Section A01
CDS 101 *Introduction to Computational and Data Sciences*

Class 17: Inference and simulation III

June 13, 2018

General

Annoucements

- Midterm reports and presentation slides due by the start of class on Thursday, June 14th
    - Each group's presentation must be given tomorrow and it cannot be made up, being absent results in an automatic zero for the presentation part of your project grade
- Written responses (not questions) to Reading 14 (to be posted) due on **June 15** by 9:00am
    - Last reading that requires you to post on Slack after completing it
- Homework 3 due by **11:59pm on Friday, June 13th**
- Homework 4 to be posted soon, will be due by **11:59pm on Wednesday, June 20th**
- Homework 5 will also be posted soon and will be an **optional extra credit assignment** also due by **11:59pm on Wednesday, June 20th**
    - Homework 4 must be submitted before you can turn in Homework 5
- Details for Final will be discussed tomorrow after the Midterm Project presentations

On the midterm project presentations

- Presentation order
  - I flipped a coin, Team 1 goes first, then Team 2
- Equipment
  - Students usually like to use one of their own laptops for the presentation
  - Make sure you bring any video adaptors you might need for HDMI or VGA outputs
  - My laptop is available: Powerpoint slides will be uploaded and presented using Powerpoint Online

On the midterm project presentations

- Reminder of expectations
  - Review the midterm project instructions for details such as presentation length and what to discuss
  - Each group member is expected to speak an approximately equal amount of time
  - **The slides must look uniform and all content from every group member must be in a single Powerpoint/Google Slides file**
  - Figures from `ggplot2` should be saved as PNG files using `ggsave()` and then imported into the slideshow
  - Tables should be in the presentation software's native format
- You will fill out a brief form I give to you to evaluate the other team's presentation
- Class will continue after the presentations, which should not take up more than the first half hour of the class

Variability of estimates

Pew Research Survey

Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

Margin of error

The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- 41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.
- 49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18–34 years olds have taken a job they didn't want just to pay the bills.

Parameter estimation

- We are often interested in **population parameters**.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Parameter estimation

- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

Confidence intervals

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss.

If we toss a net in that area, we have a good chance of catching the fish.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss.

If we toss a net in that area, we have a good chance of catching the fish.

- By analogy, if we report a point estimate (such as the mean or median), we probably won't hit the exact population parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Why do we report confidence intervals?

- A plausible range of values for the population parameter is called a **confidence interval**.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss.

If we toss a net in that area, we have a good chance of catching the fish.

- By analogy, if we report a point estimate (such as the mean or median), we probably won't hit the exact population parameter.
- If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

- We can use the bootstrap simulation from `infer`:

What is a bootstrap simulation?

Bootstrap on *Seeing Theory*

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

- We can use the bootstrap simulation from `infer`:

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

- We can use the bootstrap simulation from `infer`:

```
yawn_bootstrap <- yawn %>%
  specify(yawn ~ group, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("Treatment", "Control"))
```

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?

- We can use the bootstrap simulation from `infer`:

```
yawn_bootstrap <- yawn %>%
  specify(yawn ~ group, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("Treatment", "Control"))
```

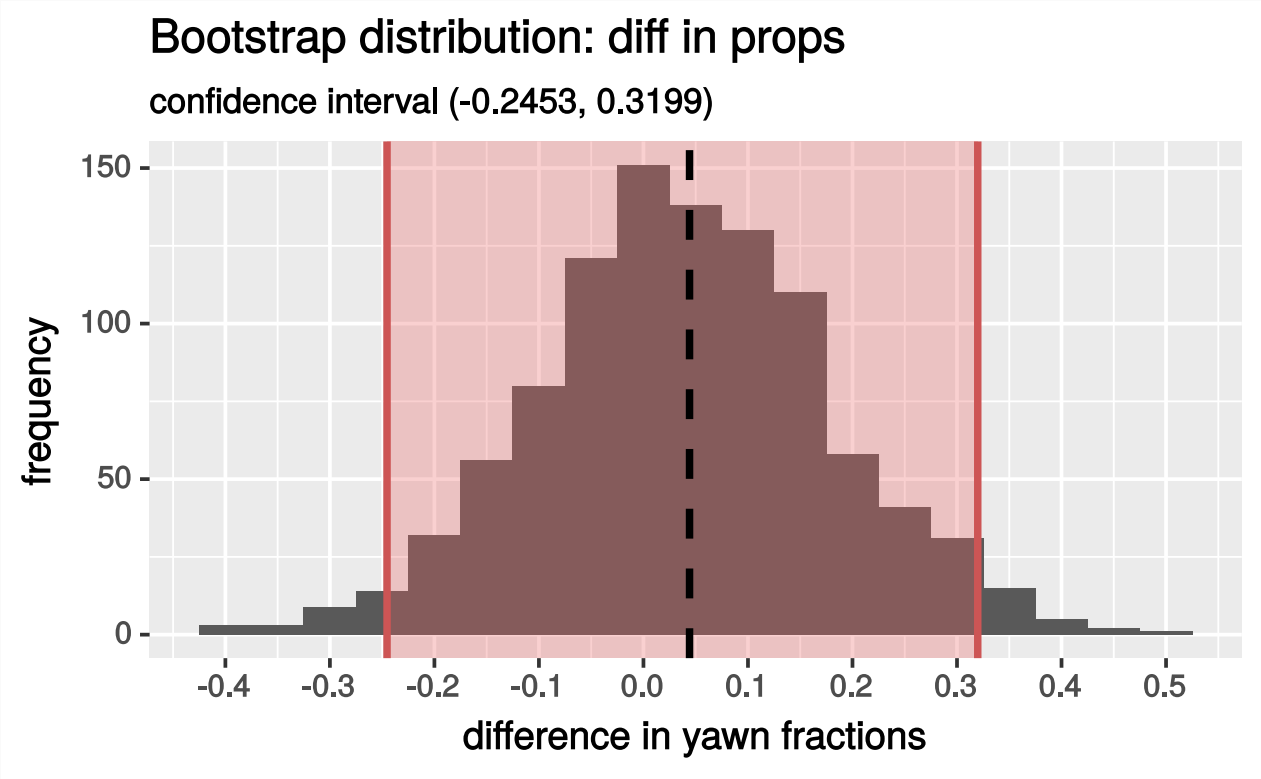
```
yawn_ci_bounds <- yawn_bootstrap %>%
  mutate(rank = min_rank(stat)) %>%
  filter(
    between(rank, n() * 0.025, n() * 0.975)
  ) %>%
  summarize(
    lower = min(stat),
    upper = max(stat)
  )
```

lower	upper
-0.2453222	0.3198529

- `min_rank(stat)` is the stat column's sorting order from smallest to largest
- `0.025 * n()` is the `rank` that defines the threshold for the 2.5th percentile
- `0.975 * n()` is the `rank` that defines the threshold for the 97.5th percentile
- `min(stat)` and `max(stat)` gives thresholds for the 2.5th and 97.5th percentiles

Example: Constructing a confidence interval

What is the 95% confidence interval for the *Mythbusters* yawning experiment?



Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns
2. People will, on average, yawn 24% less to 29% more when someone near them yawns

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns
2. People will, on average, yawn 24% less to 29% more when someone near them yawns
3. A randomly chosen person yawns 24% less to 29% more when someone near them yawns

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns
2. People will, on average, yawn 24% less to 29% more when someone near them yawns
3. A randomly chosen person yawns 24% less to 29% more when someone near them yawns
4. 95% of people yawn 24% less to 29% more when someone near them yawns

Example: Interpreting the confidence interval

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that:

1. People in this sample, on average, yawn 24% less to 29% more of the time when someone near them yawns
2. People will, on average, yawn 24% less to 29% more when someone near them yawns
3. A randomly chosen person yawns 24% less to 29% more when someone near them yawns
4. 95% of people yawn 24% less to 29% more when someone near them yawns

Demo using `infer` on the county dataset

```
library(tidyverse)
library(infer)
county <- read_rds(
  path = url("https://summer18.cds101.com/files/datasets/county_complete.rds")
)
```

Demo how to use `infer` on continuous, numerical data

- *Null hypothesis*: The average time it takes to travel to work in Virginia is the same as Maryland.
- *Alternative hypothesis*: The average time it takes to travel to work in Virginia is different from Maryland.

Do the data allow us to reject the null hypothesis in favor of the alternative hypothesis?

Credits

Content in the **Variability in estimates** and **Confidence intervals** sections was adapted from the chapter 4 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license.