Confidence Intervals via Bootstrapping

Dr. Maria Tackett

Halloween 2019 🎃

Click for PDF of slides

Announcements

- HW 03 **due TODAY at 11:59p**
- Electronic Undergraduate Research Conference on Nov 1
- Review proposal comments in the "Issue" of your GitHub repo
  - Data Analysis due Friday, November 15
- Extra credit

Packages

```
library(tidyverse)
library(infer)
```

```
library(tidyverse)
manhattan <- read_csv("data/manhattan.csv")
```

Observed sample vs. bootstrap population

Sample median = $2350

Population median = ?

Confidence intervals

Bootstrapping scheme

1. **Take a bootstrap sample** - a random sample taken with replacement from the original sample, of the same size as the original sample.

2. **Calculate the bootstrap statistic** - a statistic such as mean, median, proportion, slope, etc. computed on the bootstrap samples.

3. **Repeat steps (1) and (2) many times to create a bootstrap distribution** - a distribution of bootstrap statistics.

4. **Calculate the bounds of the XX% confidence interval** as the middle XX% of the bootstrap distribution.

Confidence intervals

- **Bootstrap**
- **Bounds**: cutoff values for the middle XX% of the distribution
- **Interpretation**: We are XX% confident that the true population parameter is in the interval.
- **Definition of confidence level**: XX% of random samples of size n are expected to produce confidence intervals that contain the true population parameter.
- `infer::generate(reps, type = "bootstrap")`

Rent in Manhattan: 95% confidence interval

```
manhattan %>%
  specify(response = rent) %>%
  generate(reps = 15000, type = "bootstrap") %>%
  calculate(stat = "median") %>%
  summarize(lower_bound = quantile(stat, 0.025), #stat = medians from bootstr
            upper_bound = quantile(stat, 0.975))
```

```
## # A tibble: 1 x 2
##   lower_bound upper_bound
##         <dbl>       <dbl>
## 1       2162.        2875
```

We are 95% confident that the median rent for a one bedroom apartment in Manhattan is between $2162 and $2875.

Calculating confidence intervals at various confidence levels

How would you modify the following code to calculate a 90% confidence interval? How would you modify it for a 99% confidence interval?

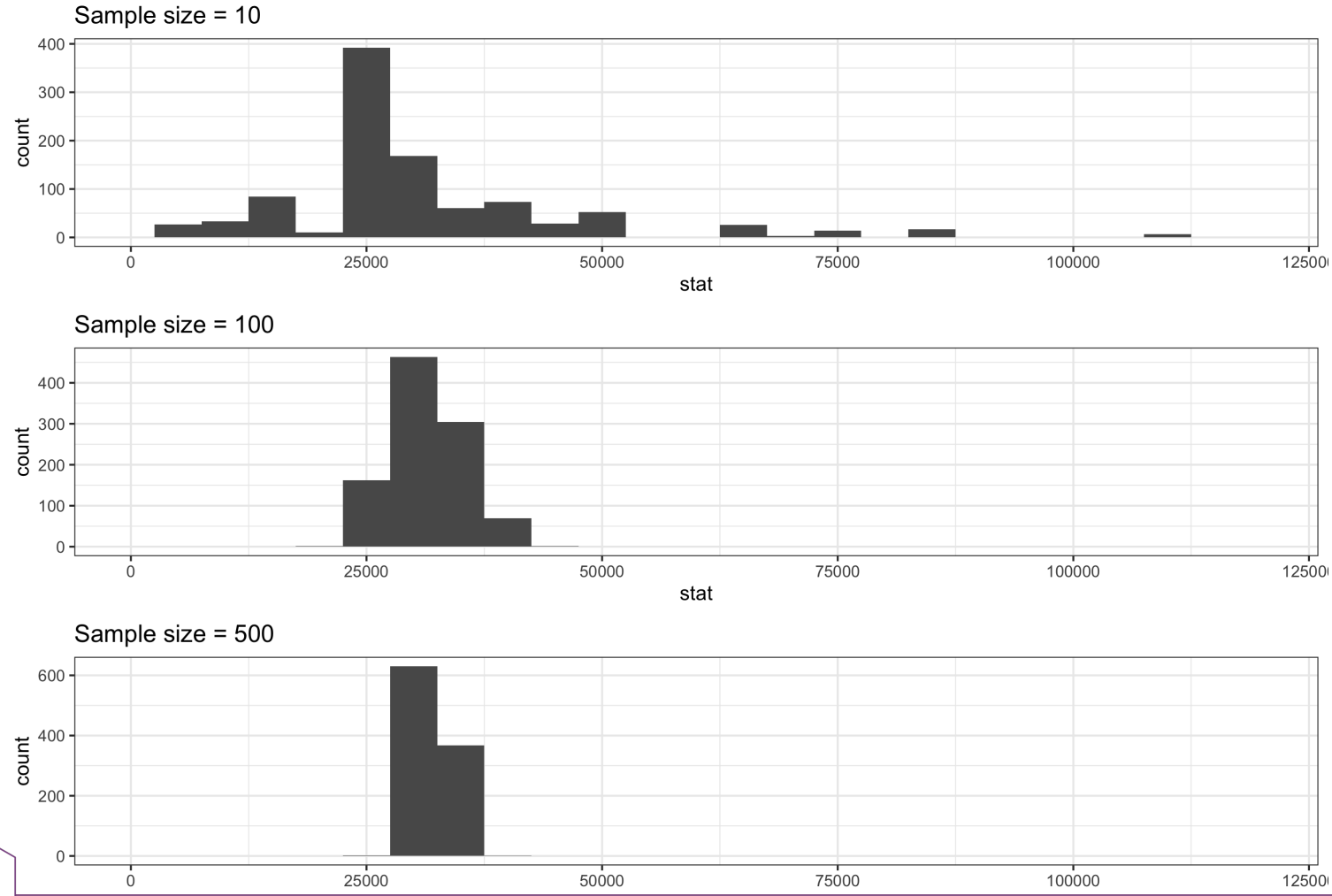
```
manhattan %>%
  specify(response = rent) %>%
  generate(reps = 15000, type = "bootstrap") %>%
  calculate(stat = "median") %>%
  summarize(lower_bound = quantile(stat, 0.025),
            upper_bound = quantile(stat, 0.975))
```

Accuracy vs. precision

What happens to the width of the confidence interval as the confidence level increases? Why?

Should we always prefer a confidence interval with a higher confidence level?

Sample size and width of intervals



Confidence Interval for standard deviation

```
sd_boot_dist <- manhattan %>%
  specify(response = rent) %>%
  generate(reps = 15000, type = "bootstrap") %>%
  calculate(stat = "sd")
```

```
visualize(sd_boot_dist)
```

Confidence interval for standard deviation

```
(percentile_ci <- get_ci(sd_boot_dist) )
```

```
## # A tibble: 1 x 2
##   `2.5%` `97.5%`
##    <dbl>   <dbl>
## 1   523.    951.
```

Confidence interval for standard deviation

```
## # A tibble: 1 x 2
##   `2.5%` `97.5%`
##    <dbl>   <dbl>
## 1   523.    951.
```

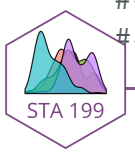
```
visualize(sd_boot_dist) +
  shade_confidence_interval(endpoints = percentile_ci)
```

Comparing visitors at National Parks

This dataset contains location and visitor information about National Parks in the United States years 1904 to 2016. We will use the data to obtain an estimate of the difference in the average number of visitors to parks in the Southeast and those in Pacific West during this time period.

```
parks <- read_csv("data/national_parks.csv")
glimpse(parks)
```

```
## Observations: 21,560
## Variables: 12
## $ year              <chr> "1904", "1941", "1961", "1935", "1982", "1919"…
## $ gnis_id           <chr> "1163670", "1531834", "2055170", "1530459", "2…
## $ geometry          <chr> "POLYGON", "MULTIPOLYGON", "MULTIPOLYGON", "MU…
## $ metadata          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…
## $ number_of_records <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ parkname          <chr> "Crater Lake", "Lake Roosevelt", "Lewis and Cl…
## $ region            <chr> "PW", "PW", "PW", "PW", "PW", "NE", "IM", "NE"…
## $ state             <chr> "OR", "WA", "WA", "WA", "CA", "ME", "TX", "MD"…
## $ unit_code         <chr> "CRLA", "LARO", "LEWI", "OLYM", "SAMO", "ACAD"…
## $ unit_name         <chr> "Crater Lake National Park", "Lake Roosevelt N…
## $ unit_type         <chr> "National Park", "National Recreation Area", "…
## $ visitors          <dbl> 1500, 0, 69000, 2200, 468144, 64000, 448000, 7…
```



Bootstrap interval to compare means of two groups

Step 1: Take a bootstrap sample from Group 1 and a bootstrap sample from Group 2. These are random samples, taken with replacement, from the original samples, of the same size as the original samples.

Step 2: Calculate the bootstrap statistic - find the mean of each bootstrap sample and take the difference between them.

Step 3: Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap differences in the means

Step 4: Calculate the bounds of the XX% confidence interval as the middle XX% of the bootstrap distribution.

Bootstrap interval to compare means in R

- This new setup will change the model we **`specify()`**
  - We will specify **`reponse =`** and **`explanatory =`**
- The **explanatory variable** is the one to be used for to split the data into groups.
- In addition to specifying the explanatory and response variables, we will also need to specify the order in which to subtract the means in Step (2) above, i.e. Group 1 Mean - Group 2 Mean, or the other way around.
- The same steps apply if you take a difference in the median, proportions, etc.

Comparing National Parks

We'd like to obtain a 95% confidence interval for the difference in the mean number of visitors to National Parks in the Southeast (SE) region and the Pacific West (PW) region between 1904 and 2016. We'll use the variables

- **`region`**: SE or PW
- **`visitors`**: Number of visitors

Open the RStudio Cloud Project **National Parks - Bootstrap Intervals**. Complete Part 1 in the .Rmd file.

Interpretation of confidence intervals

Which of the following is more informative:

- The difference in price of a gallon of milk between Whole Foods and Harris Teeter is 30 cents.
- A gallon of milk costs 30 cents more at Whole Foods compared to Harris Teeter.

What does your answer tell you about interpretation of confidence intervals for differences between two population parameters?

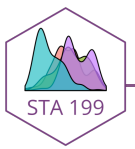
Confidence intervals exercise

Describe the simulation process for estimating the parameter assigned to your team.

- Note any assumptions you make in terms of sample size, observed sample statistic, etc.
- Imagine using index cards or color chips to represent the data.

> Lab 01: single population proportion
>
> Lab 02: difference between two population medians
>
> Lab 03: difference between two population proportions

Write your response in Part 2 of the **National Parks - Bootstrap Intervals** project in RStudio Cloud.


STA 199