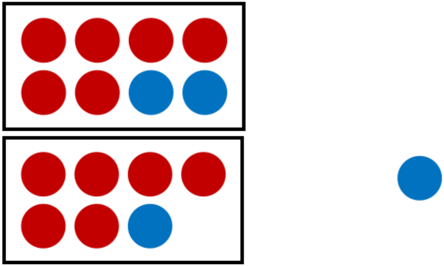
Estimating the sampling distribution

- Suppose we treat the sample (that we have access to) like the population distribution.
- Then, we can resample from this distribution with replacement, and compute the statistic on each resample.

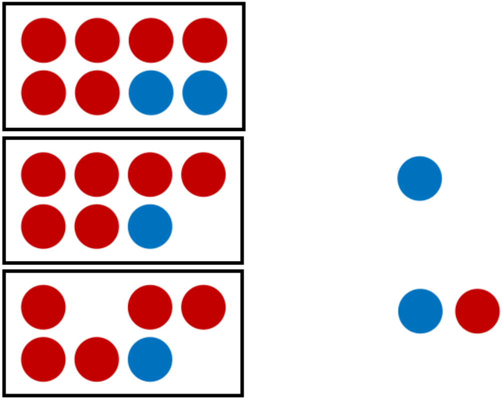
Sampling without replacement



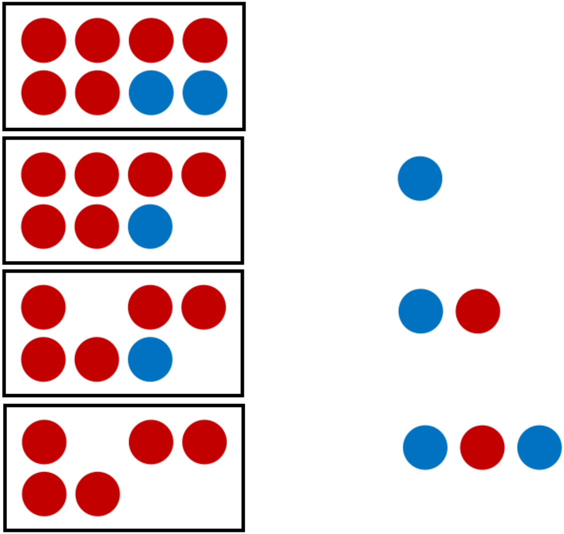
Sampling without replacement



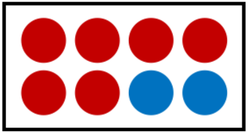
Sampling without replacement



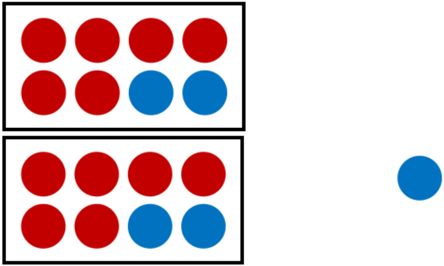
Sampling without replacement



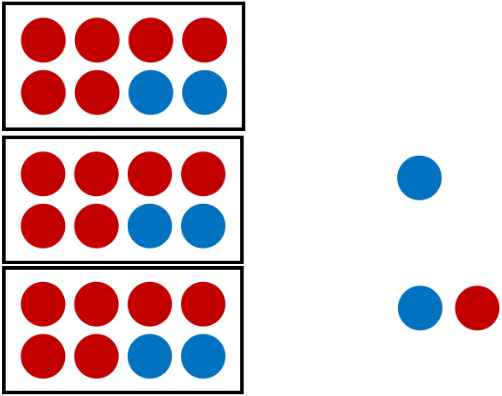
Sampling with replacement



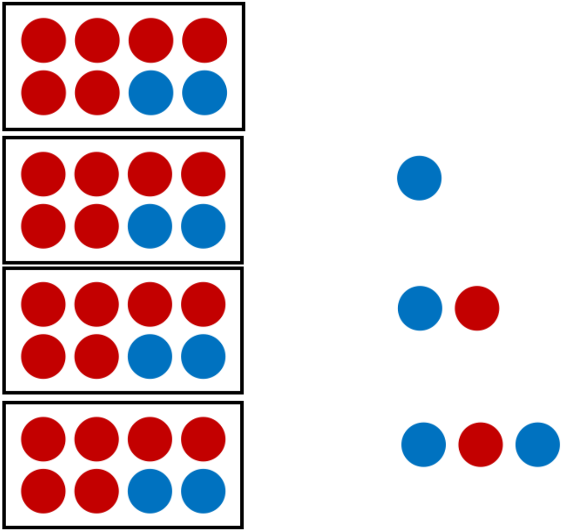
Sampling with replacement



Sampling with replacement

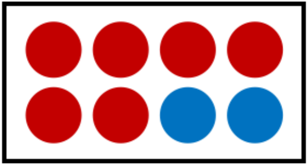


Sampling with replacement



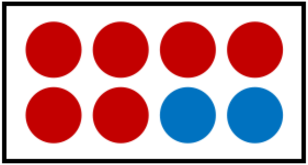
Why sampling with replacement

- Keeps the distribution the same across samples.



Why sampling with replacement

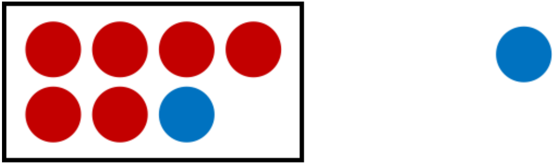
- Keeps the distribution the same across samples.



$\Pr(\text{blue}) = \frac{2}{8}$

Why sampling with replacement

- Keeps the distribution the same across samples.

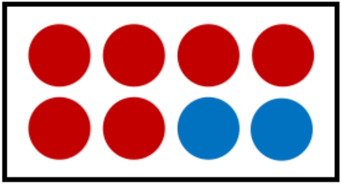


Original $\Pr(\text{blue}) = \frac{2}{8}$

New $\Pr(\text{blue}) = \frac{1}{7}$

Why sampling with replacement

- Keeps the distribution the same across samples.



Original $\Pr(\text{blue}) = \frac{2}{8}$

With replacement $\Pr(\text{blue}) = \frac{2}{8}$

Bootstrapping

- Suppose we treat the sample (that we have access to) like the population distribution.

Bootstrapping

- Suppose we treat the sample (that we have access to) like the population distribution.
- Then, we can resample from this distribution with replacement, and compute the statistic on each resample.

Bootstrapping

- Suppose we treat the sample (that we have access to) like the population distribution.
- Then, we can resample from this distribution with replacement, and compute the statistic on each resample.
- This approximates the shape and spread of the sampling distribution.

Bootstrapping

- Suppose we treat the sample (that we have access to) like the population distribution.
- Then, we can resample from this distribution with replacement, and compute the statistic on each resample.
- This approximates the shape and spread of the sampling distribution.
- We can construct an interval using percentiles of the resampled distribution.

Working through an example

Let's say I have a sample of 100 people with their party affiliations (i.e, whether they are going to vote Democrat or Republican). In this sample, 51% vote Democrat.

Working through an example

Let's say I have a sample of 100 people with their party affiliations (i.e, whether they are going to vote Democrat or Republican). In this sample, 51% vote Democrat.

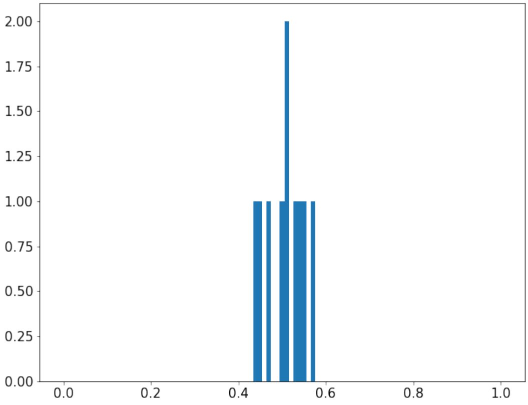
We can sample with replacement. In this resample, 45% vote Democrat.

Working through an example

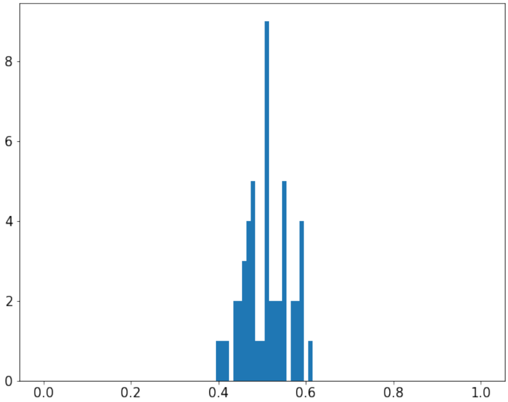
Let's say I have a sample of 100 people with their party affiliations (i.e, whether they are going to vote Democrat or Republican). In this sample, 51% vote Democrat.

We can sample with replacement. In this resample, 45% vote Democrat.

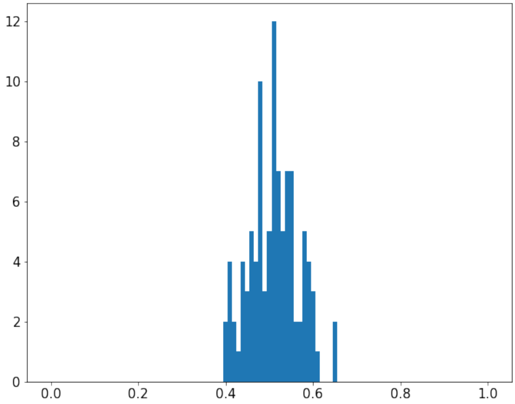
And again. This resample has 57% voting Democrat.



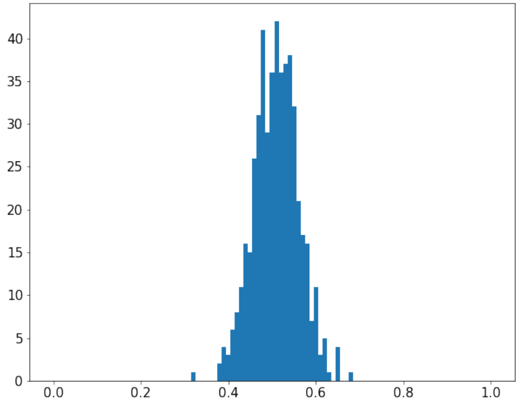
Distribution with 10 resamples.



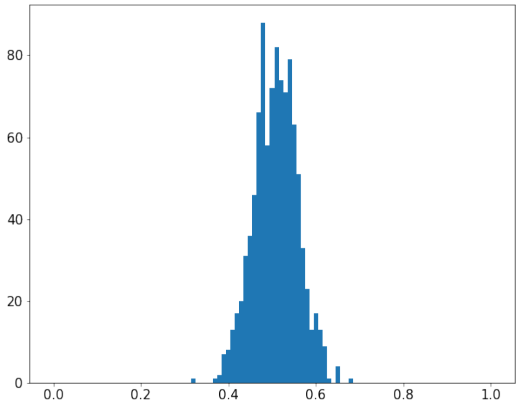
Distribution with 50 resamples.



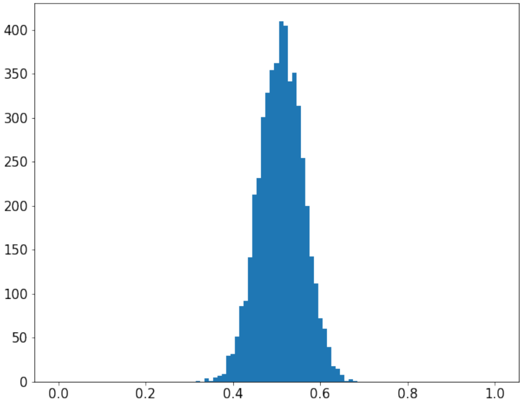
Distribution with 100 resamples.



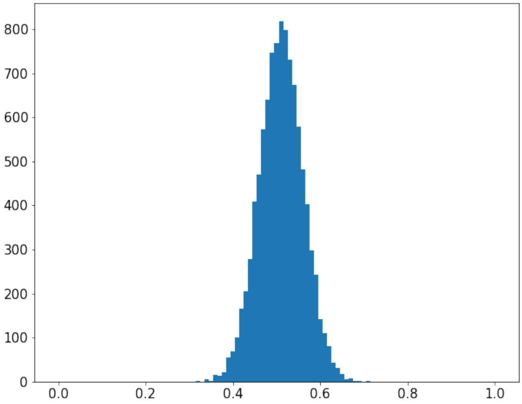
Distribution with 500 resamples.



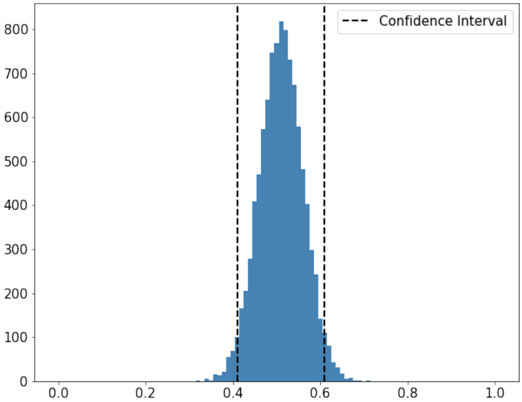
Distribution with 1000 resamples.



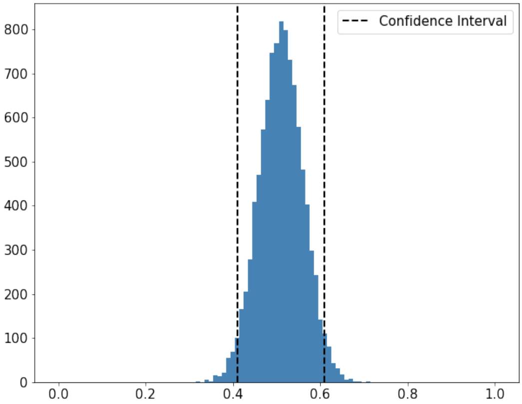
Distribution with 5000 resamples.



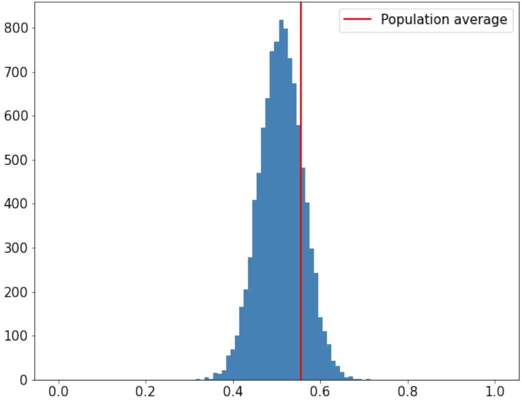
Distribution with 10000 resamples.



We can now compute the 2.5% and 97.5% percentiles.



We can now compute the 2.5% and 97.5% percentiles. This gives us a **confidence interval**: 95% of our values (fraction of Democrats) were between 41% and 61%.



Important: The mean of the resampled distribution does not always approximate the mean of the true population: just the shape and spread approximates that of the true sampling distribution.

Why does this work?

Intuitively,

- We make the assumption that each of the observations from our sample is drawn randomly from the population.

Why does this work?

Intuitively,

- We make the assumption that each of the observations from our sample is drawn randomly from the population.
- Thus, drawing observations from our sample randomly (with replacement) is similar to drawing from the population

Frequentist interpretation of a confidence interval

- From one sample of our population we have one confidence interval.

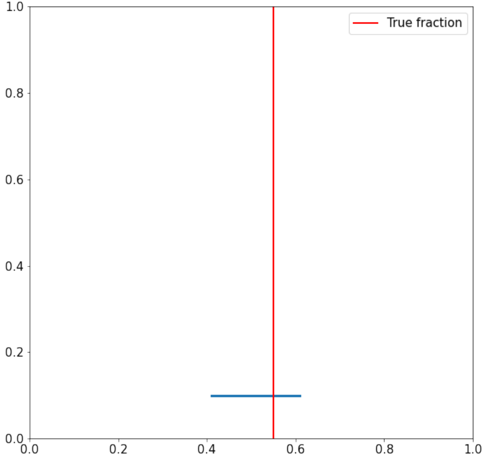
Frequentist interpretation of a confidence interval

- From one sample of our population we have one confidence interval.
- Suppose I can do this over and over again: that is, I can sample from the population, and then create another confidence interval

Frequentist interpretation of a confidence interval

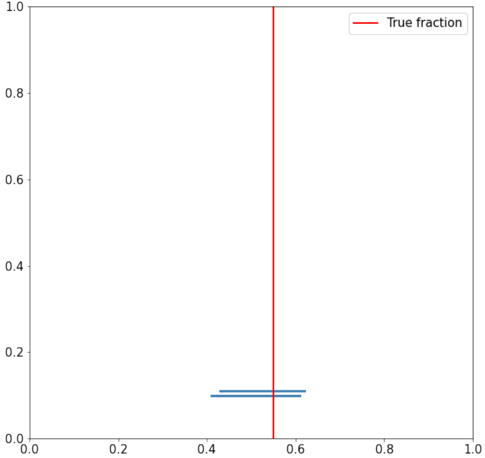
- From one sample of our population we have one confidence interval.
- Suppose I can do this over and over again: that is, I can sample from the population, and then create another confidence interval
- What would this look like?

Frequentist interpretation of a confidence interval



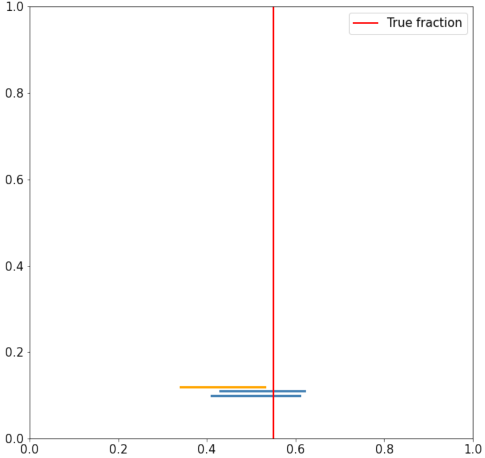
1 sample

Frequentist interpretation of a confidence interval



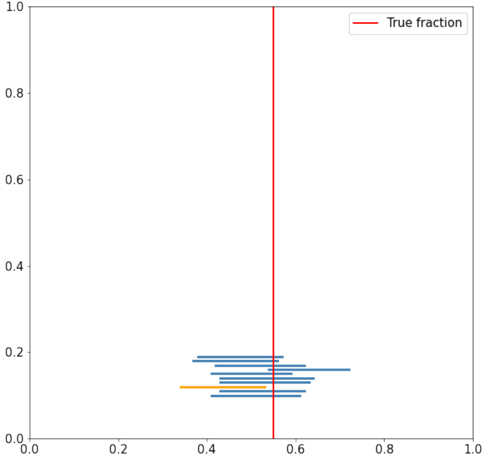
2 samples

Frequentist interpretation of a confidence interval



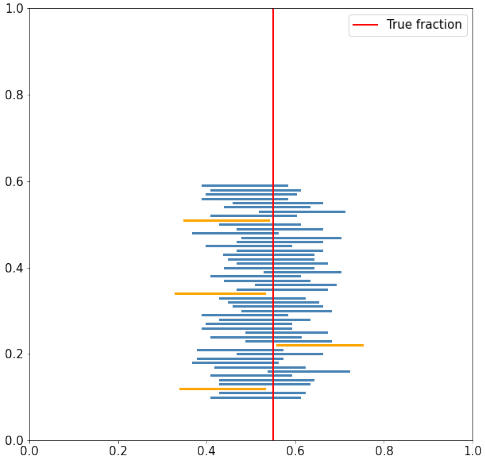
3 samples

Frequentist interpretation of a confidence interval



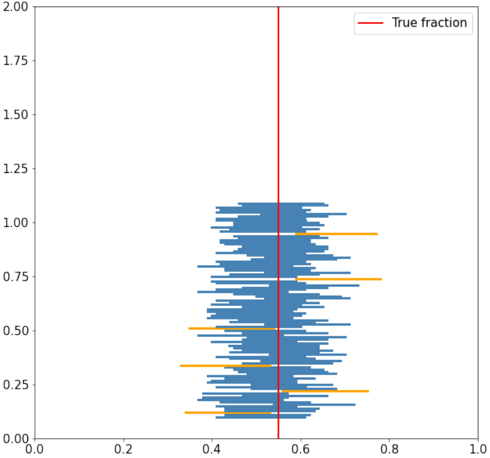
10 samples

Frequentist interpretation of a confidence interval



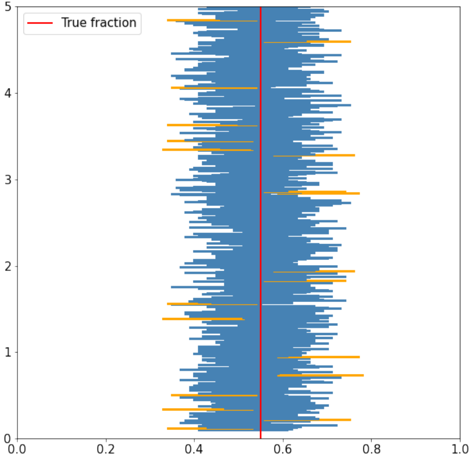
50 samples

Frequentist interpretation of a confidence interval



100 samples

Frequentist interpretation of a confidence interval



500 samples

Frequentist interpretation of a confidence interval

If we can compute enough confidence intervals, the true value lies within the confidence interval with a probability of 95%

Frequentist interpretation of a confidence interval

If we can compute enough confidence intervals, the true value lies within the confidence interval with a probability of 95%

In general, if we compute percentiles at $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$, the true value lies within the confidence interval with a probability of $1 - \alpha$.

Limitations of Bootstrapping

- This doesn't work too well when the sample size is small.

Limitations of Bootstrapping

- This doesn't work too well when the sample size is small.
  - Smaller sample size $\Rightarrow$ Harder to approximate true distribution

Limitations of Bootstrapping

- This doesn't work too well when the sample size is small.
  - Smaller sample size $\Rightarrow$ Harder to approximate true distribution
  - As an extreme consider a sample size of 1. Bootstrapping is not going to let me change the sample.

Limitations of Bootstrapping

- This doesn't work too well when the sample size is small.
  - Smaller sample size $\Rightarrow$ Harder to approximate true distribution
  - As an extreme consider a sample size of 1. Bootstrapping is not going to let me change the sample.
  - Note: This is still going to be an issue when computing confidence intervals in other ways.

Limitations of Bootstrapping

- This doesn't work too well when the sample size is small.
  - Smaller sample size $\Rightarrow$ Harder to approximate true distribution
  - As an extreme consider a sample size of 1. Bootstrapping is not going to let me change the sample.
  - Note: This is still going to be an issue when computing confidence intervals in other ways.
- More traditional methods of constructing confidence intervals have better properties/guarantees but these require stronger assumptions.

Limitations of Bootstrapping

- This doesn't work too well when the sample size is small.
  - Smaller sample size $\Rightarrow$ Harder to approximate true distribution
  - As an extreme consider a sample size of 1. Bootstrapping is not going to let me change the sample.
  - Note: This is still going to be an issue when computing confidence intervals in other ways.
- More traditional methods of constructing confidence intervals have better properties/guarantees but these require stronger assumptions.
- We'll talk more about these on Wednesday.

Where We've Been and Where We're Going

Last week and before:

- Understanding central tendency and spread of a sample

Where We've Been and Where We're Going

Last week and before:

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Where We've Been and Where We're Going

Last week and before:

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Where We've Been and Where We're Going

Last week and before:

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Today: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution

Where We've Been and Where We're Going

Last week and before:

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Today: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using Bootstrapping.

Where We've Been and Where We're Going

Last week and before:

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Today: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using Bootstrapping.
- Confidence intervals using bootstrapping

Where We've Been and Where We're Going

Last week and before:

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Today: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using Bootstrapping.
- Confidence intervals using bootstrapping

Where We've Been and Where We're Going

Last week and before:

- Understanding central tendency and spread of a sample
- Understanding association within a sample.

Today: Going from the sample to the population: How confident are we about our estimates?

- Sampling Distribution
- Estimating the sampling distribution using Bootstrapping.
- Confidence intervals using bootstrapping

Wednesday:

- Getting guarantees about confidence intervals (with more assumptions)