Confidence Intervals

Nate Wells

Math 141, 3/9/22

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Introduce confidence intervals as a method for estimating a parameter
- Use bootstrapping as means of creating confidence intervals
- Interpret confidence intervals

Section 1

Confidence Intervals

Point vs. Interval Estimates

- To estimate a population parameter, we can use a sample statistic.

Point vs. Interval Estimates

- To estimate a population parameter, we can use a sample statistic.
    - Suppose you are hosting a pizza party for 200 people, and need to know what proportion of vegetarian pizza to order.

Point vs. Interval Estimates

- To estimate a population parameter, we can use a sample statistic.
    - Suppose you are hosting a pizza party for 200 people, and need to know what proportion of vegetarian pizza to order.
    - You can ask a (random) sample of attendees about their pizza preference and use the proportion $\hat{p}$ in the sample as an estimate for the total proportion p.

Point vs. Interval Estimates

- To estimate a population parameter, we can use a sample statistic.
  - Suppose you are hosting a pizza party for 200 people, and need to know what proportion of vegetarian pizza to order.
  - You can ask a (random) sample of attendees about their pizza preference and use the proportion $\hat{p}$ in the sample as an estimate for the total proportion p.
- The sample statistic is your best guess for the population parameter, but it isn't the whole story

Point vs. Interval Estimates

- To estimate a population parameter, we can use a sample statistic.
  - Suppose you are hosting a pizza party for 200 people, and need to know what proportion of vegetarian pizza to order.
  - You can ask a (random) sample of attendees about their pizza preference and use the proportion $\hat{p}$ in the sample as an estimate for the total proportion p.
- The sample statistic is your best guess for the population parameter, but it isn't the whole story
  - Suppose after polling pizza party attendees, you find $\hat{p} = 0.33$. How certain should you be in this estimate?

Point vs. Interval Estimates

- To estimate a population parameter, we can use a sample statistic.
  - Suppose you are hosting a pizza party for 200 people, and need to know what proportion of vegetarian pizza to order.
  - You can ask a (random) sample of attendees about their pizza preference and use the proportion $\hat{p}$ in the sample as an estimate for the total proportion p.
- The sample statistic is your best guess for the population parameter, but it isn't the whole story
  - Suppose after polling pizza party attendees, you find $\hat{p} = 0.33$. How certain should you be in this estimate?
  - Your certainty might depend on the sample size. If $n = 9$, this value might not be too close to the true value. But if $n = 48$, it is probably much closer.

Point vs. Interval Estimates

- To estimate a population parameter, we can use a sample statistic.
  - Suppose you are hosting a pizza party for 200 people, and need to know what proportion of vegetarian pizza to order.
  - You can ask a (random) sample of attendees about their pizza preference and use the proportion $\hat{p}$ in the sample as an estimate for the total proportion p.
- The sample statistic is your best guess for the population parameter, but it isn't the whole story
  - Suppose after polling pizza party attendees, you find $\hat{p} = 0.33$. How certain should you be in this estimate?
  - Your certainty might depend on the sample size. If $n = 9$, this value might not be too close to the true value. But if $n = 48$, it is probably much closer.
- It might be preferable to estimate the proportion using a range of values, with smaller intervals corresponding to larger samples.

Point vs. Interval Estimates

- To estimate a population parameter, we can use a sample statistic.
  - Suppose you are hosting a pizza party for 200 people, and need to know what proportion of vegetarian pizza to order.
  - You can ask a (random) sample of attendees about their pizza preference and use the proportion $\hat{p}$ in the sample as an estimate for the total proportion p.
- The sample statistic is your best guess for the population parameter, but it isn't the whole story
  - Suppose after polling pizza party attendees, you find $\hat{p} = 0.33$. How certain should you be in this estimate?
  - Your certainty might depend on the sample size. If $n = 9$, this value might not be too close to the true value. But if $n = 48$, it is probably much closer.
- It might be preferable to estimate the proportion using a range of values, with smaller intervals corresponding to larger samples.
  - With just $n = 9$ people, you might give a range 0.03 to 0.63 for p.
  - But with $n = 48$, you might instead give the range 0.2 to 0.46.

Confidence Interval Estimates

A **confidence interval estimate** for a parameter takes the form

$$\text{Statistic} \ \pm \ \text{Margin of Error}$$

Confidence Interval Estimates

A **confidence interval estimate** for a parameter takes the form

$$\text{Statistic} \ \pm \ \text{Margin of Error}$$

- The confidence interval gives a range of plausible values for the parameter.

Confidence Interval Estimates

A **confidence interval estimate** for a parameter takes the form

$$\text{Statistic} \;\pm\; \text{Margin of Error}$$

- The confidence interval gives a range of plausible values for the parameter.
- The **Margin of Error** indicates the precision of our estimate

Confidence Interval Estimates

A **confidence interval estimate** for a parameter takes the form

$$\text{Statistic} \ \pm \ \text{Margin of Error}$$

- The confidence interval gives a range of plausible values for the parameter.
- The **Margin of Error** indicates the precision of our estimate
  - When sampling pizza preferences with $n = 48$, we estimate p using the interval

$$0.33 \pm 0.13 \qquad \text{or } 0.2 \text{ to } 0.46$$

Confidence Interval Estimates

A **confidence interval estimate** for a parameter takes the form

$$\text{Statistic} \pm \text{Margin of Error}$$

- The confidence interval gives a range of plausible values for the parameter.
- The **Margin of Error** indicates the precision of our estimate
  - When sampling pizza preferences with $n = 48$, we estimate p using the interval

$$0.33 \pm 0.13 \qquad \text{or } 0.2 \text{ to } 0.46$$

- We also report a success rate (or **confidence level**) for the estimation technique.

Confidence Interval Estimates

A **confidence interval estimate** for a parameter takes the form

$$\text{Statistic} \pm \text{Margin of Error}$$

- The confidence interval gives a range of plausible values for the parameter.
- The **Margin of Error** indicates the precision of our estimate
  - When sampling pizza preferences with $n = 48$, we estimate p using the interval

$$0.33 \pm 0.13 \qquad \text{or } 0.2 \text{ to } 0.46$$

- We also report a success rate (or **confidence level**) for the estimation technique.
- The confidence level corresponds to the proportion of sample statistics within the margin of error of the true parameter.

Confidence Interval Estimates

A **confidence interval estimate** for a parameter takes the form

$$\text{Statistic} \pm \text{Margin of Error}$$

- The confidence interval gives a range of plausible values for the parameter.
- The **Margin of Error** indicates the precision of our estimate
  - When sampling pizza preferences with $n = 48$, we estimate p using the interval

$$0.33 \pm 0.13 \qquad \text{or } 0.2 \text{ to } 0.46$$

- We also report a success rate (or **confidence level**) for the estimation technique.
- The confidence level corresponds to the proportion of sample statistics within the margin of error of the true parameter.
  - When sampling pizza preference with $n = 48$, we estimate p using the interval

$$0.33 \pm 0.13 \qquad \text{with confidence } 95\%.$$

Confidence Interval Estimates

A **confidence interval estimate** for a parameter takes the form

$$\text{Statistic} \pm \text{Margin of Error}$$

- The confidence interval gives a range of plausible values for the parameter.
- The **Margin of Error** indicates the precision of our estimate
  - When sampling pizza preferences with $n = 48$, we estimate p using the interval
    $$0.33 \pm 0.13 \qquad \text{or } 0.2 \text{ to } 0.46$$
- We also report a success rate (or **confidence level**) for the estimation technique.
- The confidence level corresponds to the proportion of sample statistics within the margin of error of the true parameter.
  - When sampling pizza preference with $n = 48$, we estimate p using the interval
    $$0.33 \pm 0.13 \qquad \text{with confidence } 95\%.$$
- To get the margin of error and the confidence level, we make use of the sampling distribution (or the bootstrap approximation).

The Sampling Distribution

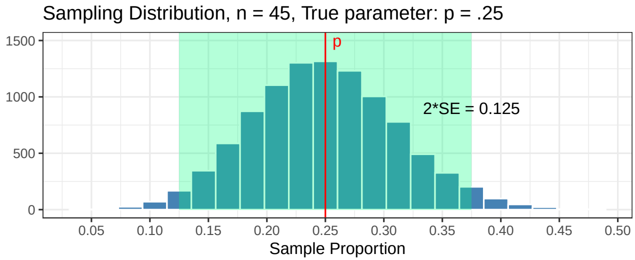
- Suppose that in truth, $p = .25$ (i.e. exactly 25% of all party attendees prefer vegetarian)

The Sampling Distribution

- Suppose that in truth, $p = .25$ (i.e. exactly 25% of all party attendees prefer vegetarian)
- For approximately bell-shaped sampling distributions, 95% of all sample statistics are within 2 SE of the parameter.

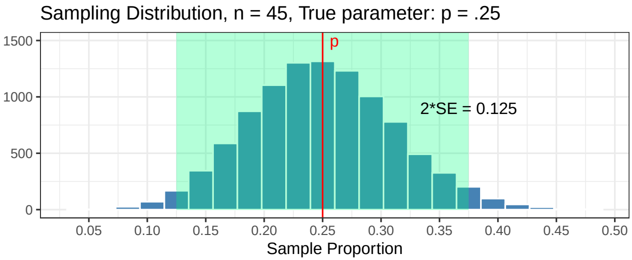
The Sampling Distribution

- Suppose that in truth, $p = .25$ (i.e. exactly 25% of all party attendees prefer vegetarian)
- For approximately bell-shaped sampling distributions, 95% of all sample statistics are within 2 SE of the parameter.



The Sampling Distribution

- Suppose that in truth, $p = .25$ (i.e. exactly 25% of all party attendees prefer vegetarian)
- For approximately bell-shaped sampling distributions, 95% of all sample statistics are within 2 SE of the parameter.



- But this means that for 95% of all samples, the true *parameter* will be within a distance of 2 SE of the sample *statistic* (every sample in the green region)

Interval Estimates

- To estimate the parameter, build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \mathrm{SE}$$

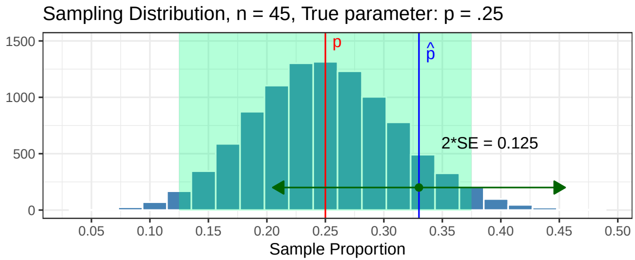
  - This interval will contain the parameter p for 95% of all samples

Interval Estimates

- To estimate the parameter, build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \mathrm{SE}$$

  - This interval will contain the parameter p for 95% of all samples



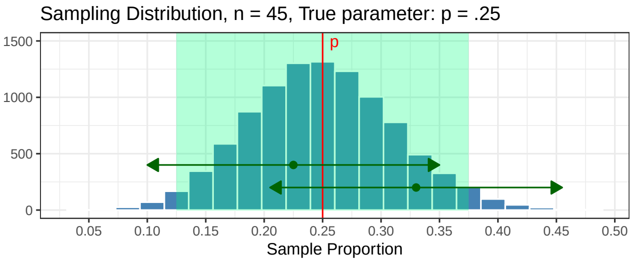
- The interval for our sample was $.33 \pm .125$, which *does* contain the parameter p

Interval Estimates

- To estimate the parameter, build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \mathrm{SE}$$

  - This interval will contain the parameter p for 95% of all samples



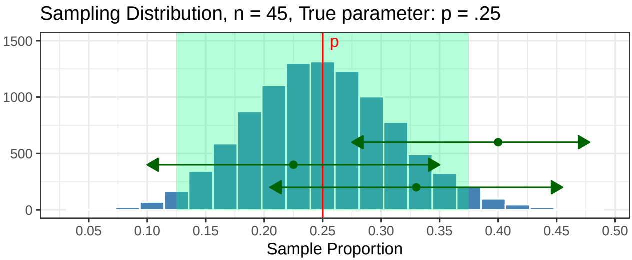
- Samples with $\hat{p}$ in the green region have intervals that also contain the parameter p

Interval Estimates

- To estimate the parameter, build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \mathrm{SE}$$

  - This interval will contain the parameter p for 95% of all samples



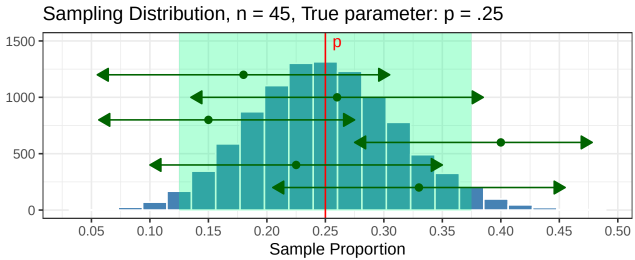
- Samples with $\hat{p}$ outside the green region have intervals that don't contain p

Interval Estimates

- To estimate the parameter, build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \mathrm{SE}$$

  - This interval will contain the parameter p for 95% of all samples



- But 95% of all samples will have intervals that do contain p

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.
- What does confidence mean?

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.
- What does confidence mean?
  - It gives the success rate for our method. For 95% of all possible samples, the interval we construct will actually contain the population parameter.

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.
- What does confidence mean?
  - It gives the success rate for our method. For 95% of all possible samples, the interval we construct will actually contain the population parameter.
- The problem?
  - We only have 1 sample, and we don't know if it belongs to the 95% of "good" samples, or the 5% of "bad" ones

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.
- What does confidence mean?
  - It gives the success rate for our method. For 95% of all possible samples, the interval we construct will actually contain the population parameter.
- The problem?
  - We only have 1 sample, and we don't know if it belongs to the 95% of "good" samples, or the 5% of "bad" ones
- The consolation?

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.
- What does confidence mean?
  - It gives the success rate for our method. For 95% of all possible samples, the interval we construct will actually contain the population parameter.
- The problem?
  - We only have 1 sample, and we don't know if it belongs to the 95% of "good" samples, or the 5% of "bad" ones
- The consolation?
  - If I go through my life constructing 95% confidence intervals, I will be telling the truth about 95% of the time (I'll take that rate!)

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.
- What does confidence mean?
  - It gives the success rate for our method. For 95% of all possible samples, the interval we construct will actually contain the population parameter.
- The problem?
  - We only have 1 sample, and we don't know if it belongs to the 95% of "good" samples, or the 5% of "bad" ones
- The consolation?
  - If I go through my life constructing 95% confidence intervals, I will be telling the truth about 95% of the time (I'll take that rate!)
- We do have more more problem:

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.
- What does confidence mean?
  - It gives the success rate for our method. For 95% of all possible samples, the interval we construct will actually contain the population parameter.
- The problem?
  - We only have 1 sample, and we don't know if it belongs to the 95% of "good" samples, or the 5% of "bad" ones
- The consolation?
  - If I go through my life constructing 95% confidence intervals, I will be telling the truth about 95% of the time (I'll take that rate!)
- We do have more more problem:
  - To make a confidence interval, we need the sampling distribution in order to compute the standard error. But in practice, we don't have direct access to this.

Confidence Level

- Confidence intervals consists of both an interval estimate and a confidence level.
  - Based on our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.2 and 4.7, with 95% confidence.
- What does confidence mean?
  - It gives the success rate for our method. For 95% of all possible samples, the interval we construct will actually contain the population parameter.
- The problem?
  - We only have 1 sample, and we don't know if it belongs to the 95% of "good" samples, or the 5% of "bad" ones
- The consolation?
  - If I go through my life constructing 95% confidence intervals, I will be telling the truth about 95% of the time (I'll take that rate!)
- We do have more more problem:
  - To make a confidence interval, we need the sampling distribution in order to compute the standard error. But in practice, we don't have direct access to this.
  - The solution is to approximate the sampling distribution via bootstrapping!

Section 2

Bootstrap Confidence Intervals

Reproduction Rate for Covid-19

Researchers are interested in the COVID-19 **reproduction rate** (the average number of individuals each infected person further infects)

- 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 COVID-19 **reproduction rate** (the average number of individuals each infected person further infects)

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

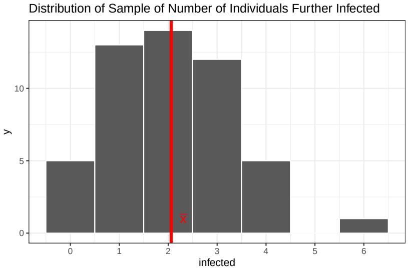
```
##   infected  n
## 1        0  5
## 2        1 13
## 3        2 14
## 4        3 12
## 5        4  5
## 6        6  1
##   mean_infected
## 1          2.06
```

Reproduction Rate for Covid-19

Researchers are interested in the COVID-19 **reproduction rate** (the average number of individuals each infected person further infects)

- 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
```



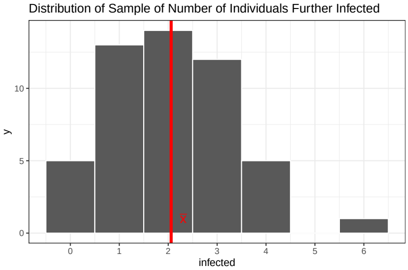
- Is the true reproduction rate exactly 2.06?

Reproduction Rate for Covid-19

Researchers are interested in the COVID-19 **reproduction rate** (the average number of individuals each infected person further infects)

- 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
```



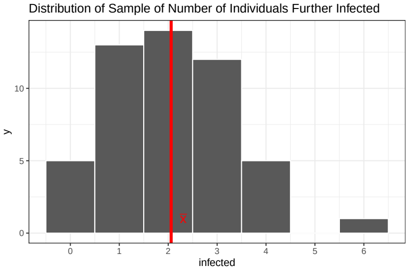
- Is the true reproduction rate exactly 2.06?
  - Surely not! This is just one sample of size 50

Reproduction Rate for Covid-19

Researchers are interested in the COVID-19 **reproduction rate** (the average number of individuals each infected person further infects)

- 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:

```
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 2
## # Groups:   replicate [5,000]
##    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 249,990 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 2
## # Groups:   replicate [5,000]
##    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 249,990 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
```

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 2
## # Groups:   replicate [5,000]
##    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 249,990 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
```

- 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(infected))
```

Bootstrap Reproduction Rate

Compute bootstrap statistics:

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

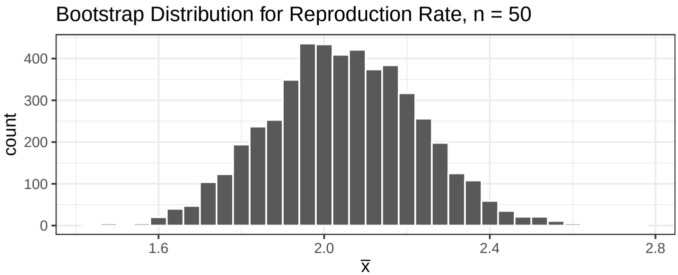
```
bootstrap_stats
```

```
## # A tibble: 5,000 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 4,990 more rows
```

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

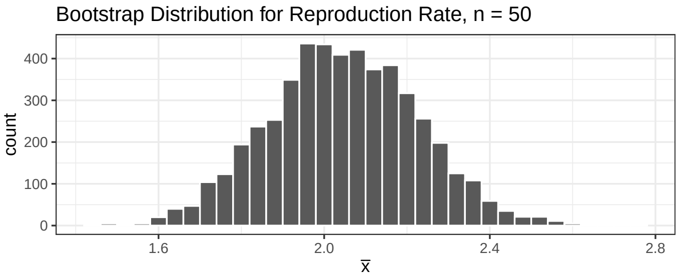
Bootstrap Reproduction Rate

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 0.181
```

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 $SE = 0.181$.

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 $SE = 0.181$.
- Our 95% confidence interval for the true reproduction rate of COVID-19 is

$$2.06 \pm 2 \cdot 0.181$$

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 $SE = 0.181$.
- Our 95% confidence interval for the true reproduction rate of COVID-19 is

$$2.06 \pm 2 \cdot 0.181$$

- Our best guess for the reproduction rate is between 1.698 and 2.422. 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 $SE = 0.181$.
- Our 95% confidence interval for the true reproduction rate of COVID-19 is

$$2.06 \pm 2 \cdot 0.181$$

- Our best guess for the reproduction rate is between 1.698 and 2.422. 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

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

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 5% 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 5% 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

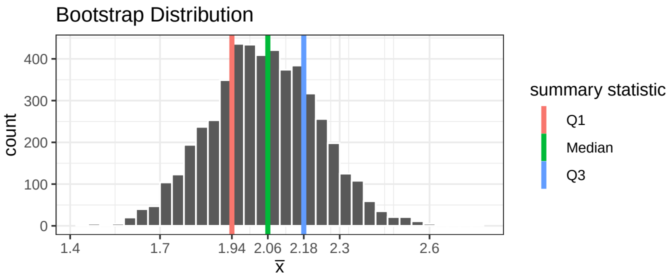
- For a number k between 0 and 100, the kth **percentile** of a distribution is the value so that 5% of the data is less than or equal to that value.
  - The median is the 50th percentile of a distribution, and the 1st/3rd quartiles are the 25th and 75th percentiles, respectively.
- For a number p between 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 5% 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.

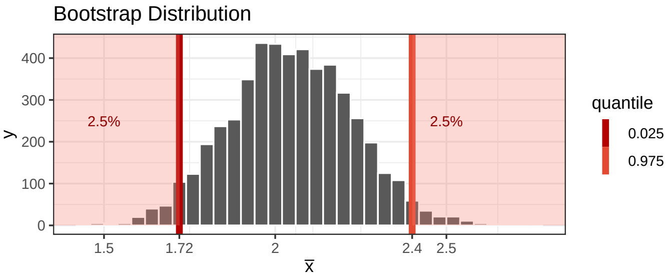
Review: Percentiles and Quantiles

- For a number k between 0 and 100, the kth **percentile** of a distribution is the value so that 5% 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.



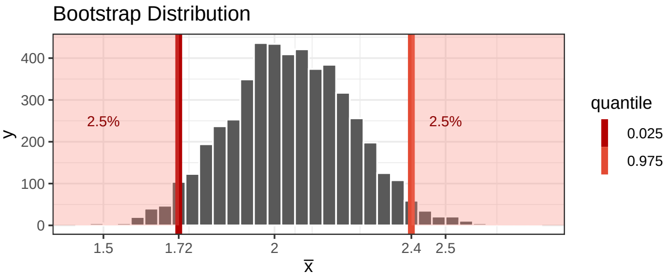
Quantiles and Percentiles

- 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

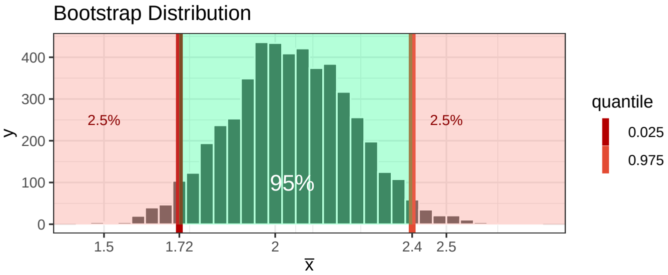
- By definition, 2.5% of the data is less than the .025 quantile, and 2.5% of the data is greater than the .975 quantile



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

Quantiles and Percentiles

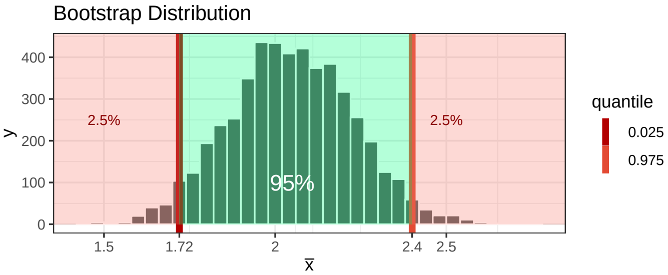
- By definition, 2.5% of the data is less than the .025 quantile, and 2.5% of the data is greater than the .975 quantile



- 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

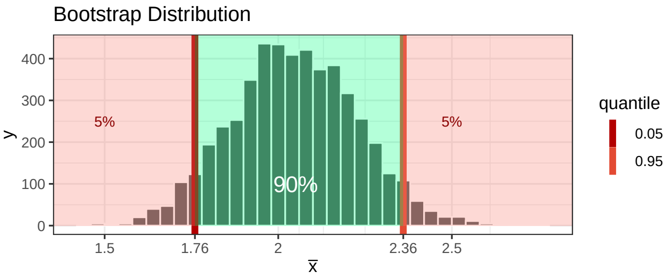
- Suppose we want to construct a 90% confidence interval for the reproduction rate

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

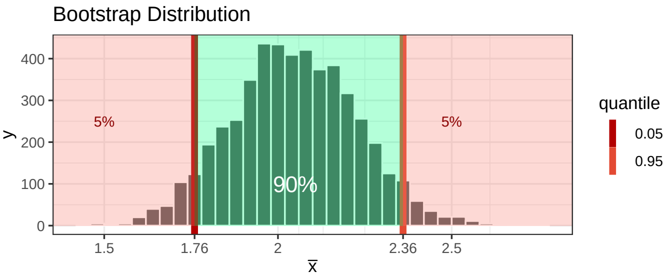
The Percentile Method

- 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



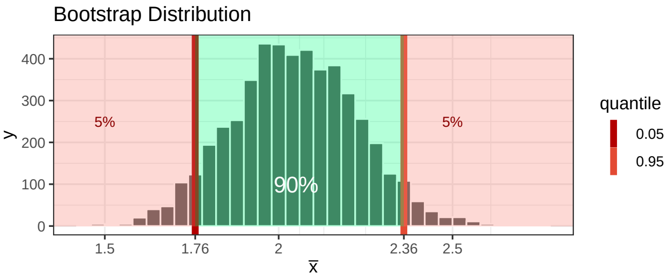
- We can use the `quantile` function in R to calculate the .05 and .95 quantiles

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

```
##   5%  95%
## 1.76 2.36
```

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 2.36
```

Precision

How can we increase the precision of our confidence interval (i.e. decrease the margin of error)?

Precision

How can we increase the precision of our confidence interval (i.e. decrease the margin of error)?

- 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 (i.e. decrease the margin of error)?

- 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.

Section 3

Confidence Interval Misunderstandings

Common Confidence Interval Misunderstandings

Suppose we wish to estimate the number of hours a Reed student sleeps on a typical night. We obtain the following 95% confidence interval: $(7.86, 8.34)$

Common Confidence Interval Misunderstandings

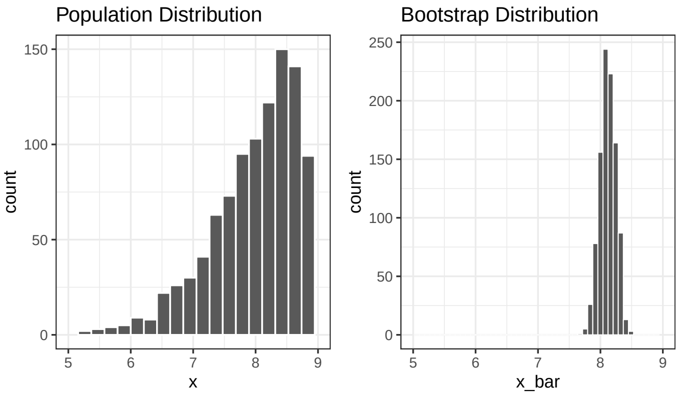
Suppose we wish to estimate the number of hours a Reed student sleeps on a typical night. We obtain the following 95% confidence interval: $(7.86, 8.34)$

1. A 95% confidence interval **does not** contain 95% of observations in the population.

Common Confidence Interval Misunderstandings

Suppose we wish to estimate the number of hours a Reed student sleeps on a typical night. We obtain the following 95% confidence interval: $(7.86, 8.34)$

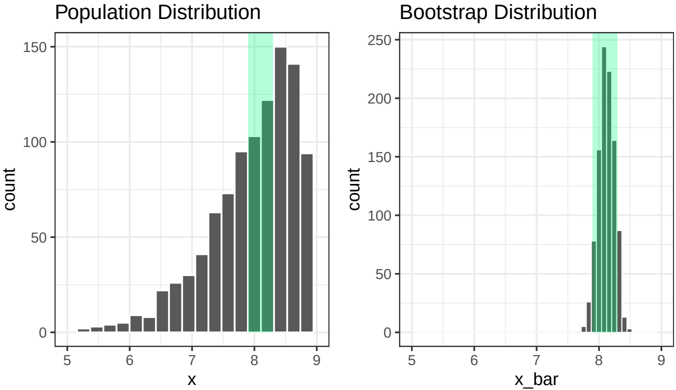
1. A 95% confidence interval **does not** contain 95% of observations in the population.



Common Confidence Interval Misunderstandings

Suppose we wish to estimate the number of hours a Reed student sleeps on a typical night. We obtain the following 95% confidence interval: $(7.86, 8.34)$

1. A 95% confidence interval **does not** contain 95% of observations in the population.

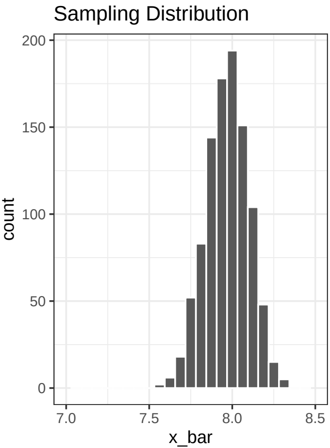


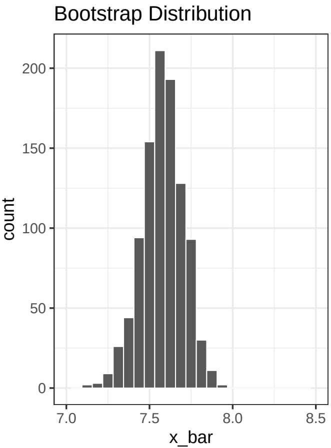
Common Confidence Interval Misunderstandings

2. A 95% confidence interval **does not** mean that 95% of all sample means fall within the given range.

Common Confidence Interval Misunderstandings

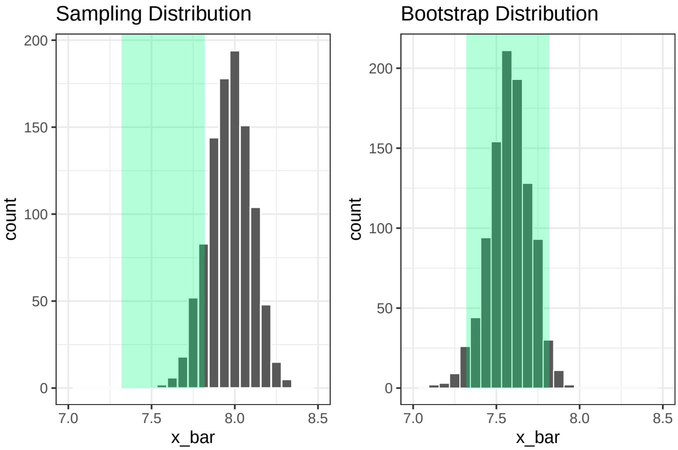
2. A 95% confidence interval **does not** mean that 95% of all sample means fall within the given range.





Common Confidence Interval Misunderstandings

2. A 95% confidence interval **does not** mean that 95% of all sample means fall within the given range.



Common Confidence Interval Misunderstandings

3. A 95% confidence interval **does not** mean that there is a 95% chance that the true parameter falls in the given range.

Common Confidence Interval Misunderstandings

3. A 95% confidence interval **does not** mean that there is a 95% chance that the true parameter falls in the given range.

- Once a random sample has been observed and the confidence interval calculated, there is no more randomness in the process. We cannot make probabilistic statements about the outcome.

Common Confidence Interval Misunderstandings

3. A 95% confidence interval **does not** mean that there is a 95% chance that the true parameter falls in the given range.

- Once a random sample has been observed and the confidence interval calculated, there is no more randomness in the process. We cannot make probabilistic statements about the outcome.
  - At this point, the interval either does or does not contain the fixed (but unknown) parameter

Common Confidence Interval Misunderstandings

3. A 95% confidence interval **does not** mean that there is a 95% chance that the true parameter falls in the given range.

- Once a random sample has been observed and the confidence interval calculated, there is no more randomness in the process. We cannot make probabilistic statements about the outcome.
  - At this point, the interval either does or does not contain the fixed (but unknown) parameter
  - One sample (of 10000) had a sample mean of 4.9 and produced a confidence interval of $(4.6, 5.2)$.

Common Confidence Interval Misunderstandings

3. A 95% confidence interval **does not** mean that there is a 95% chance that the true parameter falls in the given range.

- Once a random sample has been observed and the confidence interval calculated, there is no more randomness in the process. We cannot make probabilistic statements about the outcome.
  - At this point, the interval either does or does not contain the fixed (but unknown) parameter
  - One sample (of 10000) had a sample mean of 4.9 and produced a confidence interval of $(4.6, 5.2)$.
  - Based on what you know about sleep patterns, do you think there is a 95% chance this interval contains the true parameter?