Bootstrapping

Prof. Wells

Math 209, 3/13/23

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Discuss bootstrapping as means of approximating the sampling distribution
- Use the bootstrap distribution to create general confidence intervals

Confidence Interval Review and Conundrum

Confidence Interval Review and Conundrum

- We do have one more problem:

Confidence Interval Review and Conundrum

- We do have one more problem:
  - To make a confidence interval, we need the sampling distribution in order to compute the standard error.

Confidence Interval Review and Conundrum

- We do have one more problem:
  - To make a confidence interval, we need the sampling distribution in order to compute the standard error.
  - But in order to visualize the sampling distribution, and compute it's standard deviation, we would need to obtain thousands of samples.

Confidence Interval Review and Conundrum

- We do have one more problem:
  - To make a confidence interval, we need the sampling distribution in order to compute the standard error.
  - But in order to visualize the sampling distribution, and compute it's standard deviation, we would need to obtain thousands of samples.
- In practice, we just have 1 sample! And if we had the time / funding to obtain thousands of samples, we could probably just conduct a census of the population.

Confidence Interval Review and Conundrum

- We do have one more problem:
  - To make a confidence interval, we need the sampling distribution in order to compute the standard error.
  - But in order to visualize the sampling distribution, and compute it's standard deviation, we would need to obtain thousands of samples.
- In practice, we just have 1 sample! And if we had the time / funding to obtain thousands of samples, we could probably just conduct a census of the population.
- Miraculously, it turns out we can assess the variability and shape of the sampling distribution using just a single sample!

Confidence Interval Review and Conundrum

- We do have one more problem:
  - To make a confidence interval, we need the sampling distribution in order to compute the standard error.
  - But in order to visualize the sampling distribution, and compute it's standard deviation, we would need to obtain thousands of samples.
- In practice, we just have 1 sample! And if we had the time / funding to obtain thousands of samples, we could probably just conduct a census of the population.
- Miraculously, it turns out we can assess the variability and shape of the sampling distribution using just a single sample!
  - This process is called **Bootstrapping**, which we'll investigate next week!

Section 1

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 is relatively similar to 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 is relatively similar to the population
- Resampling (**with replacement**) from the *original* sample approximates sampling from the population (**without replacement**)

The Bootstrap Trick

The Impossible Task:

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

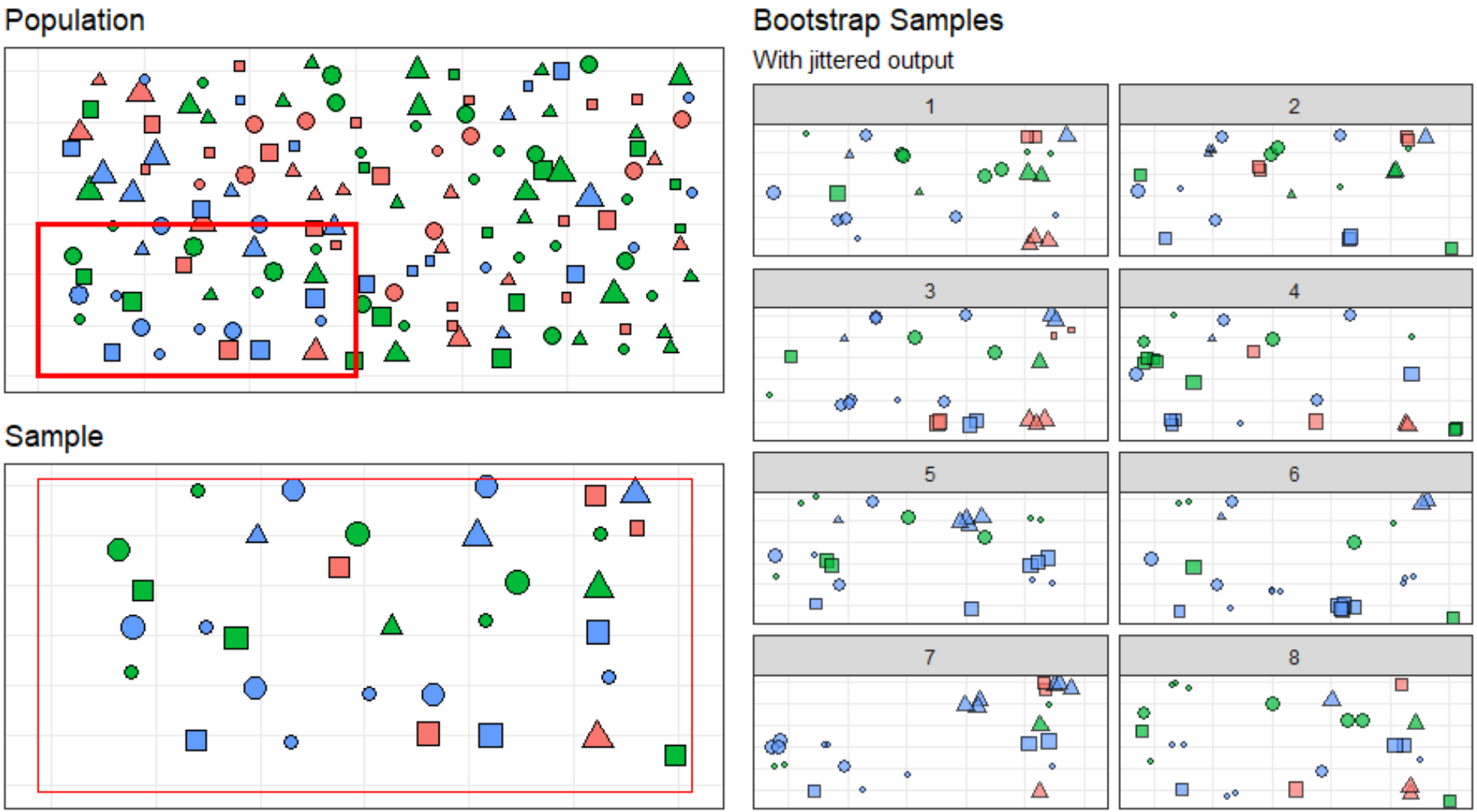
The "Ludicrous" Solution obtained without outside help:

- 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 is relatively similar to the population
- Resampling (**with replacement**) from the *original* sample approximates sampling from the population (**without replacement**)
- The distribution of sample statistics calculated from resamples should look like the sampling distribution

Bootstrapping Visualized



The Bootstrap Procedure

To generate a **bootstrap distribution**:

1. Obtain an SRS of size n from the population.
2. Draw a sample of size n *with* replacement from the original sample (called a **bootstrap 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.

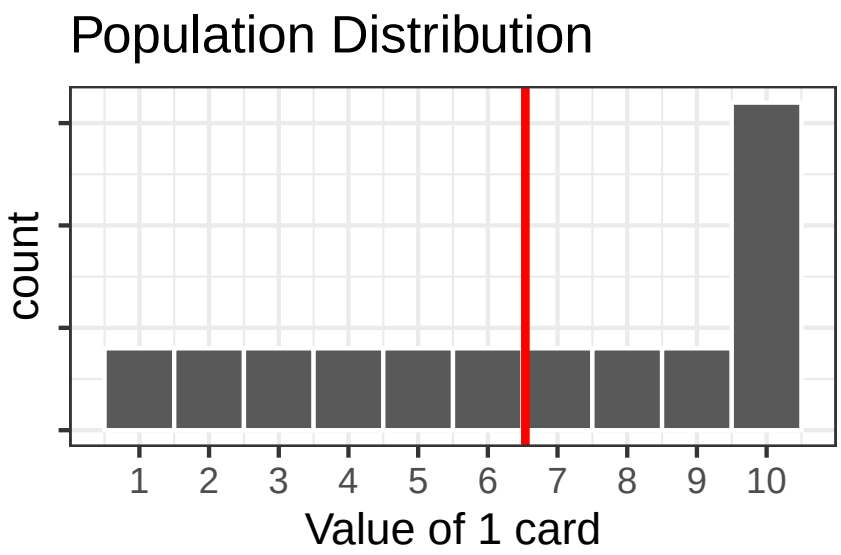
Proof of Concept

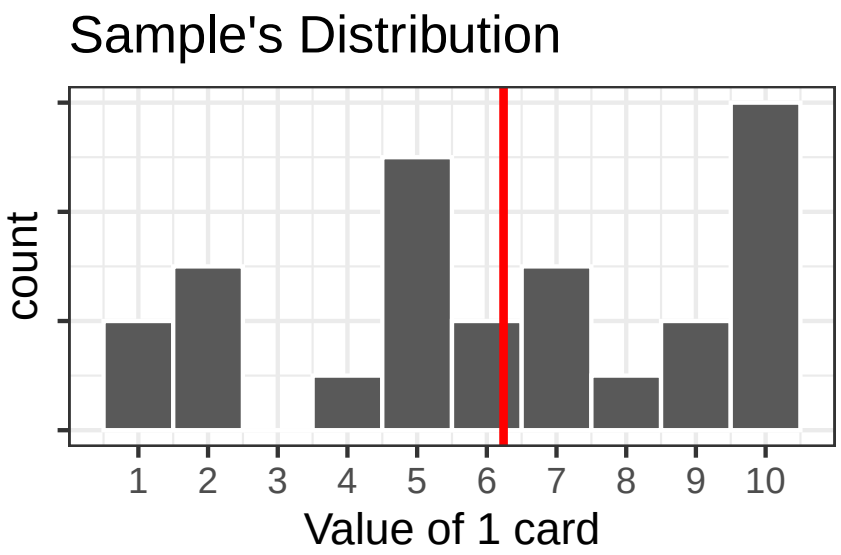
- Consider a very large deck of cards (5200 cards) with 100 of each standard card.
- Suppose we draw a sample hand of size 25 and calculate the mean value of the hand.

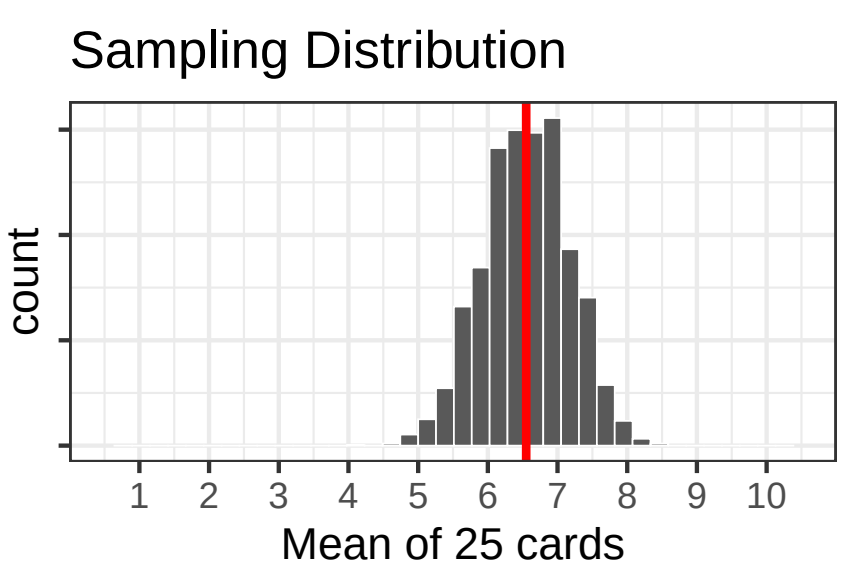
Proof of Concept

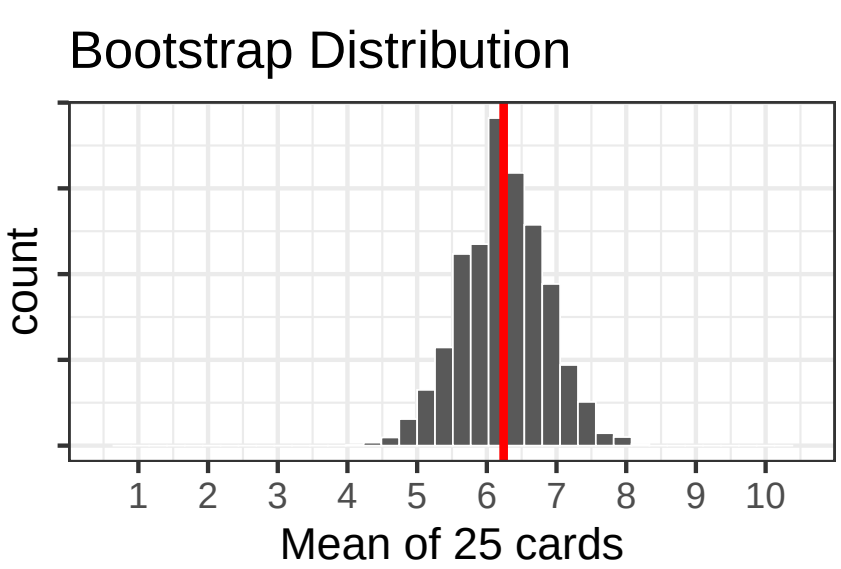
- Consider a very large deck of cards (5200 cards) with 100 of each standard card.
- Suppose we draw a sample hand of size 25 and calculate the mean value of the hand.
- Since we have the deck of cards, we can look at:
  1. The population distribution
  2. The single sample's distribution
  3. The sampling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

Section 2

Bootstrapping Example

Reproduction Rate for Covid-19

When COVID-19 first emerged in early 2020, researchers were interested in estimating the reproduction rate of this new virus to determine the risk it posed.

Reproduction Rate for Covid-19

When COVID-19 first emerged in early 2020, researchers were interested in estimating the reproduction rate of this new virus to determine the risk it posed.

- **Reproduction Rate** is the average number of cases directly generated by one case in a population where all individuals are susceptible.

Reproduction Rate for Covid-19

When COVID-19 first emerged in early 2020, researchers were interested in estimating the reproduction rate of this new virus to determine the risk it posed.

- **Reproduction Rate** is the average number of cases directly generated by one case in a population where all individuals are susceptible.
  - Over the past 3 years, scientists have gathered significant data on the Reproduction Rate of COVID-19 and its variants.
  - But at the start of the pandemic data on the Reproduction Rate was relatively limited.

Reproduction Rate for Covid-19

When COVID-19 first emerged in early 2020, researchers were interested in estimating the reproduction rate of this new virus to determine the risk it posed.

- **Reproduction Rate** is the average number of cases directly generated by one case in a population where all individuals are susceptible.
  - Over the past 3 years, scientists have gathered significant data on the Reproduction Rate of COVID-19 and its variants.
  - But at the start of the pandemic data on the Reproduction Rate was relatively limited.
- **Parameter of Interest**: Reproduction Rate μ for COVID-19 (original strain)
- **Population**: Theoretical population of all COVID-19 susceptible individuals
- **Sample**: 50 individuals infected with COVID-19 in February 2020
- **Statistic**: Average number of cases $\bar{x}$ generated by each individual in sample

Reproduction Rate for Covid-19

When COVID-19 first emerged in early 2020, researchers were interested in estimating the reproduction rate of this new virus to determine the risk it posed.

- **Reproduction Rate** is the average number of cases directly generated by one case in a population where all individuals are susceptible.
  - Over the past 3 years, scientists have gathered significant data on the Reproduction Rate of COVID-19 and its variants.
  - But at the start of the pandemic data on the Reproduction Rate was relatively limited.
- **Parameter of Interest**: Reproduction Rate μ for COVID-19 (original strain)
- **Population**: Theoretical population of all COVID-19 susceptible individuals
- **Sample**: 50 individuals infected with COVID-19 in February 2020
- **Statistic**: Average number of cases $\bar{x}$ generated by each individual in sample
- **Sampling Method**: Does this represent an SRS?
  - Individuals selected had non-overlapping contact networks (Independence)
  - But did each member of the population have equal chance to be selected?

Reproduction Rate for Covid-19 Sample

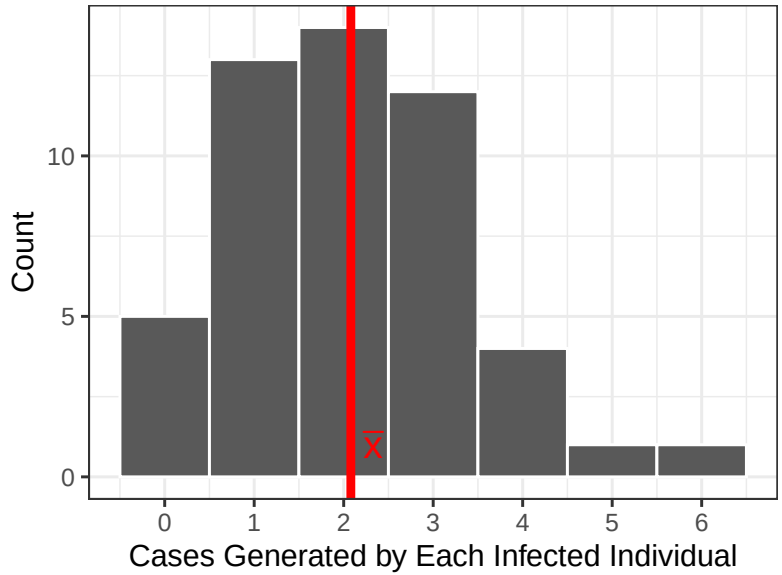
- We have a sample (`covid`) of 50 infected individuals and perform contract tracing to determine how many other individuals each infects (`n_infected`)

Reproduction Rate for Covid-19 Sample

- We have a sample (`covid`) of 50 infected individuals and perform contract tracing to determine how many other individuals each infects (`n_infected`)

```
covid %>% group_by(n_infected) %>%
  summarize(number = n())
```

```
## # A tibble: 7 x 2
##   n_infected number
##        <int>  <int>
## 1          0      5
## 2          1     13
## 3          2     14
## 4          3     12
## 5          4      4
## 6          5      1
## 7          6      1
```

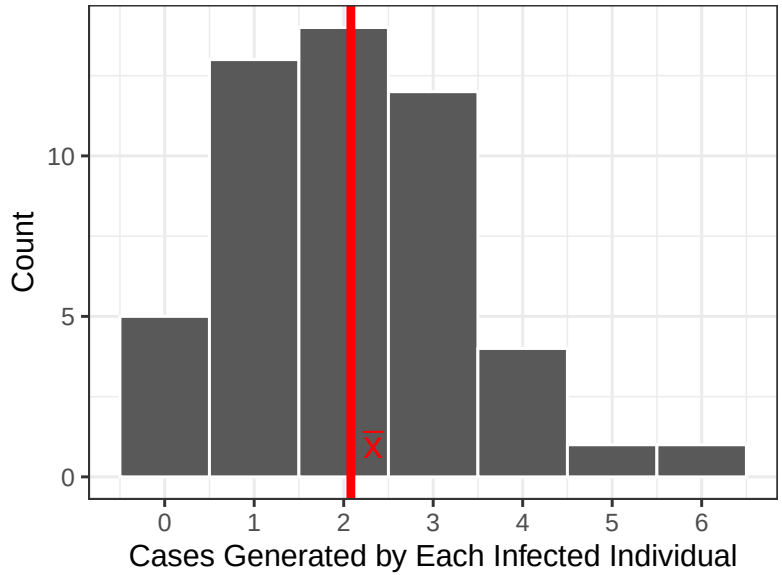


Reproduction Rate for Covid-19 Sample

- We have a sample (`covid`) of 50 infected individuals and perform contract tracing to determine how many other individuals each infects (`n_infected`)

```
covid %>% group_by(n_infected) %>%
  summarize(number = n())
```

```
## # A tibble: 7 x 2
##   n_infected number
##        <int>  <int>
## 1          0      5
## 2          1     13
## 3          2     14
## 4          3     12
## 5          4      4
## 6          5      1
## 7          6      1
```




```
covid %>% summarize(mean_number_infected = mean(n_infected))
```

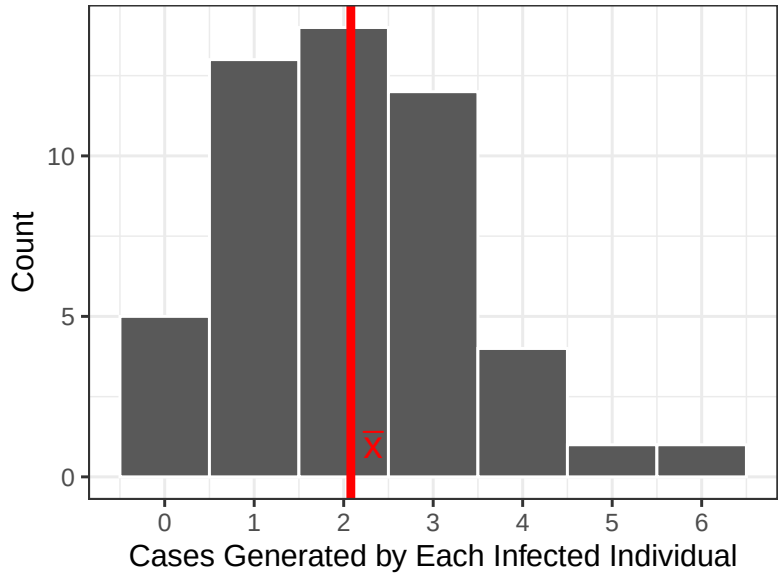
```
##   mean_number_infected
## 1                 2.08
```

Reproduction Rate for Covid-19 Sample

- We have a sample (`covid`) of 50 infected individuals and perform contract tracing to determine how many other individuals each infects (`n_infected`)

```
covid %>% group_by(n_infected) %>%
  summarize(number = n())
```

```
## # A tibble: 7 x 2
##   n_infected number
##        <int>  <int>
## 1          0      5
## 2          1     13
## 3          2     14
## 4          3     12
## 5          4      4
## 6          5      1
## 7          6      1
```




```
covid %>% summarize(mean_number_infected = mean(n_infected))
```

```
##   mean_number_infected
## 1                 2.08
```

- Is the true reproduction rate exactly 2.06? Probably not. Instead, we'll create a confidence interval to estimate it.

Bootstrap Reproduction Rate

Create the bootstrap samples:

```
set.seed(121)
bootstrap_samples <- covid %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 5000)
```

Bootstrap Reproduction Rate

Create the bootstrap samples:

```
set.seed(121)
bootstrap_samples <- covid %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 5000)
```

```
bootstrap_samples
```

```
## # A tibble: 250,000 x 3
## # Groups:   replicate [5,000]
##    replicate    id n_infected
##        <int> <int>      <int>
##  1         1    28          2
##  2         1    12          1
##  3         1     7          1
##  4         1     4          0
##  5         1     9          1
##  6         1     1          0
##  7         1    27          2
##  8         1    39          3
##  9         1    37          3
## 10         1    44          3
## # ... with 249,990 more rows
## # i Use `print(n = ...)` to see more rows
```

Bootstrap Reproduction Rate

Create the bootstrap samples:

```
set.seed(121)
bootstrap_samples <- covid %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 5000)
```

```
bootstrap_samples
```

```
## # A tibble: 250,000 x 3
## # Groups:   replicate [5,000]
##    replicate    id n_infected
##        <int> <int>      <int>
##  1         1    28          2
##  2         1    12          1
##  3         1     7          1
##  4         1     4          0
##  5         1     9          1
##  6         1     1          0
##  7         1    27          2
##  8         1    39          3
##  9         1    37          3
## 10         1    44          3
## # ... with 249,990 more rows
## # i Use `print(n = ...)` to see more rows
```

```
bootstrap_samples %>% group_by(replicate) %>%
  summarize(n = n())
```

```
## # A tibble: 5,000 x 2
##    replicate     n
##        <int> <int>
##  1         1    50
##  2         2    50
##  3         3    50
##  4         4    50
##  5         5    50
##  6         6    50
##  7         7    50
##  8         8    50
##  9         9    50
## 10        10    50
## # ... with 4,990 more rows
## # i Use `print(n = ...)` to see more rows
```

Bootstrap Reproduction Rate

Create the bootstrap samples:

```
set.seed(121)
bootstrap_samples <- covid %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 5000)
```

```
bootstrap_samples
```

```
## # A tibble: 250,000 x 3
## # Groups:   replicate [5,000]
##    replicate    id n_infected
##        <int> <int>      <int>
##  1         1    28          2
##  2         1    12          1
##  3         1     7          1
##  4         1     4          0
##  5         1     9          1
##  6         1     1          0
##  7         1    27          2
##  8         1    39          3
##  9         1    37          3
## 10         1    44          3
## # ... with 249,990 more rows
## # i Use `print(n = ...)` to see more rows
```

```
bootstrap_samples %>% group_by(replicate) %>%
  summarize(n = n())
```

```
## # A tibble: 5,000 x 2
##    replicate     n
##        <int> <int>
##  1         1    50
##  2         2    50
##  3         3    50
##  4         4    50
##  5         5    50
##  6         6    50
##  7         7    50
##  8         8    50
##  9         9    50
## 10        10    50
## # ... with 4,990 more rows
## # i Use `print(n = ...)` to see more rows
```

- Each bootstrap sample consists of 50 observations sampled *with replacement* from the original sample (`size = 50`)
- We have a total of 5000 bootstrap samples (`reps = 5000`)

Bootstrap Reproduction Rate

Compute bootstrap statistics:

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

Bootstrap Reproduction Rate

Compute bootstrap statistics:

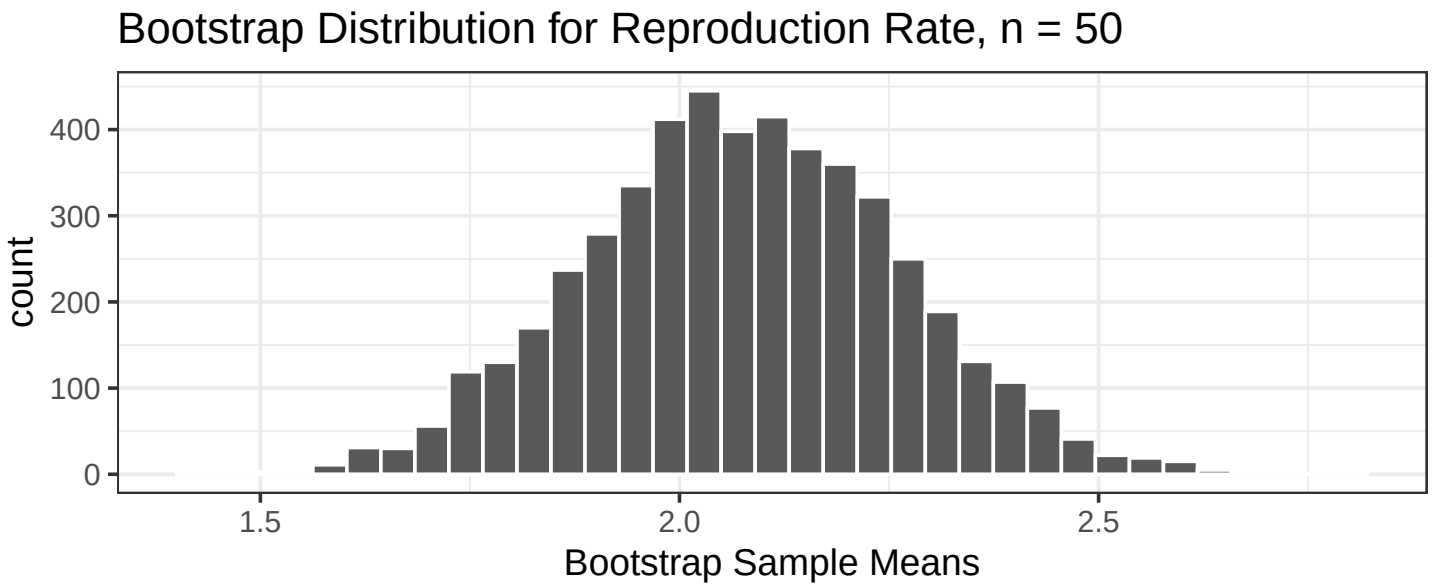
```
bootstrap_stats <- bootstrap_samples %>%
  group_by(replicate) %>%
  summarize(x_bar = mean(n_infected))
```

```
## # A tibble: 5,000 x 2
##    replicate x_bar
##        <int> <dbl>
##  1         1  1.88
##  2         2  2.38
##  3         3  2.24
##  4         4  1.88
##  5         5  1.88
##  6         6  1.6
##  7         7  2.02
##  8         8  2.16
##  9         9  2.22
## 10        10  1.8
## # ... with 4,990 more rows
## # i Use `print(n = ...)` to see more rows
```

- We now have 5000 sample means based on the bootstrap samples, and can assess their variability

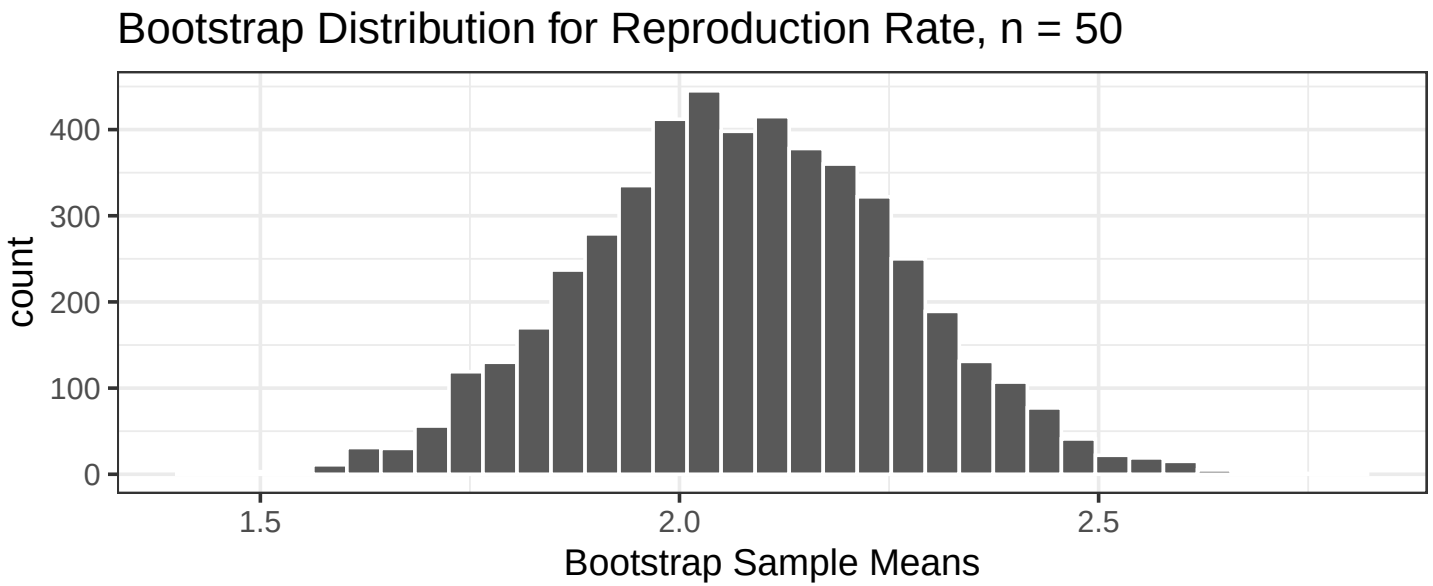
Bootstrap Reproduction Rate

Graph the bootstrap distribution:



Bootstrap Reproduction Rate

Graph the bootstrap distribution:



- Use the bootstrap distribution to estimate the standard error:

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

```
## # A tibble: 1 x 1
##      SE
##   <dbl>
## 1 0.186
```

Confidence Interval for Reproduction Rate

- Our sample reproduction rate was $\bar{x} = 2.04$.

Confidence Interval for Reproduction Rate

- Our sample reproduction rate was $\bar{x} = 2.04$.
- Based on the bootstrap distribution, this statistic has a standard error of $SE = 0.186$.

Confidence Interval for Reproduction Rate

- Our sample reproduction rate was $\bar{x} = 2.04$.
- Based on the bootstrap distribution, this statistic has a standard error of $SE = 0.186$.
- Recall that a 95% confidence interval has the form

$$\bar{x} \pm 2 \cdot SE$$

Confidence Interval for Reproduction Rate

- Our sample reproduction rate was $\bar{x} = 2.04$.
- Based on the bootstrap distribution, this statistic has a standard error of $SE = 0.186$.
- Recall that a 95% confidence interval has the form

$$\bar{x} \pm 2 \cdot SE$$

- Our 95% confidence interval for the true reproduction rate of COVID-19 is

$$2.06 \pm 2 \cdot 0.186$$

Confidence Interval for Reproduction Rate

- Our sample reproduction rate was $\bar{x} = 2.04$.
- Based on the bootstrap distribution, this statistic has a standard error of $SE = 0.186$.
- Recall that a 95% confidence interval has the form

$$\bar{x} \pm 2 \cdot SE$$

- Our 95% confidence interval for the true reproduction rate of COVID-19 is

$$2.06 \pm 2 \cdot 0.186$$

- Our best guess for the reproduction rate is between 1.688 and 2.432. This method has a success rate of 95%.

Confidence Interval for Reproduction Rate

- Our sample reproduction rate was $\bar{x} = 2.04$.
- Based on the bootstrap distribution, this statistic has a standard error of $SE = 0.186$.
- Recall that a 95% confidence interval has the form

$$\bar{x} \pm 2 \cdot SE$$

- Our 95% confidence interval for the true reproduction rate of COVID-19 is

$$2.06 \pm 2 \cdot 0.186$$

- Our best guess for the reproduction rate is between 1.688 and 2.432. This method has a success rate of 95%.
- For reference, this interval matches the one provided by the WHO on 1/23/20.

Generalized Confidence Intervals

- In the previous example, we used the fact that for approximately bell-shaped sampling distributions, 95% of of sample statistics are within 2 SE of the population parameter

Generalized Confidence Intervals

- In the previous example, we used the fact that for approximately bell-shaped sampling distributions, 95% of of sample statistics are within 2 SE of the population parameter
  - But suppose we instead want a different success rate for our estimation method

Generalized Confidence Intervals

- In the previous example, we used the fact that for approximately bell-shaped sampling distributions, 95% of of sample statistics are within 2 SE of the population parameter
  - But suppose we instead want a different success rate for our estimation method
  - Or suppose we want to create interval estimates for sampling distributions that are not bell-shaped

Generalized Confidence Intervals

- In the previous example, we used the fact that for approximately bell-shaped sampling distributions, 95% of of sample statistics are within 2 SE of the population parameter
  - But suppose we instead want a different success rate for our estimation method
  - Or suppose we want to create interval estimates for sampling distributions that are not bell-shaped
- We can make these modifications again using the bootstrap approximation to the sampling distribution

Section 3

General Confidence Intervals

General Confidence Intervals

General Confidence Intervals

The $C\%$ confidence interval for a parameter is an interval estimate that is computed from sample data by a method that captures the parameter for $C\%$ of all samples.

Review: Percentiles and Quantiles

- For a number k between 0 and 100, the kth **percentile** of a distribution is the value so that $k\%$ of the data is less than or equal to that value.

Review: Percentiles and Quantiles

- For a number k between 0 and 100, the kth **percentile** of a distribution is the value so that k% of the data is less than or equal to that value.
  - The median is the 50th percentile of a distribution, and the 1st/3rd quartiles are the 25th and 75th percentiles, respectively.

Review: Percentiles and Quantiles

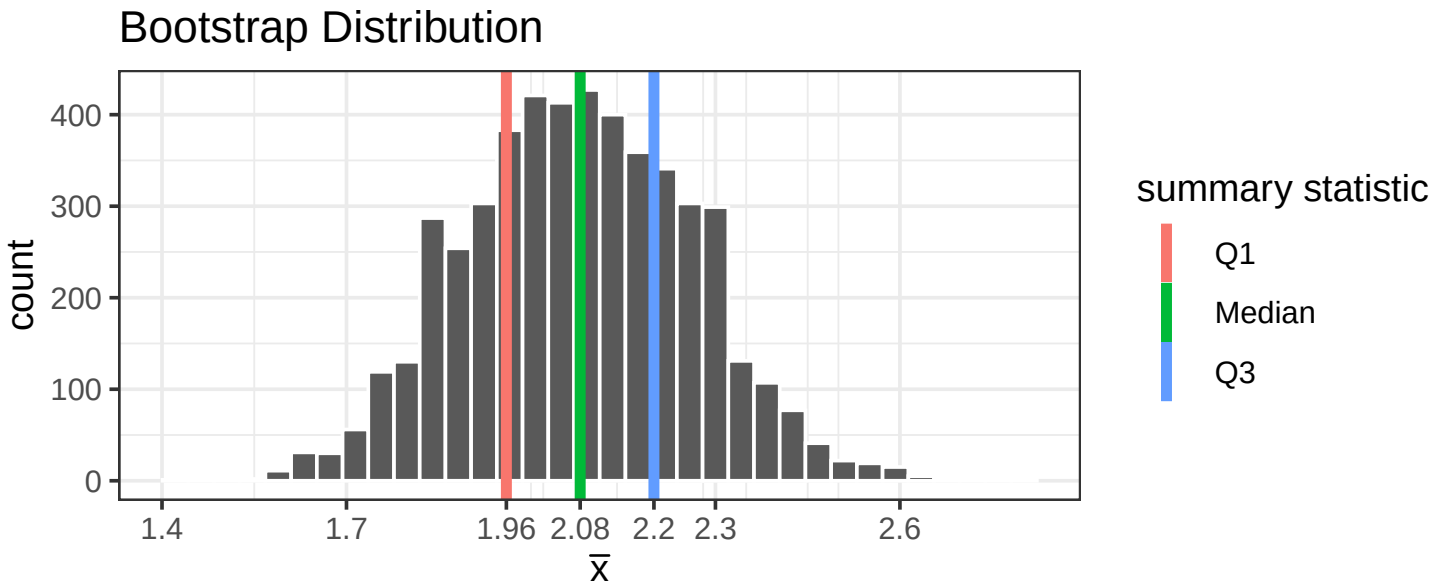
- For a number k between 0 and 100, the kth **percentile** of a distribution is the value so that $k\%$ of the data is less than or equal to that value.
  - The median is the 50th percentile of a distribution, and the 1st/3rd quartiles are the 25th and 75th percentiles, respectively.
- For a number p between 0 and 1, the p **quantile** of a distribution is the value so that a proportion p of the data is less than or equal to that value.

Review: Percentiles and Quantiles

- For a number k between 0 and 100, the kth **percentile** of a distribution is the value so that k% of the data is less than or equal to that value.
  - The median is the 50th percentile of a distribution, and the 1st/3rd quartiles are the 25th and 75th percentiles, respectively.
- For a number p between 0 and 1, the p **quantile** of a distribution is the value so that a proportion p of the data is less than or equal to that value.
  - The median is the 0.5 quantile of a distribution, and the 1st/3rd quartiles are the 0.25 and 0.75 quantiles, respectively.

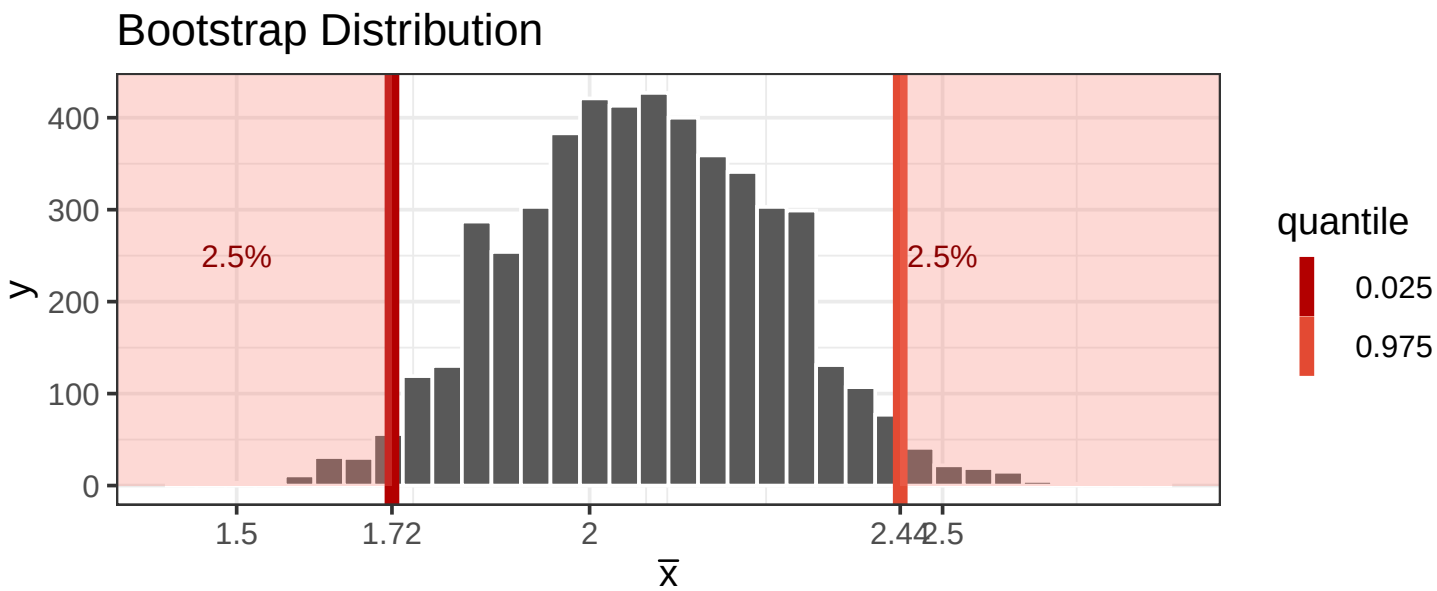
Review: Percentiles and Quantiles

- For a number k between 0 and 100, the kth **percentile** of a distribution is the value so that k% of the data is less than or equal to that value.
  - The median is the 50th percentile of a distribution, and the 1st/3rd quartiles are the 25th and 75th percentiles, respectively.
- For a number p between 0 and 1, the p **quantile** of a distribution is the value so that a proportion p of the data is less than or equal to that value.
  - The median is the 0.5 quantile of a distribution, and the 1st/3rd quartiles are the 0.25 and 0.75 quantiles, respectively.



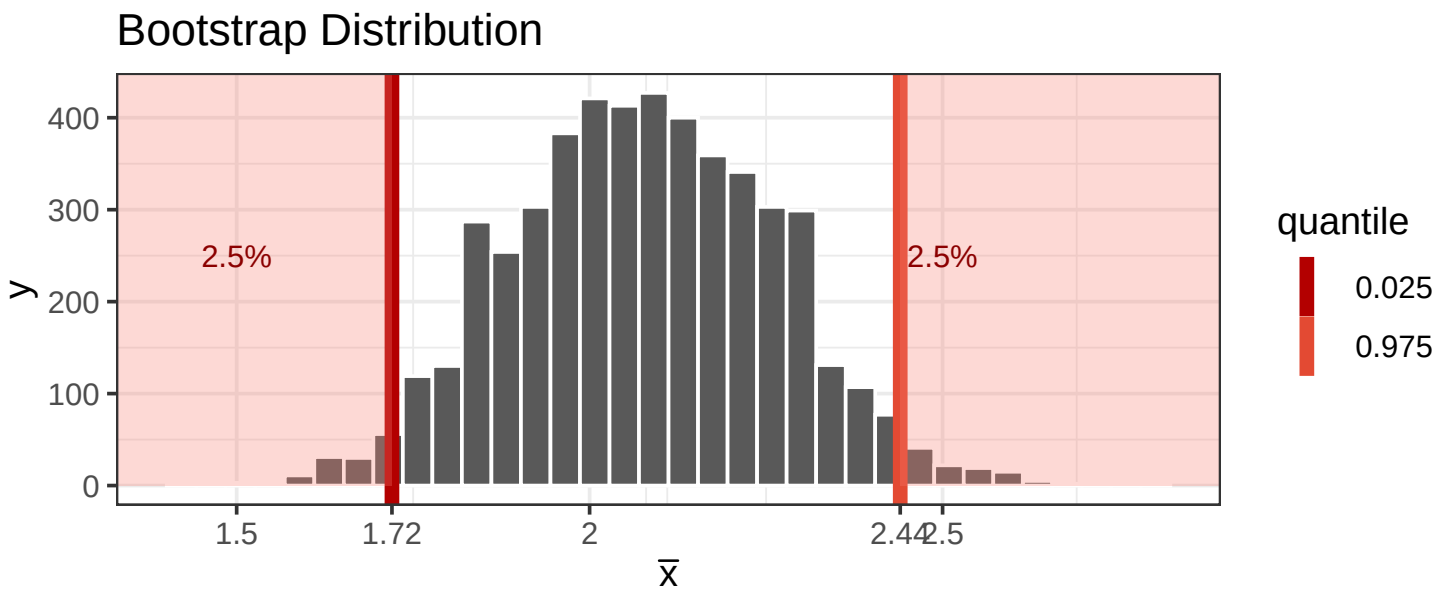
Quantiles and Percentiles

- By definition, 2.5% of the data is less than the .025 quantile, and 2.5% of the data is greater than the .975 quantile



Quantiles and Percentiles

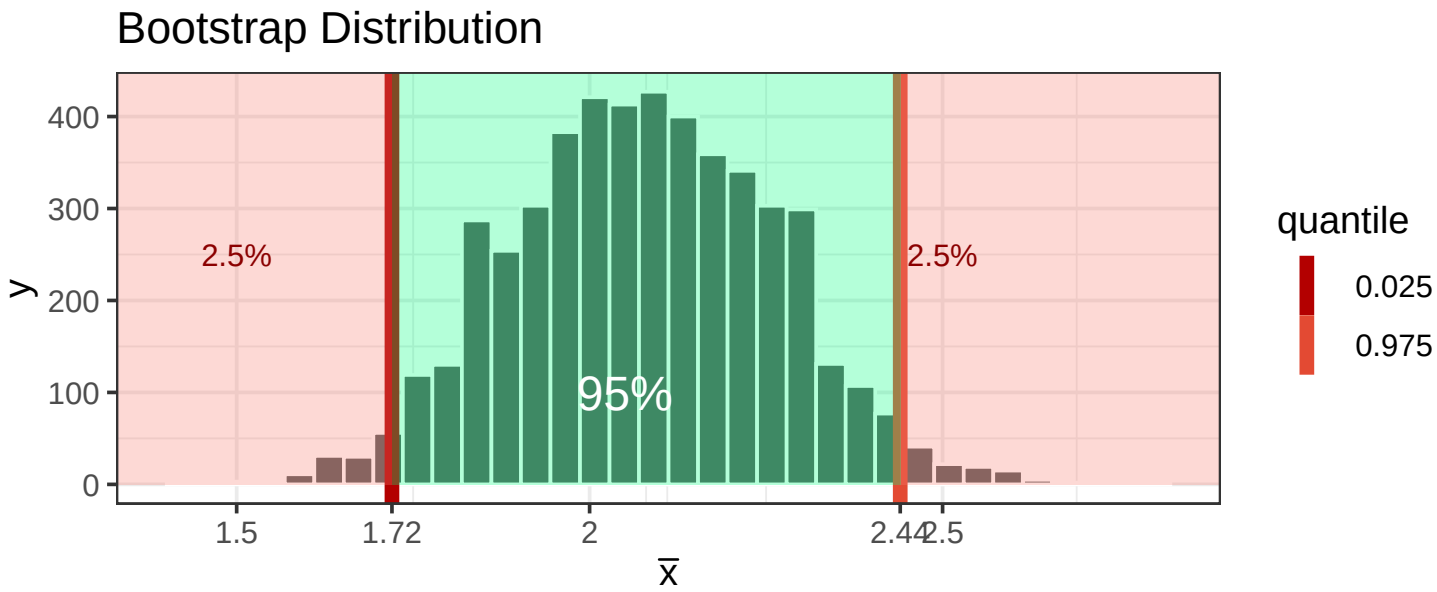
- By definition, 2.5% of the data is less than the .025 quantile, and 2.5% of the data is greater than the .975 quantile



- But this means that 95% of the data is between the .025 and the .975 quantiles

Quantiles and Percentiles

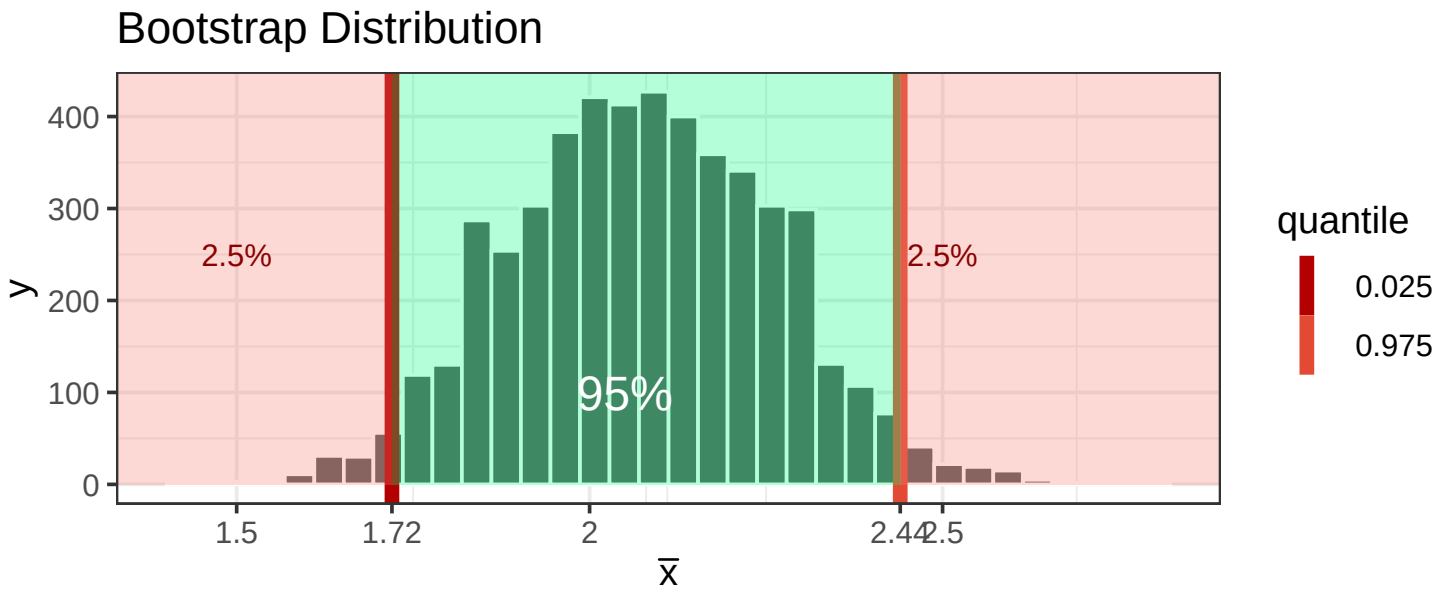
- By definition, 2.5% of the data is less than the .025 quantile, and 2.5% of the data is greater than the .975 quantile



- But this means that 95% of the data is between the .025 and the .975 quantiles

Quantiles and Percentiles

- By definition, 2.5% of the data is less than the .025 quantile, and 2.5% of the data is greater than the .975 quantile



- But this means that 95% of the data is between the .025 and the .975 quantiles
  - For a sampling distribution that is approximately bell-shaped, the .025 quantile is about $2 \cdot SE$ below the mean, and the .975 quantile is about $2 \cdot SE$ above the mean

The Percentile Method

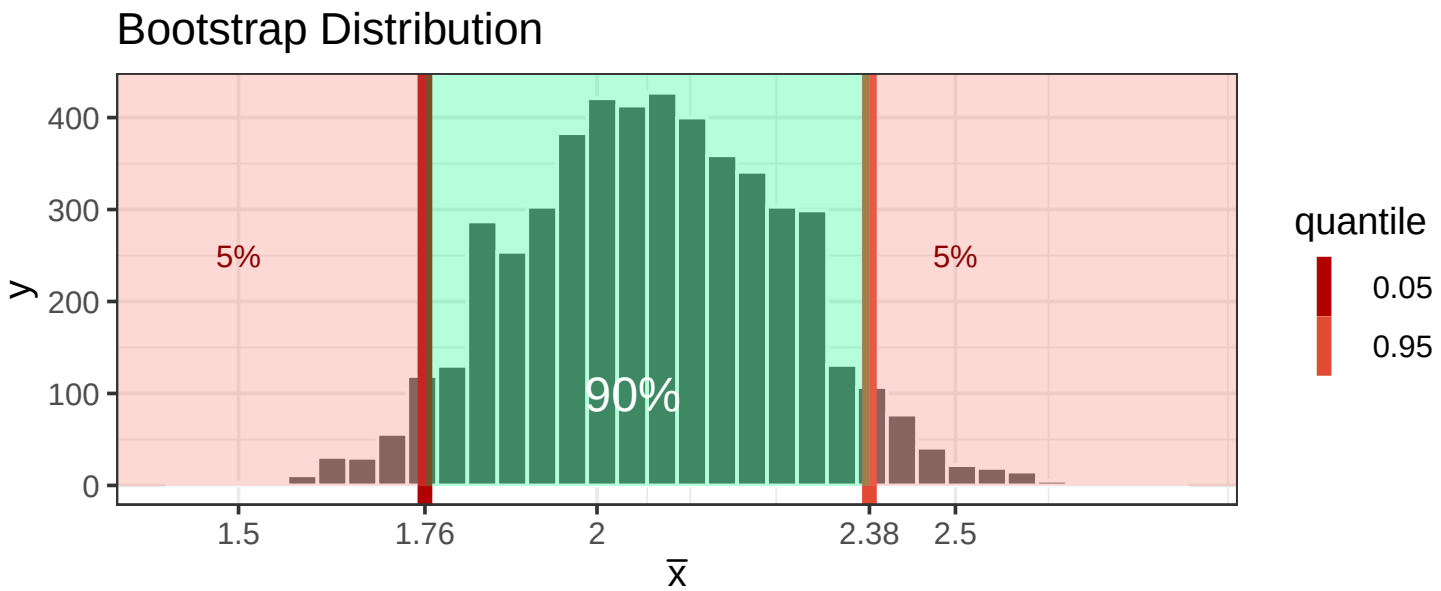
- Suppose we want to construct a 90% confidence interval for the reproduction rate

The Percentile Method

- Suppose we want to construct a 90% confidence interval for the reproduction rate
  - Instead of adding/subtracting $2 * SE$, find the 0.05 and .95 quantiles in the bootstrap distribution. Then 90% of bootstrap sample statistics will be between these values

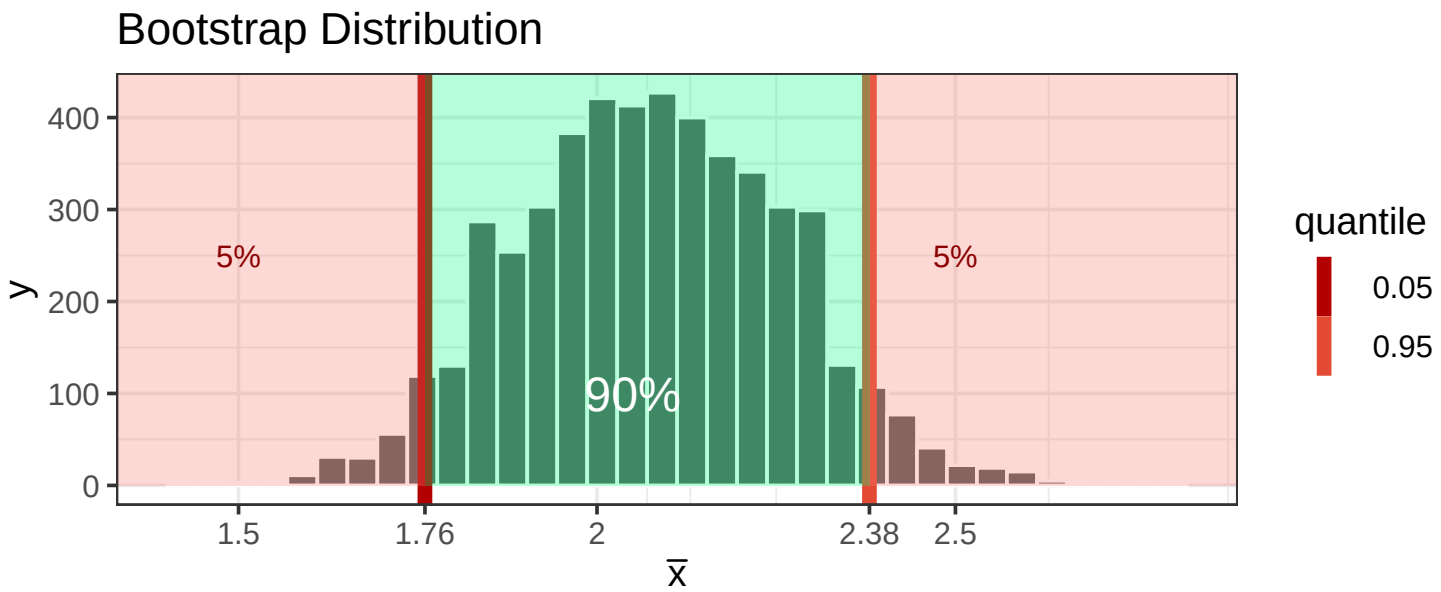
The Percentile Method

- Suppose we want to construct a 90% confidence interval for the reproduction rate
  - Instead of adding/subtracting $2 * SE$, find the 0.05 and .95 quantiles in the bootstrap distribution. Then 90% of bootstrap sample statistics will be between these values



The Percentile Method

- Suppose we want to construct a 90% confidence interval for the reproduction rate
  - Instead of adding/subtracting $2 * SE$, find the 0.05 and .95 quantiles in the bootstrap distribution. Then 90% of bootstrap sample statistics will be between these values



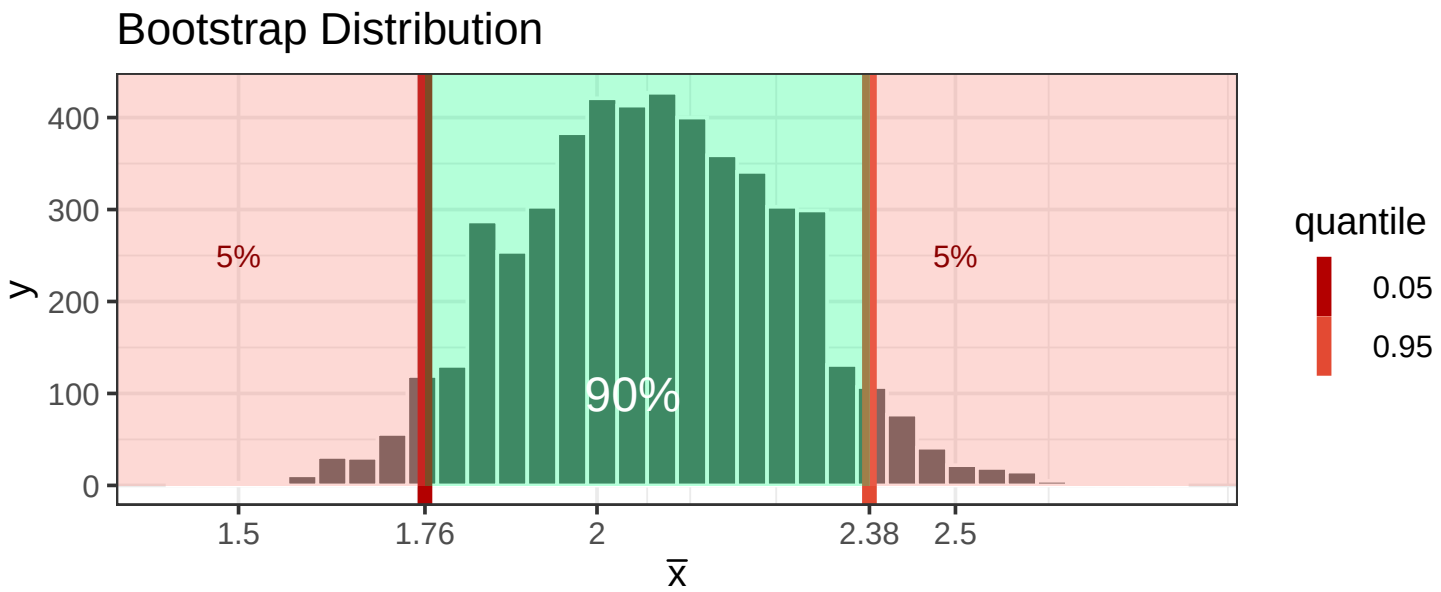
- We can use the `quantile` function in R to calculate the .05 and .95 quantiles

```
quantile(bootstrap_stats$x_bar, c(.05, .95))
```

```
##   5%  95%
## 1.76 2.38
```

The Percentile Method

- Suppose we want to construct a 90% confidence interval for the reproduction rate
  - Instead of adding/subtracting $2 * SE$, find the 0.05 and .95 quantiles in the bootstrap distribution. Then 90% of bootstrap sample statistics will be between these values



- Our 90% confidence interval is therefore 1.76 to 2.36

```
quantile(bootstrap_stats$x_bar, c(.05, .95))
```

```
##   5%  95%
## 1.76 2.38
```

Precision

How can we increase the precision of our confidence interval? That is, decrease the margin of error of the interval?

Precision

How can we increase the precision of our confidence interval? That is, decrease the margin of error of the interval?

- Increase sample size.
  - The standard deviation of the sampling distribution decreases as sample size increases. More sample means are closer to the true parameter

Precision

How can we increase the precision of our confidence interval? That is, decrease the margin of error of the interval?

- Increase sample size.
  - The standard deviation of the sampling distribution decreases as sample size increases. More sample means are closer to the true parameter
- Decrease confidence level.
  - The margin of error is determined by the percentiles. A 95% confidence interval is formed by the 2.5th and 97.5th percentiles in the bootstrap distribution.
  - Decreasing confidence level brings the percentiles closer to the 50th percentile, decreasing the width of the interval.

Precision

How can we increase the precision of our confidence interval? That is, decrease the margin of error of the interval?

- Increase sample size.
  - The standard deviation of the sampling distribution decreases as sample size increases. More sample means are closer to the true parameter
- Decrease confidence level.
  - The margin of error is determined by the percentiles. A 95% confidence interval is formed by the 2.5th and 97.5th percentiles in the bootstrap distribution.
  - Decreasing confidence level brings the percentiles closer to the 50th percentile, decreasing the width of the interval.
- Note that **accuracy** (i.e. success rate) $\neq$ **precision** (i.e. margin of error)

Precision

How can we increase the precision of our confidence interval? That is, decrease the margin of error of the interval?

- Increase sample size.
  - The standard deviation of the sampling distribution decreases as sample size increases. More sample means are closer to the true parameter
- Decrease confidence level.
  - The margin of error is determined by the percentiles. A 95% confidence interval is formed by the 2.5th and 97.5th percentiles in the bootstrap distribution.
  - Decreasing confidence level brings the percentiles closer to the 50th percentile, decreasing the width of the interval.
- Note that **accuracy** (i.e. success rate) $\neq$ **precision** (i.e. margin of error)
  - We can have confidence intervals with high precision and low accuracy, if we have a low confidence level.

Precision

How can we increase the precision of our confidence interval? That is, decrease the margin of error of the interval?

- Increase sample size.
  - The standard deviation of the sampling distribution decreases as sample size increases. More sample means are closer to the true parameter
- Decrease confidence level.
  - The margin of error is determined by the percentiles. A 95% confidence interval is formed by the 2.5th and 97.5th percentiles in the bootstrap distribution.
  - Decreasing confidence level brings the percentiles closer to the 50th percentile, decreasing the width of the interval.
- Note that **accuracy** (i.e. success rate) $\neq$ **precision** (i.e. margin of error)
  - We can have confidence intervals with high precision and low accuracy, if we have a low confidence level.
  - Similarly, we can have confidence intervals with low precision and high accuracy, if we use a high confidence level.