Bootstrapping

Nate Wells

Math 141, 3/7/22

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Review how the sampling distribution can be used to assess sampling variability
- Discuss bootstrapping as means of approximating the sampling distribution

Section 1

Sampling Distribution

Polling Example

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

Polling Example

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided

Polling Example

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.

Polling Example

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)

Polling Example

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Trump/Pence. Based on election results, $p = 0.4884$.

Polling Example

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking

  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Trump/Pence. Based on election results, $p = 0.4884$.
- **Sampling Method**: SRS(?) of size $n = 1020$ obtained using phone-numbers

Polling Example

- A Oct 29 - Nov 1 2020 poll by Marist College surveyed 1020 registered voters in Pennsylvania by landline or mobile number, asking
  If November's election were held today, whom would you support?

  The options were: Joe Biden/Kamala Harris, Donald Trump/Mike Pence, Other, Undecided.
- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
- **Population**: All registered voters in Pennsylvania ($N \approx 9 \text{ million}$)
- **Population Parameter**: The proportion p of registered voters who plan to vote for Trump/Pence. Based on election results, $p = 0.4884$.
- **Sampling Method**: SRS(?) of size $n = 1020$ obtained using phone-numbers
- **Point Estimate/Sample Statistic**: The sample proportion $\hat{p}$ of Americans who plan to vote for Trump/Pence. In this case, $\hat{p} = 0.46$.

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
    - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
  - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
  - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
    - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
    - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.
    - But if we just want an estimate that is likely close true proportion, then we should be very confidant.

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
    - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
    - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.
    - But if we just want an estimate that is likely close true proportion, then we should be very confidant.
- The sampling distribution tells us how much variability to expect from sample to sample.

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
    - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
    - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.
    - But if we just want an estimate that is likely close true proportion, then we should be very confidant.
- The sampling distribution tells us how much variability to expect from sample to sample.
    - Using probability theory, we can show that the standard error for the sampling distribution of the proportion with sample size n is at most $\frac{1}{\sqrt{4n}}$

Sampling Variability

- How confidant should we be in the accuracy of our estimate of $\hat{p} = 0.46$?
    - There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
    - If we want to claim our estimate is exactly equal to true proportion, we should be skeptical.
    - But if we just want an estimate that is likely close true proportion, then we should be very confidant.
- The sampling distribution tells us how much variability to expect from sample to sample.
    - Using probability theory, we can show that the standard error for the sampling distribution of the proportion with sample size n is at most $\frac{1}{\sqrt{4n}}$
    - For a sample of size $n = 1020$, the standard error is at most $SE = 0.016$.

Sampling Variability

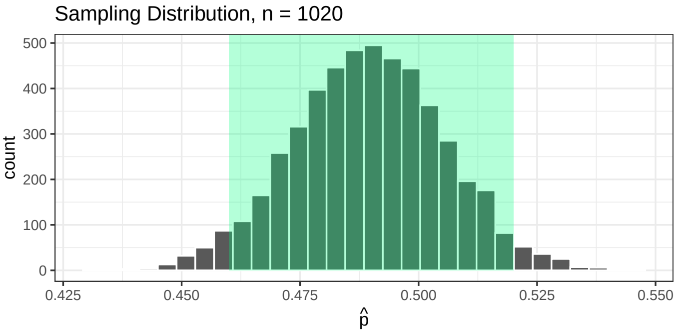
- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$

Sampling Variability

- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$
- We can simulate 5000 samples of size 1020 to see how many have $\hat{p}$ far from $p = .49$.

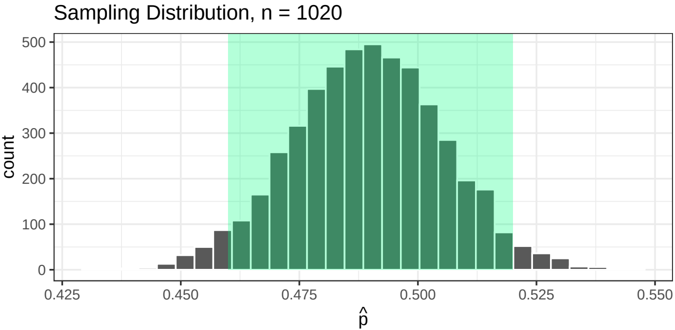
Sampling Variability

- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$
- We can simulate 5000 samples of size 1020 to see how many have $\hat{p}$ far from $p = .49$.



Sampling Variability

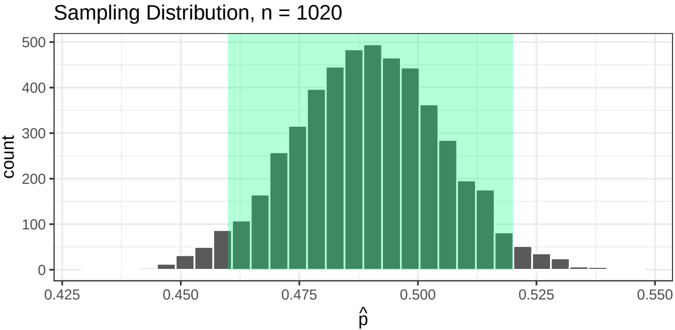
- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$
- We can simulate 5000 samples of size 1020 to see how many have $\hat{p}$ far from $p = .49$.



- Of these, only 6% differed from the true value $p = .49$ by more than .03

Sampling Variability

- Suppose the true proportion of support for Trump/Pence were actually $p = 0.49$
- We can simulate 5000 samples of size 1020 to see how many have $\hat{p}$ far from $p = .49$.



- Of these, only 6% differed from the true value $p = .49$ by more than .03
- But this also means that for 94% of samples, the true proportion p is within 0.03 of the sample proportion $\hat{p}$.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

- In order to form the sampling distribution, we need to collect a large number of samples.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

- In order to form the sampling distribution, we need to collect a large number of samples.
    - But if we can collect enough samples to form the sampling distribution, we probably can just take a census of the population.

The Problem

- The sampling distribution tells us how much variability to expect from sample to sample.
- For sampling distributions that are approximately bell-shaped (usually true if $n \geq 30$), 95% of all sample means will be within 2 standard error units of the true parameter.
- We can use the standard error (i.e. standard deviation of sampling distribution) to assess how close the typical sample statistic will be to the population parameter

What is the problem in practice?

- In order to form the sampling distribution, we need to collect a large number of samples.
    - But if we can collect enough samples to form the sampling distribution, we probably can just take a census of the population.
- The fix?

Section 2

Bootstrapping

Bootstrapping

- The term *bootstrapping* refers to the phrase “to pull oneself up by one’s bootstraps”

Bootstrapping

- The term *bootstrapping* refers to the phrase “to pull oneself up by one’s bootstraps”
    - The phrase originated in the 19th century as reference to a ludicrous or impossible feat

Bootstrapping

- The term *bootstrapping* refers to the phrase “to pull oneself up by one’s bootstraps”
    - The phrase originated in the 19th century as reference to a ludicrous or impossible feat
    - By the mid 20th century, its meaning had changed to suggest a success by one’s own efforts, without outside help

Bootstrapping

- The term *bootstrapping* refers to the phrase “to pull oneself up by one’s bootstraps”
    - The phrase originated in the 19th century as reference to a ludicrous or impossible feat
    - By the mid 20th century, its meaning had changed to suggest a success by one’s own efforts, without outside help
- Its use in statistics alludes to both interpretations.

The Bootstrap Trick

The Impossible Task:

The Bootstrap Trick

The Impossible Task:

- How can we learn about the sampling distribution, if we only have 1 sample?

The Bootstrap Trick

The Impossible Task:

- How can we learn about the sampling distribution, if we only have 1 sample?

The “Ludicrous” Solution obtained without outside help:

The Bootstrap Trick

The Impossible Task:

- How can we learn about the sampling distribution, if we only have 1 sample?

The “Ludicrous” Solution obtained without outside help:

- Draw repeated samples from the original sample at hand; compute the statistic of interest for each; plot the resulting distribution

The Bootstrap Trick

The Impossible Task:

- How can we learn about the sampling distribution, if we only have 1 sample?

The “Ludicrous” Solution obtained without outside help:

- Draw repeated samples from the original sample at hand; compute the statistic of interest for each; plot the resulting distribution

The Main Idea:

The Bootstrap Trick

The Impossible Task:

- How can we learn about the sampling distribution, if we only have 1 sample?

The “Ludicrous” Solution obtained without outside help:

- Draw repeated samples from the original sample at hand; compute the statistic of interest for each; plot the resulting distribution

The Main Idea:

- The original sample approximates the population

The Bootstrap Trick

The Impossible Task:

- How can we learn about the sampling distribution, if we only have 1 sample?

The “Ludicrous” Solution obtained without outside help:

- Draw repeated samples from the original sample at hand; compute the statistic of interest for each; plot the resulting distribution

The Main Idea:

- The original sample approximates the population
- Resampling from the sample approximates sampling many times from the population

The Bootstrap Trick

The Impossible Task:

- How can we learn about the sampling distribution, if we only have 1 sample?

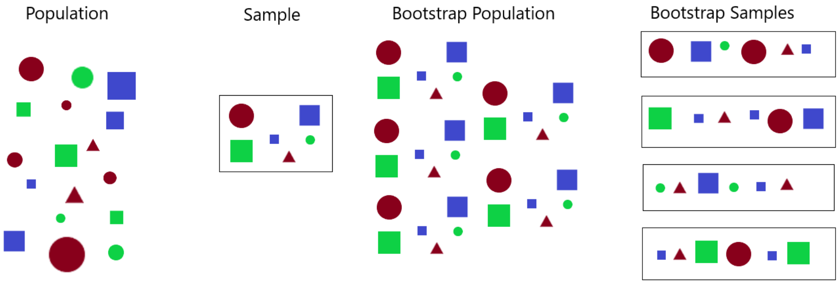
The “Ludicrous” Solution obtained without outside help:

- Draw repeated samples from the original sample at hand; compute the statistic of interest for each; plot the resulting distribution

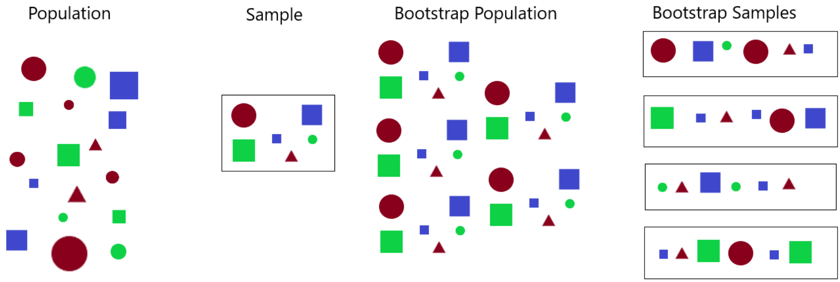
The Main Idea:

- The original sample approximates the population
- Resampling from the sample approximates sampling many times from the population
- The distribution of statistics from the resamples approximates the sampling distribution

Theory

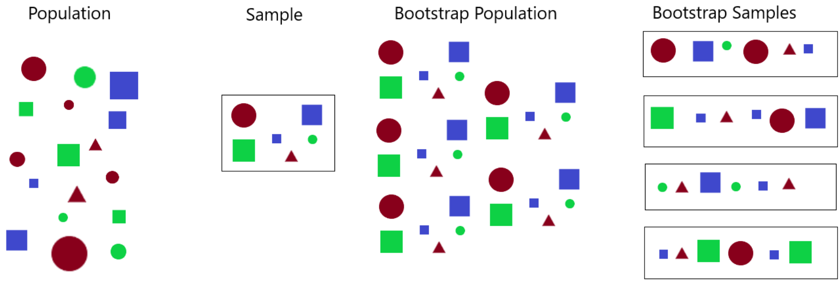

Population
Sample
Bootstrap Population
Bootstrap Samples

Theory



- We could copy the original sample many times to create a bootstrap population, and then sample *without replacement* to get bootstrap samples

Theory



- We could copy the original sample many times to create a bootstrap population, and then sample *without replacement* to get bootstrap samples
- But this is the same as sampling *with replacement* from the original sample

The Bootstrap Procedure

To generate a **bootstrap distribution**:

1. Obtain an SRS of size n from the population.
2. Generate a bootstrap sample of size n by resampling *with* replacement from the original sample
3. Repeat (2) a large number of times (with technology, at least 1000 times)
4. For each bootstrap sample, calculate the appropriate statistic (called the **bootstrap statistic**)
5. The collection of the bootstrap statistics form the **bootstrap distribution**

Proof of Concept

- Consider a very large deck of cards (5200 cards) with 100 of each standard card.

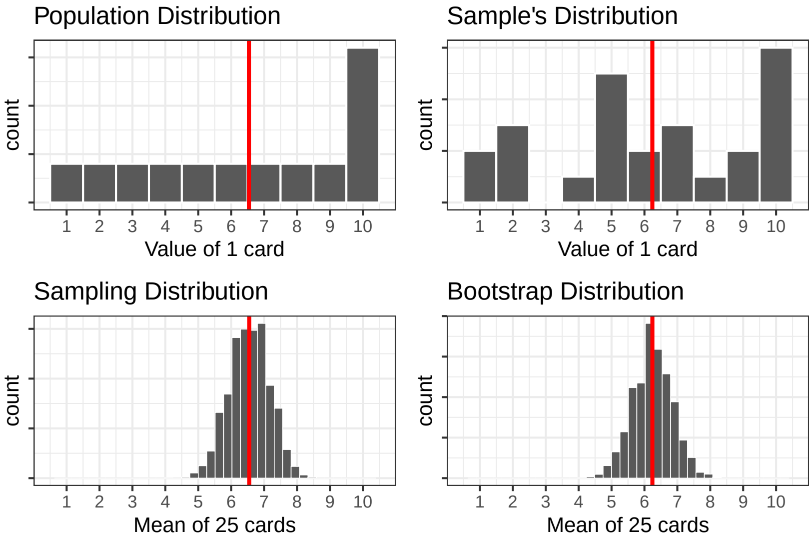
Proof of Concept

- Consider a very large deck of cards (5200 cards) with 100 of each standard card.
- Suppose we draw a sample hand of size 25 and calculate the mean value of the hand.

Proof of Concept

- Consider a very large deck of cards (5200 cards) with 100 of each standard card.
- Suppose we draw a sample hand of size 25 and calculate the mean value of the hand.
- Since we have the deck of cards, we can look at:
  1. The population distribution
  2. The single sample's distribution
  3. The sampling distribution for sample means
  4. The bootstrap distribution for sample means

House of Cards



House of Cards

We can compute some relevant statistics:

Population:

mean_value	sd_value
6.538462	3.153211

Sample:

mean_value	sd_value
6.24	3.072458

Sampling Distribution:

mean_xbar	sd_xbar
6.55047	0.6162582

Bootstrap Distribution:

mean_xbar	sd_xbar
6.24119	0.604233

Reproduction Rate for Covid-19

Researchers are interested in the reproduction rate of COVID-19.

- We have a sample of 50 infected individuals and perform contract tracing to determine how many other individuals each infects.

Reproduction Rate for Covid-19

Researchers are interested in the reproduction rate of COVID-19.

- We have a sample of 50 infected individuals and perform contract tracing to determine how many other individuals each infects.

```
##   infected  n
## 1        0  5
## 2        1 13
## 3        2 14
## 4        3 12
## 5        4  5
## 6        6  1
```

```
##   mean_infected
## 1          2.06
```

Reproduction Rate for Covid-19

Researchers are interested in the reproduction rate of COVID-19.

- We have a sample of 50 infected individuals and perform contract tracing to determine how many other individuals each infects.

```
##   infected  n
## 1        0  5
## 2        1 13
## 3        2 14
## 4        3 12
## 5        4  5
## 6        6  1
```

```
##   mean_infected
## 1          2.06
```

- Is the true reproduction rate exactly 2.06?

Reproduction Rate for Covid-19

Researchers are interested in the reproduction rate of COVID-19.

- We have a sample of 50 infected individuals and perform contract tracing to determine how many other individuals each infects.

```
##   infected  n
## 1        0  5
## 2        1 13
## 3        2 14
## 4        3 12
## 5        4  5
## 6        6  1
```

```
##   mean_infected
## 1          2.06
```

- Is the true reproduction rate exactly 2.06?
  - Surely not! This is just one sample of size 50

Reproduction Rate for Covid-19

Researchers are interested in the reproduction rate of COVID-19.

- We have a sample of 50 infected individuals and perform contract tracing to determine how many other individuals each infects.

```
##   infected  n
## 1        0  5
## 2        1 13
## 3        2 14
## 4        3 12
## 5        4  5
## 6        6  1
```

```
##   mean_infected
## 1          2.06
```

- Is the true reproduction rate exactly 2.06?
  - Surely not! This is just one sample of size 50
- But how much does the reproduction rate vary from sample to sample?

Bootstrap Reproduction Rate

Create the bootstrap samples:

```
bootstrap_samples <- covid %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 2000)

head(bootstrap_samples)
```

```
## # A tibble: 100,000 x 2
## # Groups:   replicate [2,000]
##    replicate infected
##        <int>    <int>
##  1         1        2
##  2         1        1
##  3         1        1
##  4         1        0
##  5         1        1
##  6         1        0
##  7         1        2
##  8         1        3
##  9         1        3
## 10         1        3
## # ... with 99,990 more rows
```

Bootstrap Reproduction Rate

Compute bootstrap statistics:

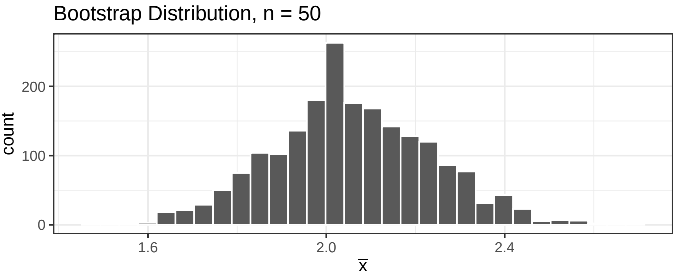
```
bootstrap_stats <- bootstrap_samples %>%
  group_by(replicate) %>%
  summarize(x_bar = mean(infected))

bootstrap_stats
```

```
## # A tibble: 2,000 x 2
##    replicate x_bar
##        <int> <dbl>
##  1         1  1.86
##  2         2  2.36
##  3         3  2.22
##  4         4  1.86
##  5         5  1.88
##  6         6  1.6
##  7         7  2.02
##  8         8  2.16
##  9         9  2.2
## 10        10  1.8
## # ... with 1,990 more rows
```

Bootstrap Reproduction Rate

Graph the bootstrap distribution:

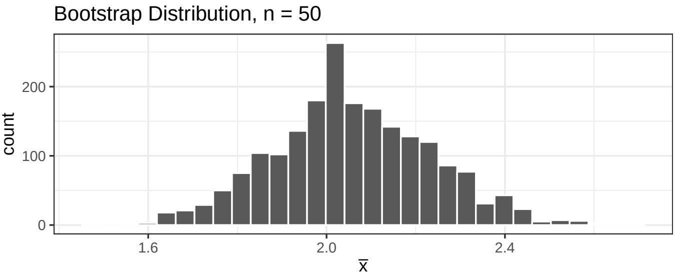
```
ggplot(bootstrap_stats, aes(x = x_bar))+
  geom_histogram(bins = 30, color = "white")+
  labs(title = "Bootstrap Distribution, n = 50", x = expression(bar(x)))
```



Bootstrap Reproduction Rate

Graph the bootstrap distribution:

```
ggplot(bootstrap_stats, aes(x = x_bar))+
  geom_histogram(bins = 30, color = "white")+
  labs(title = "Bootstrap Distribution, n = 50", x = expression(bar(x)))
```



Use the bootstrap distribution to estimate the standard error:

```
bootstrap_stats %>% summarize(SE = sd(x_bar))
```

```
## # A tibble: 1 x 1
##       SE
##    <dbl>
## 1 0.177
```