STAT 102: Week 5

Ricky’s Section

Introductions and Attendance

Introduction: Name

Question of the Week: Which best describes you: Data Visualizer (🎨), Data Wrangler (🧹), or Data Collector (🔍)?

Important Reminders

My Office Hours

- For this week, **Sat, 03/01 at 11 AM - 12 PM, 1 - 2 PM** in Adams D-Hall
- Most up-to-date information on Slack and Spreadsheet
  - https://docs.google.com/spreadsheets/d/1AgnpomB7qUGtyRTI8Ansothj-2LbZnKJYO7qobr6c18/edit?usp=sharing
- Sorry for all the changes! Hopefully, things will be normal starting next week

Midterm

- **Written Component:** Wed, 03/12 from 6 to 9 PM in Science Center 705 and 706
- **Oral Component:** Over Zoom afterwards on 03/13 and 03/14 (10 minute sessions)
- No class/section on Thur, Mar 13
- **You all got this!** 🙂

Content Review: Week 3

Data Wrangling: Notes on summarize() and mutate()

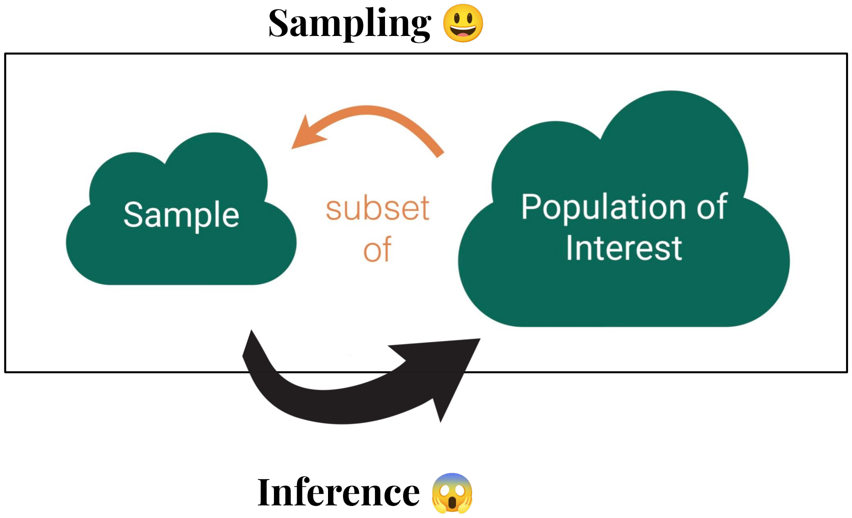
- `summarize()` - Allows for use of summary functions, such as `mean()`, `sd()`, `cor()`, `IQR()`, and `n()`, which can be set equal to new variables
    - *summarize(mean_INCOME = mean(INCOME), mean_IRAX = mean(IRAX), households = n())*

- `mutate()` - Modifies existing variables and/or create new ones
    - *hpi %>%*
      mutate(LogFootprint = log(Footprint))

Content Review: Week 5

Introduction to Inference

- Last week, we went **from population to sample**. Moving forward, we'll go **from sample to population**!
- Why? Recall the difficulty of obtaining a **census**
- We have data from a **sample** and are interested in concluding something about the **population**



Parameter vs. Statistic

Population parameter:

- Typically **unknown** (what we're interested in finding)
- For **population proportion**, it's denoted as p
  - This is for binary categorical variables
  - There are many other parameters, which we'll soon learn about!
- *Ex: Out of all 67 million viewers of the debate, how many believed Harris won? I don't know!*

Sample statistic:

- **Known**/calculated from the **sample**
- For **sample proportion**, it's denoted as $\hat{p}$
- *Ex: From my (random) sample of 600 viewers, how many believed Harris won? Let's say it was 300, so* $\hat{p} = 0.5$

A **sample statistic** is a **point estimate** of the **population parameter** (i.e., our best guess, but we could be wrong)

Other Parameters and Statistics

	Response Variable		Numeric Quantity	Sample Statistic	Population Parameter
1 variable	Numerical		Mean	$\bar{x}$	μ
	Binary Categorical		Proportion	$\hat{p}$	p
	Response variable	**Explanatory Variable**	**Numeric Quantity**	**Sample Statistic**	**Population Parameter**
2 variables	Numerical	Binary Categorical	Difference in Means	$\bar{x}_1 - \bar{x}_2$	$\mu_1 - \mu_2$
	Binary Categorical	Binary Categorical	Difference in Proportions	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$
	Numerical	Numerical	Correlation	r	ρ

What is unknown here? Does it make sense to have a confidence interval for the sample statistic?

Question:

What is unknown here? Does it make sense to have a confidence interval for the sample statistic?

The **population parameter** is unknown while the **sample statistic** is known (it's a number we calculate, like $\bar{x}$ = $210,000), so it doesn't make sense to have a **confidence interval** for the sample statistic.

Conversely, we want to know more about the (unknown) population parameter.

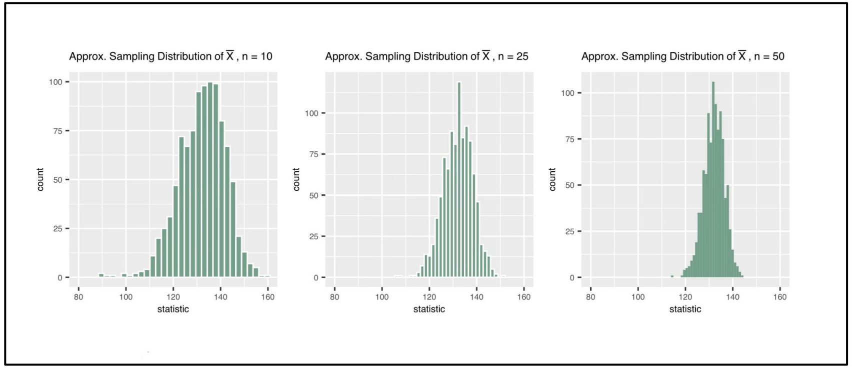
Sampling Variability

- We could've taken a different **sample** of 600 people from the **population** of 67 million viewers
  - The **sample proportion** (probably) would've differed
- **Sampling variability** refers to the **differences in the sample statistic** from sample to sample
  - If we take many samples, how much would the **sample proportion** vary?
  - $\hat{p} = 0.5$ in this sample, but $\hat{p} = 0.4$ in that sample, and so on

Sampling Distribution

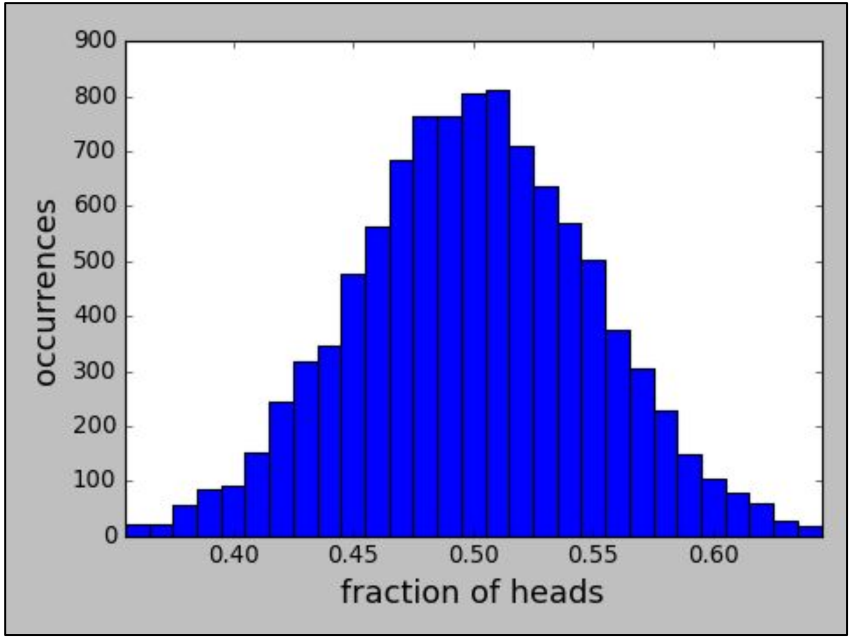
Sampling distribution of a statistic:

- Graph of **sample statistics** from **repeated samples** (requires access to entire **population**)
- **Center** of **sampling distribution** is **population parameter**
- As n, sample size of each rep, increases…
  - **Standard error** (standard deviation of sampling distribution) **decreases** (indicated by less spread)
  - **Sampling distribution** becomes **more bell-shaped and symmetric**



Coin Flips: An Intuition behind Sampling Distributions

- Let's flip a fair coin 10 times and record the proportion of heads
- Will our sample statistic always be 0.5? No!
- The center is the "theoretical" population proportion ($p = 0.5$)
- We're graphing a bunch of sample proportions ($\hat{p}_1 = 0.4$, $\hat{p}_2 = 0.5$, $\hat{p}_3 = 0.6$, ...)



What is the “problem” with the sampling distribution? I.e., what do we need access to?

Question:

What is the “problem” with the sampling distribution? I.e., what do we need access to?

To construct a **sampling distribution**, we need access to the entire **population** from which to draw **repeated samples**.

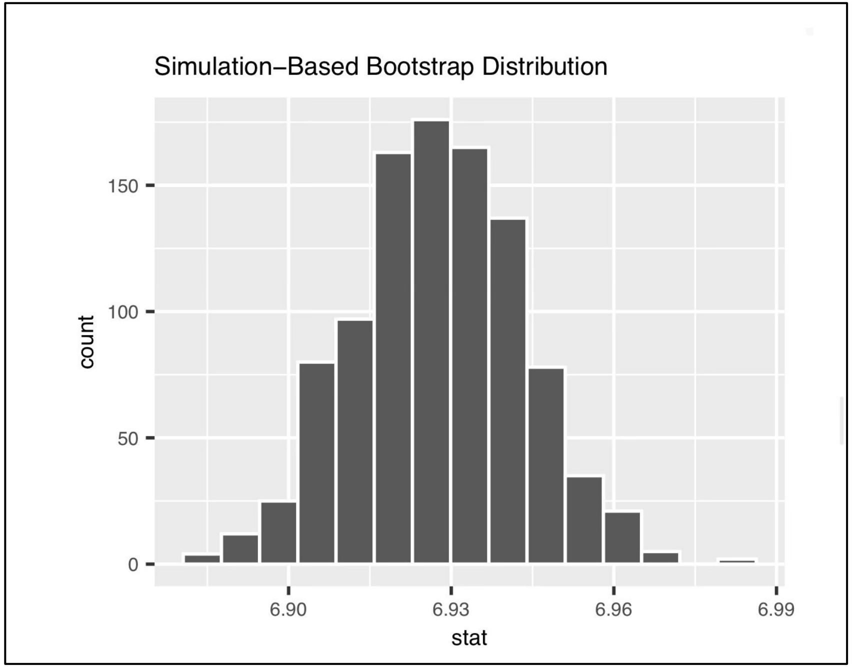
This is not always practical.

Here’s where the **bootstrap distribution** comes in!

Bootstrap Distribution

Bootstrap distribution of a sample statistic:

- **Procedure**: Take a **sample** of size *n* (with **replacement**) from the **original sample**, compute the statistic on this **bootstrap sample**, and repeat many times to get many **bootstrap statistics** (basically, **sampling the sample**)
  - We no longer need the entire **population**
- **Bootstrap distribution** graphs these **bootstrap statistics**
- **Center** of bootstrap **distribution** is the **original sample statistic**



Example of Bootstrapping

Population: {100, 250, 75, 30, 50, 75, 100, 300, 120, 55, 80, 90}, μ = 110.416...

Original Sample (n = 4): {250, 75, 75, 120}, $\bar{x} = 130$

Bootstrap sample #1 (n = 4): {250, 120, 120, 250}, $b_1 = 185$

Bootstrap sample #2 (n = 4): {75, 120, 75, 250}, $b_2 = 130$

Bootstrap sample #3 (n = 4): {75, 75, 120, 75}, $b_3 = 86.25$

and so on...

Sampling Distribution vs. Bootstrap Distribution

Sampling distribution:

- Requires access to the entire **population**
- Its **center** is the **population parameter**
- Its **spread/standard deviation** is the **standard error**, which we need to compute a CI

Bootstrap distribution:

- Does NOT require access to the entire **population**
  - We only need **1 sample**
- Its **center** is the **sample statistic**
- Its **spread/standard deviation** is a **good estimate for standard error**

Confidence Interval

Confidence interval: Range of **plausible** values (around the **sample statistic**) that may contain the **population parameter**

- **SE method: CI = statistic ± z* × ($\hat{S}\hat{E}$)**
  - z* is critical value, $\hat{S}\hat{E}$ is **standard deviation** of bootstrapped statistics (spread of **bootstrap distribution**)
  - *Ex: 95% CI = statistic ± 1.96($\hat{S}\hat{E}$)*
- **Percentile method: CI = the middle (CL)% of the bootstrap distribution**
  - CL = confidence level
  - *Ex: 95% CI = the middle 95% of the bootstrap distribution*

I tell you we’re 100% confident the true average hours of sleep Harvard students get every night (our population parameter) is between 0 and 24 hours. Is this informative?

Question:

I tell you we're 100% confident the true average hours of sleep Harvard students get every night (our population parameter) is between 0 and 24 hours. Is this informative?

No, not really!

This is why it's important we choose a good confidence level (usually, but not always, 95%).

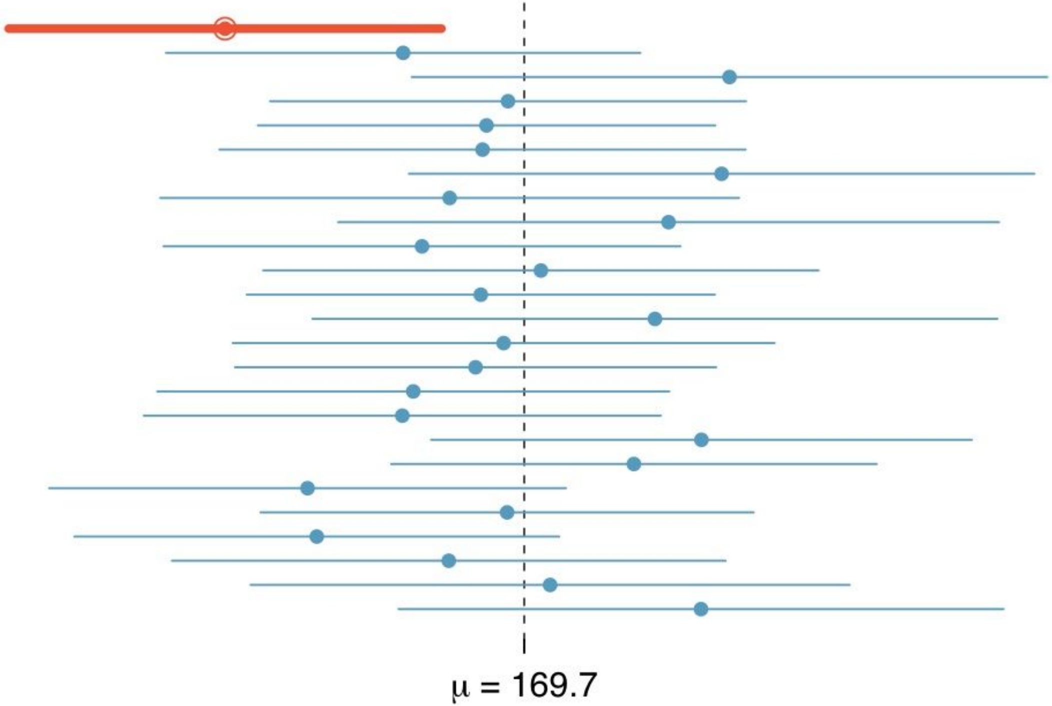
There's a trade-off between how narrow/informative our interval is and how confident we are.

Interpreting Confidence Intervals

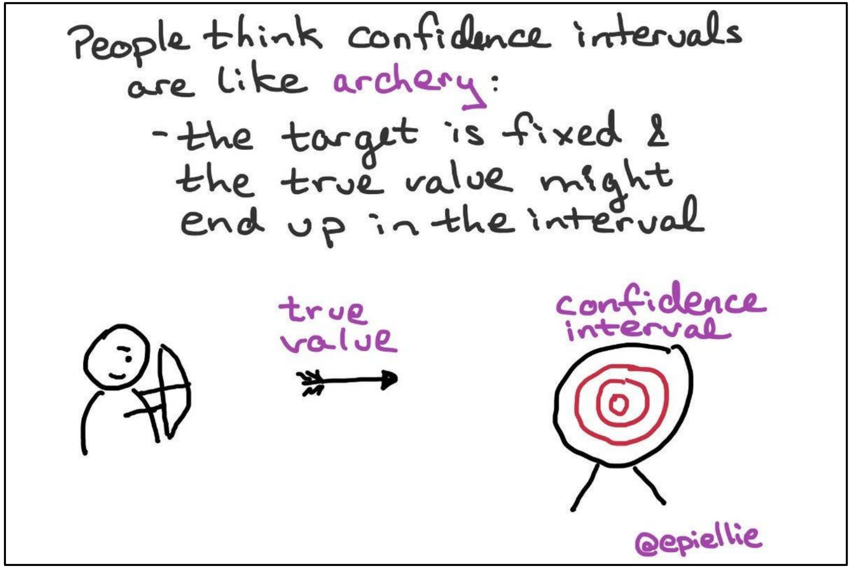
- **"We are {<u>confidence level</u>}% confident that the interval ({<u>lower bound</u>}, {<u>upper bound</u>}) captures the true {<u>population parameter</u>}."**
  - **Confidence** is NOT **probability**
  - Either the **parameter** is in the CI (100% probability) or it's not (0% probability)
  - For a 95% CI, we expect it to succeed (for it to capture the population parameter) **95/100 times**

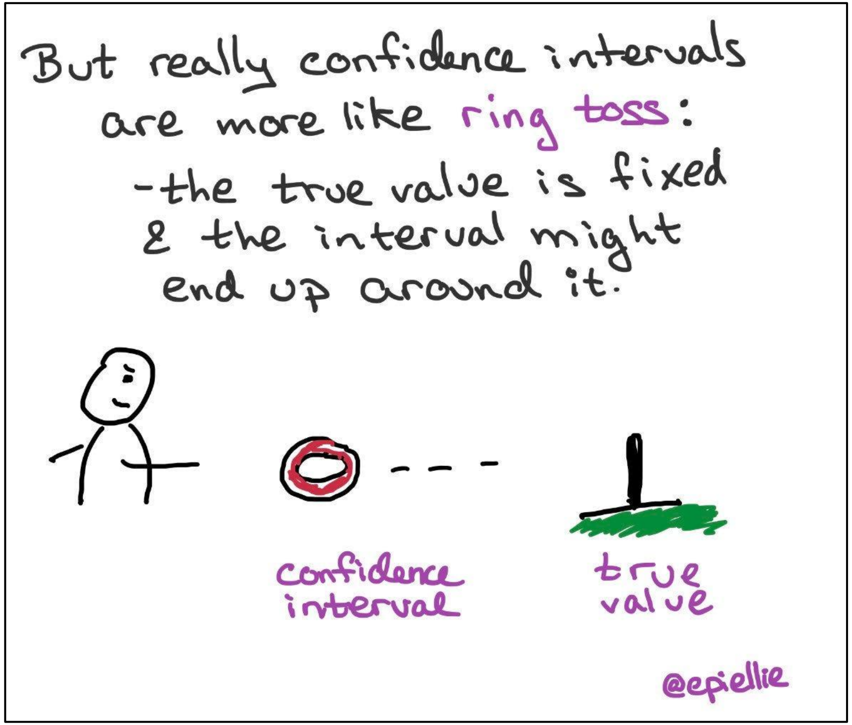
The meaning of confidence. . .

Twenty-five samples of size $n = 60$ were taken from the 'artificial' population, then a 95% CI for μ was computed based on each sample. Only 1 of these 25 intervals did not contain μ.



An Analogy with Ring Toss





https://medium.com/@EpiEllie/having-confidence-in-confidence-intervals-8f881712d837

Why does the sampling dist. get narrower as we increase n?

Question:

Why does the sampling dist. get narrower as we increase n?

n is the **sample size** of each rep.

When ***n* is small** (e.g., n = 10), we're drawing small samples, so a single **outlier** can drastically skew our **sample statistic**. As ***n* increases**, **outliers** become less "powerful."

Also, we know when ***n* is the population**, the **sampling statistic** is just the **population parameter**.

Important Code for Week 5

https://drive.google.com/file/d/1dl7IlLhz9u4cAhKio_zj7Fkxxs-bxVfk/view?usp=drive_link

Questions?

Midterm Review (Weeks 1-5)

Week 2: Data Visualization

- **Grammar of graphics**: Dataset, geom, aesthetic
- **Color palettes**: Sequential, diverging, qualitative
- **Choosing the right graph**

Week 3: Data Wrangling

- **Data joins**: Left, (right), inner, full
- **Creating/modifying variables**
- **Grouping/filtering/selecting data**
- **Summary statistics**: Mean, median, SD, IQR
- **Handling missing values (NA)**
- **Interpreting code in English**

Week 4: Data Collection

- **Groups**: Sample, census, population
- **Observational study vs. experiment**
- **Two types of bias**: Sampling, nonresponse
- **Four sampling methods**: Simple, systematic, cluster, stratified

Week 5: Simulation-Based Inference

- **<u>Parameter vs. statistic</u>**
- **<u>Distributions</u>**: Sampling, bootstrap
- **<u>Confidence intervals</u>**: Constructing, interpreting

Midterm Tips

- **PLEASE SET A TIMER FOR THE ORAL!** There should be 3 questions in 10 minutes, so try not to "ramble"
- If you haven't already, make a **study guide**
- **Partial credit** counts (so don't delete all your code)
- Remember to **load all relevant libraries**
- **Pace yourself**—if a question is taking too long, move on
- Sign up for **practice oral exams** (usually not 3 questions)

Midterm Practice

Practice 1: Everybody

https://drive.google.com/file/d/1GjMautF3pZowG6DIgLDWcDoVt3DlFJyO/view?usp=drive_link

Practice 2: Person A (Grade Q1, Answer Q2)

https://drive.google.com/file/d/1101gLTgPXgkWN3JtudLDn05xphQ3ky3T/view?usp=drive_link

Practice 2: Person B (Answer Q1, Grade Q2)

https://drive.google.com/file/d/1w6EkCBEeNHPVdPBhmUVxF64zA0HZ1pQK/view?usp=drive_link

Questions?

P-Set 4

Have a great rest of your week!