Bootstrapping Confidence Intervals (with Confidence!)

Aaron Kaufman

November 16, 2018

Previously: Calculating Confidence Intervals

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty,

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval:

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Procedure:

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Procedure:

1. Collect a sample of data from the population

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Procedure:

1. Collect a sample of data from the population
2. Calculate a quantity of interest, **X**

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Procedure:

1. Collect a sample of data from the population
2. Calculate a quantity of interest, **X**
3. Calculate a standard deviation, s

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Procedure:

1. Collect a sample of data from the population
2. Calculate a quantity of interest, **X**
3. Calculate a standard deviation, s
4. Calculate a standard **error** σ: $\frac{s}{\sqrt{n}}$

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Procedure:

1. Collect a sample of data from the population
2. Calculate a quantity of interest, **X**
3. Calculate a standard deviation, s
4. Calculate a standard **error** σ: $\frac{s}{\sqrt{n}}$
5. Use a Z-table to figure out how many SEs you need

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Procedure:

1. Collect a sample of data from the population
2. Calculate a quantity of interest, **X**
3. Calculate a standard deviation, s
4. Calculate a standard **error** σ: $\frac{s}{\sqrt{n}}$
5. Use a Z-table to figure out how many SEs you need (1.96 for 95%, 2.58 for 99%)

Previously: Calculating Confidence Intervals

Statistics: science of uncertainty, and CIs are our most important tool

95% Confidence Interval: If we construct 100 such intervals, 95% will contain the truth

Procedure:

1. Collect a sample of data from the population
2. Calculate a quantity of interest, **X**
3. Calculate a standard deviation, s
4. Calculate a standard **error** σ: $\frac{s}{\sqrt{n}}$
5. Use a Z-table to figure out how many SEs you need (1.96 for 95%, 2.58 for 99%)
6. For 95%: $[X - 1.96 \times \sigma, X + 1.96 \times \sigma]$

Previously: Calculating Confidence Intervals

Previously: Calculating Confidence Intervals

Data:

Example:

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example:

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example: What percentage of adult US residents are registered to vote?

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example: What percentage of adult US residents are registered to vote?

1. **Average:** 85.7%

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example: What percentage of adult US residents are registered to vote?

1. **Average:** 85.7%
2. **Standard Deviation:** 35.0%

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example: What percentage of adult US residents are registered to vote?

1. **Average:** 85.7%
2. **Standard Deviation:** 35.0%
3. **Standard Error:** $\frac{34.9}{\sqrt{4{,}142}} = 0.5\%$

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example: What percentage of adult US residents are registered to vote?

1. **Average:** 85.7%
2. **Standard Deviation:** 35.0%
3. **Standard Error:** $\frac{34.9}{\sqrt{4{,}142}} = 0.5\%$
4. **95% Confidence Interval:** $[85.7 - 1.96 \times 0.5, 85.7 + 1.96 \times 0.5] =$ [84.6%, 86.8%]

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example: What percentage of adult US residents are registered to vote?

1. **Average:** 85.7%
2. **Standard Deviation:** 35.0%
3. **Standard Error:** $\frac{34.9}{\sqrt{4{,}142}} = 0.5\%$
4. **95% Confidence Interval:** $[85.7 - 1.96 \times 0.5, 85.7 + 1.96 \times 0.5] =$ [84.6%, 86.8%]

Interpretation?

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example: What percentage of adult US residents are registered to vote?

1. **Average:** 85.7%
2. **Standard Deviation:** 35.0%
3. **Standard Error:** $\frac{34.9}{\sqrt{4{,}142}} = 0.5\%$
4. **95% Confidence Interval:** $[85.7 - 1.96 \times 0.5, 85.7 + 1.96 \times 0.5] =$ [84.6%, 86.8%]

Interpretation? If we take 100 samples of 4,142 voters and construct 95% CIs, then 95% of them will contain the true percentage of (self-reported) registered voters

Previously: Calculating Confidence Intervals

Data: 2016 American National Election Study (4,142 US adults)

Example: What percentage of adult US residents are registered to vote?

1. **Average:** 85.7%
2. **Standard Deviation:** 35.0%
3. **Standard Error:** $\frac{34.9}{\sqrt{4{,}142}} = 0.5\%$
4. **95% Confidence Interval:** $[85.7 - 1.96 \times 0.5, 85.7 + 1.96 \times 0.5] =$ [84.6%, 86.8%]

Interpretation? If we take 100 samples of 4,142 voters and construct 95% CIs, then 95% of them will contain the true percentage of (self-reported) registered voters

This is a pain!

An Alternative Approach: The Bootstrap

An Alternative Approach: The Bootstrap

Bootstrap:

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:**

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*
2. **Recalculate:**

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*
2. **Recalculate:** recalculate the original statistic of interest *for each resample*

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*
2. **Recalculate:** recalculate the original statistic of interest *for each resample*
3. **Find Quantiles:**

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*
2. **Recalculate:** recalculate the original statistic of interest *for each resample*
3. **Find Quantiles:** take the 2.5% and 97.5% quantiles of the bootstrapped statistics

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*
2. **Recalculate:** recalculate the original statistic of interest *for each resample*
3. **Find Quantiles:** take the 2.5% and 97.5% quantiles of the bootstrapped statistics

Why sample with replacement?

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*
2. **Recalculate:** recalculate the original statistic of interest *for each resample*
3. **Find Quantiles:** take the 2.5% and 97.5% quantiles of the bootstrapped statistics

Why sample with replacement?

- We need variation in the quantity of interest

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*
2. **Recalculate:** recalculate the original statistic of interest *for each resample*
3. **Find Quantiles:** take the 2.5% and 97.5% quantiles of the bootstrapped statistics

Why sample with replacement?

- We need variation in the quantity of interest
- Sampling without replacement

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

1. **Resample:** resample the original data *with replacement*
2. **Recalculate:** recalculate the original statistic of interest *for each resample*
3. **Find Quantiles:** take the 2.5% and 97.5% quantiles of the bootstrapped statistics

Why sample with replacement?

- We need variation in the quantity of interest
- Sampling without replacement $\rightsquigarrow$ identical samples

An Alternative Approach: The Bootstrap

Bootstrap: A computational solution to confidence intervals (much less math)

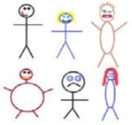
1. **Resample:** resample the original data *with replacement*
2. **Recalculate:** recalculate the original statistic of interest *for each resample*
3. **Find Quantiles:** take the 2.5% and 97.5% quantiles of the bootstrapped statistics

Why sample with replacement?

- We need variation in the quantity of interest
- Sampling without replacement $\rightsquigarrow$ identical samples $\rightsquigarrow$ same QOI every time!

The Bootstrap: Terminology

The Bootstrap: Terminology

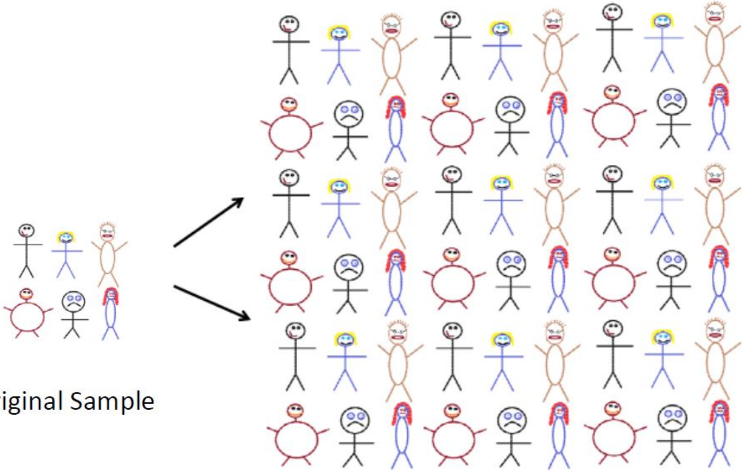


Original Sample

The Bootstrap: Terminology

What is the CI for the average height?

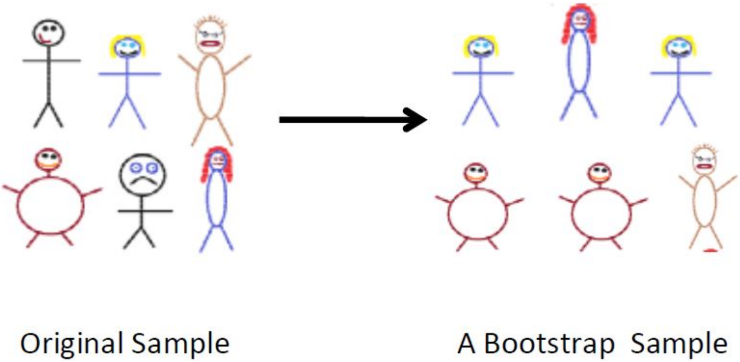
The Bootstrap: Terminology



Original Sample

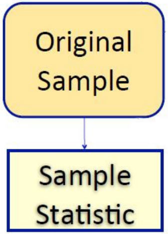
A simulated “population” to sample from

The Bootstrap: Terminology

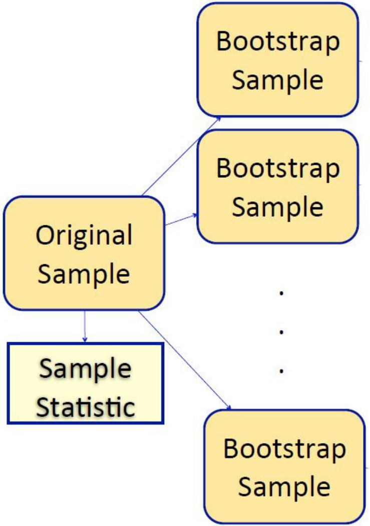


The Bootstrap: Terminology

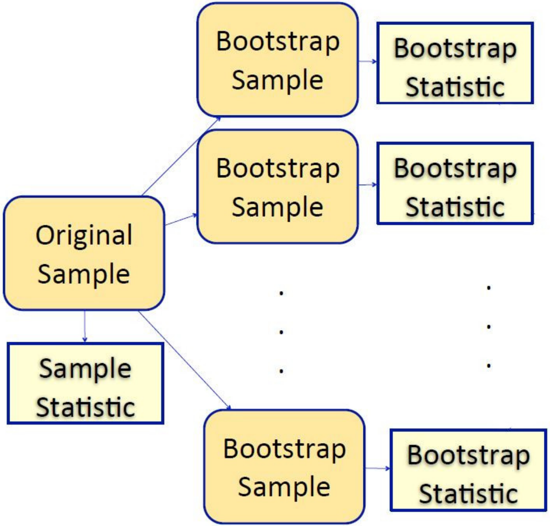
The Bootstrap: Terminology



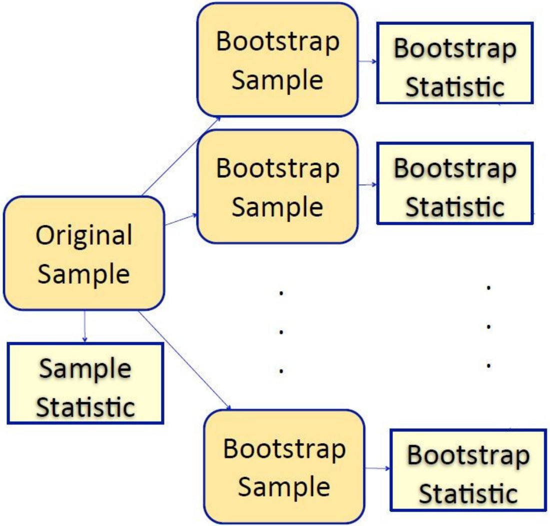
The Bootstrap: Terminology



The Bootstrap: Terminology

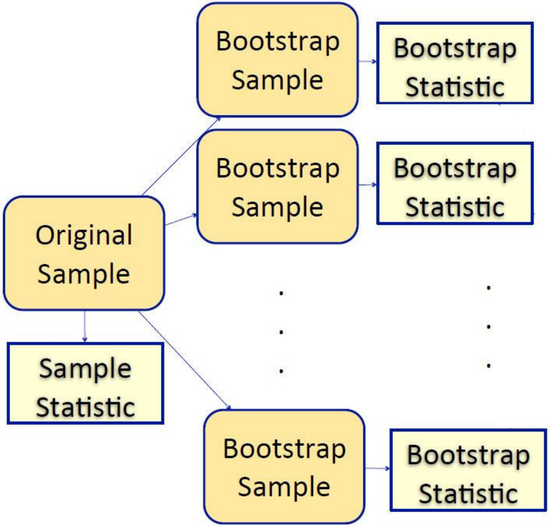


The Bootstrap: Terminology



86.3%

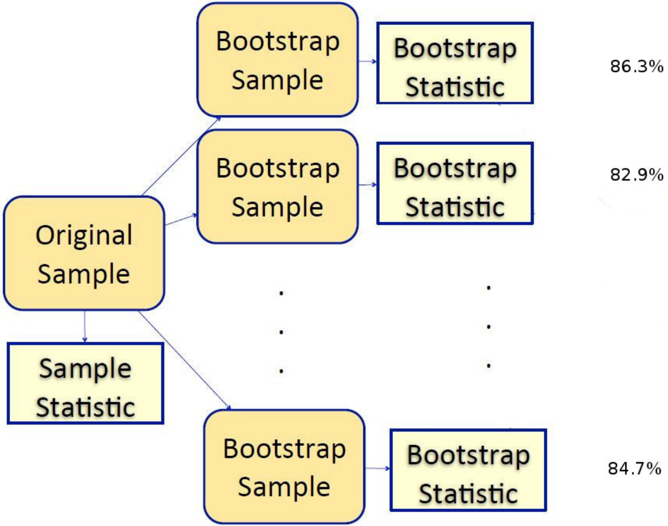
The Bootstrap: Terminology



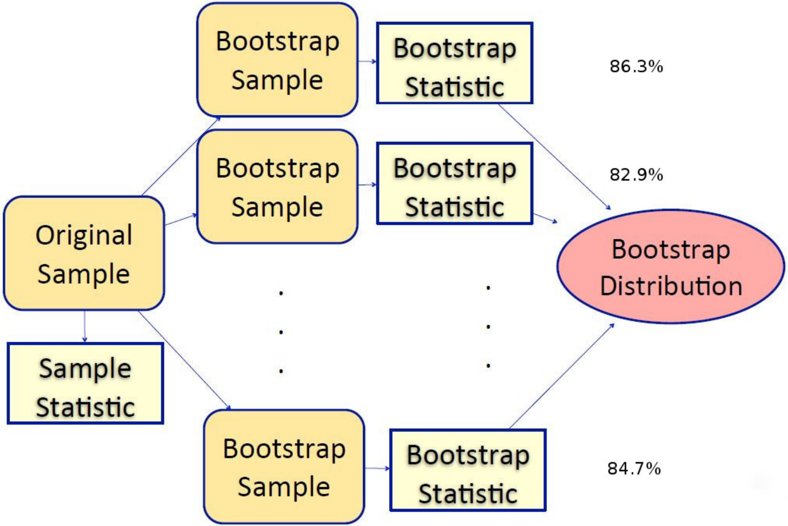
86.3%

82.9%

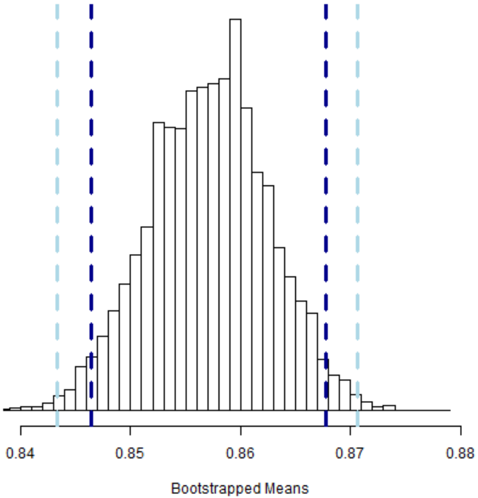
The Bootstrap: Terminology



The Bootstrap: Terminology



The Bootstrap: Terminology



R Coding: The American National Election Study

R Coding: The American National Election Study

```
X = mean(anes16$registered)
X # 0.857
```

R Coding: The American National Election Study

```
X = mean(anes16$registered)
X # 0.857

SD = sd(anes16$registered)
SD # 0.350
```

R Coding: The American National Election Study

```
X = mean(anes16$registered)
X # 0.857

SD = sd(anes16$registered)
SD # 0.350

SE = SD/sqrt(nrow(anes16))
SE # 0.005
```

R Coding: The American National Election Study

```
X = mean(anes16$registered)
X # 0.857

SD = sd(anes16$registered)
SD # 0.350

SE = SD/sqrt(nrow(anes16))
SE # 0.005

X + 1.96*SE # 0.868
X - 1.96*SE # 0.848
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
mean(bootstrapped$registered)
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
mean(bootstrapped$registered)
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
mean(bootstrapped$registered)
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
mean(bootstrapped$registered)
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
mean(bootstrapped$registered)
```

R Coding: The American National Election Study

```
bootstrapped.distribution = replicate(100, {
  bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
  mean(bootstrapped$registered)
})
```

R Coding: The American National Election Study

```
bootstrapped.distribution = replicate(100, {
   bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
   mean(bootstrapped$registered)
})
```

R Coding: The American National Election Study

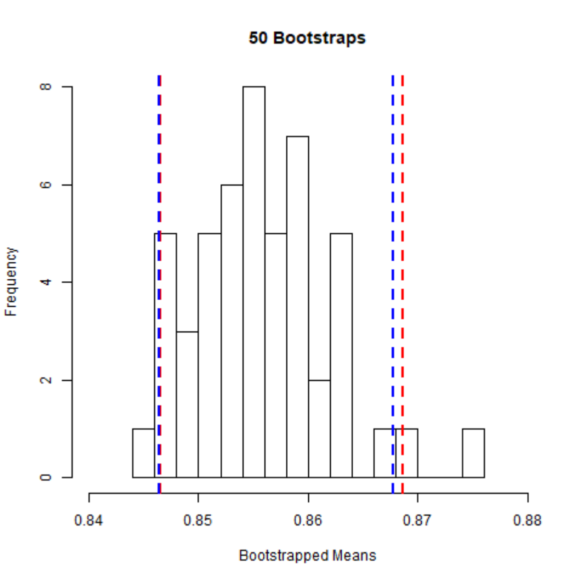
```
bootstrapped.distribution = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
    mean(bootstrapped$registered)
})
quants = quantile(bootstrapped.distribution, c(0.025, 0.975))
quants # 0.846, 0.867
```

R Coding: The American National Election Study

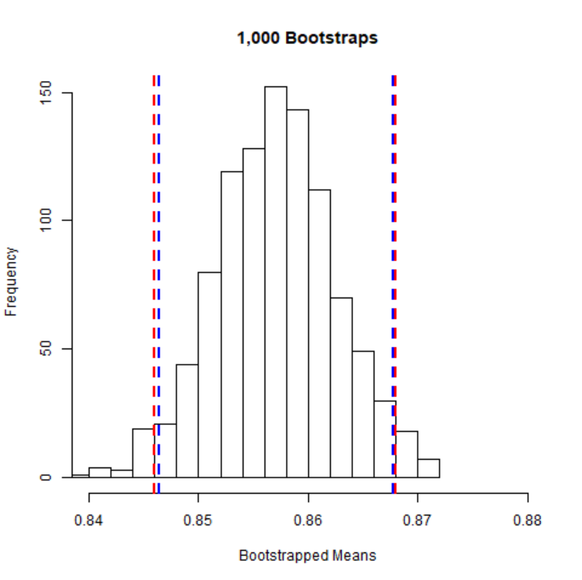
```
bootstrapped.distribution = replicate(100, {
   bootstrapped.sample = mosaic::shuffle(anes16, replace=TRUE)
   mean(bootstrapped$registered)
})
quants = quantile(bootstrapped.distribution, c(0.025, 0.975))
quants # 0.846, 0.867
```

How many bootstrapped samples do we generate?

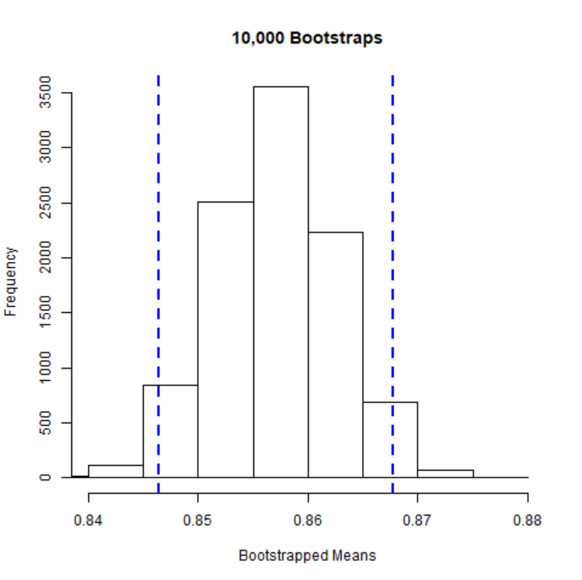
R Coding: The American National Election Study



R Coding: The American National Election Study



R Coding: The American National Election Study



R Coding: The American National Election Study

But we can also bootstrap more complicated stuff!

```
cor(anes16$age, anes16$registered) # 0.199
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16,replace=TRUE)
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16,replace=TRUE)
cor(bootstrapped.sample$registered, bootstrapped.sample$age)
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16,replace=TRUE)
cor(bootstrapped.sample$registered, bootstrapped.sample$age)
```

R Coding: The American National Election Study

```
bootstrapped.sample = mosaic::shuffle(anes16,replace=TRUE)
cor(bootstrapped.sample$registered, bootstrapped.sample$age)
```

R Coding: The American National Election Study

```
bootstrapped.corr = replicate(100, {
   bootstrapped.sample = mosaic::shuffle(anes16,replace=TRUE)
   cor(bootstrapped.sample$registered, bootstrapped.sample$age)
})
```

R Coding: The American National Election Study

```
bootstrapped.corr = replicate(100, {
   bootstrapped.sample = mosaic::shuffle(anes16,replace=TRUE)
   cor(bootstrapped.sample$registered, bootstrapped.sample$age)
})
corrs = quantile(bootstrapped.corr, c(0.025, 0.975))
corrs # 0.173, 0.230
```

Even more complicated stuff

Even more complicated stuff

How do they “call” an election? How do they call elections with only 1% of the data?

Even more complicated stuff

How do they "call" an election? How do they call elections with only 1% of the data?

- Build a model to predict the election

Even more complicated stuff

How do they "call" an election? How do they call elections with only 1% of the data?

- Build a model to predict the election
- How do calculate uncertainty?

Even more complicated stuff

How do they “call” an election? How do they call elections with only 1% of the data?

- Build a model to predict the election
- How do calculate uncertainty?
- **Bootstrap!**

Even more complicated stuff

When can we "call" an election?

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

$$\text{bootstrapped.corr} = \text{replicate}(100, \{$$

$$\})$$

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)




})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)




})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
        bootstrapped.sample)



})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
      bootstrapped.sample)



})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
        bootstrapped.sample)



})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
        bootstrapped.sample)



})
```

Even more complicated stuff

When can we “call” an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
        bootstrapped.sample)



})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
        bootstrapped.sample)
    preds = predict(m, newdata = data.frame(VoteSoFar = 0.65,
        DemPct16 = 0.75, pctWhite = 0.4)
})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
      bootstrapped.sample)
    preds = predict(m, newdata = data.frame(VoteSoFar = 0.65,
      DemPct16 = 0.75, pctWhite = 0.4)
})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
        bootstrapped.sample)
    preds = predict(m, newdata = data.frame(VoteSoFar = 0.65,
        DemPct16 = 0.75, pctWhite = 0.4)
})
```

Even more complicated stuff

When can we “call” an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
        bootstrapped.sample)
    preds = predict(m, newdata = data.frame(VoteSoFar = 0.65,
        DemPct16 = 0.75, pctWhite = 0.4)
})
```

Even more complicated stuff

When can we "call" an election? When 99% CI only contains one winner

```
bootstrapped.corr = replicate(100, {
    bootstrapped.sample = mosaic::shuffle(election.returns,replace=TRUE)
    model = lm(DemWins18 ~ VoteSoFar + DemPct16 + pctWhite,
        bootstrapped.sample)
    preds = predict(m, newdata = data.frame(VoteSoFar = 0.65,
        DemPct16 = 0.75, pctWhite = 0.4)
})
```

Summary & Coding tips

Summary & Coding tips

The bootstrap...

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too:

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations,

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations, predictions,

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations, predictions, modeling accuracy...

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations, predictions, modeling accuracy...

To make it easier...

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations, predictions, modeling accuracy...

To make it easier...

- Use the `replicate()` function to repeat your resampling lots of times

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations, predictions, modeling accuracy...

To make it easier...

- Use the `replicate()` function to repeat your resampling lots of times
- Use the `shuffle()` function from the `mosaic` library:

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations, predictions, modeling accuracy...

To make it easier...

- Use the `replicate()` function to repeat your resampling lots of times
- Use the `shuffle()` function from the `mosaic` library: `mosaic::shuffle(dat, replace=FALSE)`

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations, predictions, modeling accuracy...

To make it easier...

- Use the `replicate()` function to repeat your resampling lots of times
- Use the `shuffle()` function from the `mosaic` library: `mosaic::shuffle(dat, replace=FALSE)`
- Calculate quantiles:

Summary & Coding tips

The bootstrap...

1. Involves resampling the full data with replacement
2. Is useful when analytic approaches fail (or when we are feeling lazy)
3. Works for complicated sample statistics too: correlations, predictions, modeling accuracy...

To make it easier...

- Use the `replicate()` function to repeat your resampling lots of times
- Use the `shuffle()` function from the `mosaic` library: `mosaic::shuffle(dat, replace=FALSE)`
- Calculate quantiles: `quantile(bootstrapped.avgs, c(0.025, 0.975))`

Thank you!

LaTeX, R code, and data at:

`http://www.github.com/aaronrkaufman/bootstrap`