STAT 100: 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

Anonymous Feedback

https://docs.google.com/forms/d/e/1FAIpQLSfKv_FGvs00qm-IvtxKx3Vf6bBzSJE2jamK1gklAzL6NkXE8w/viewform

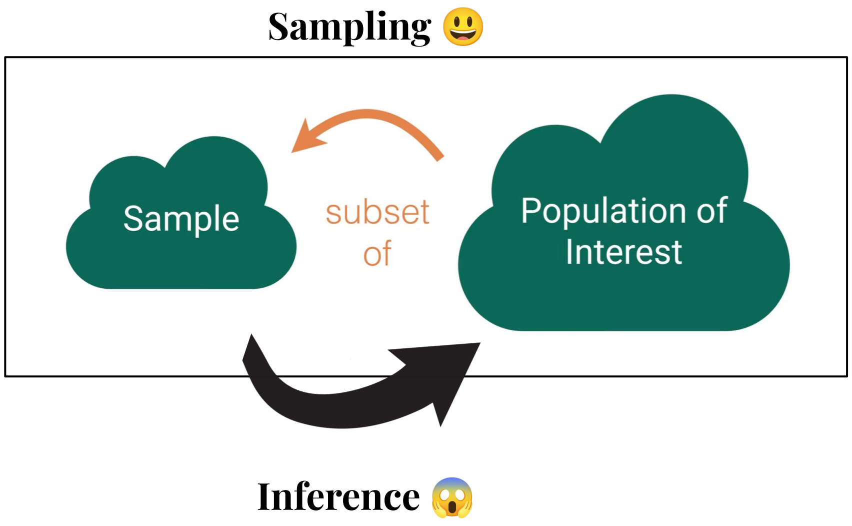
Midterm

- **Written Component:** Wed, Oct. 16 from 6 to 9 PM in SC Hall B
- **Oral Component**: Over Zoom afterwards (10 minute sessions)
- Let me know if you have any questions
- **You all got this!** 🙂

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

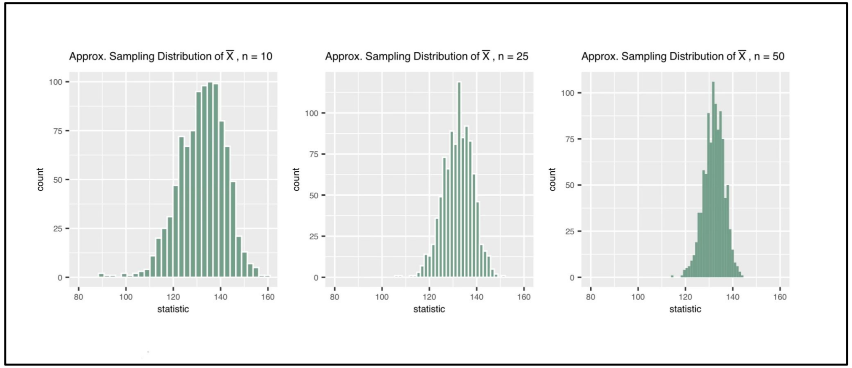
Sampling Variability

- 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

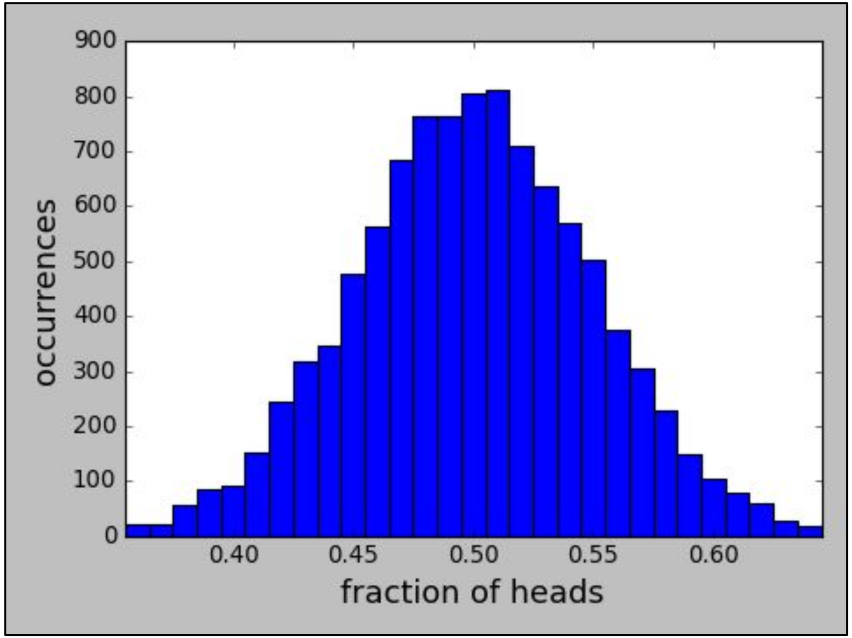
Sampling distribution of a statistic:

- 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?

Question:

What is the “problem” with the sampling distribution?

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

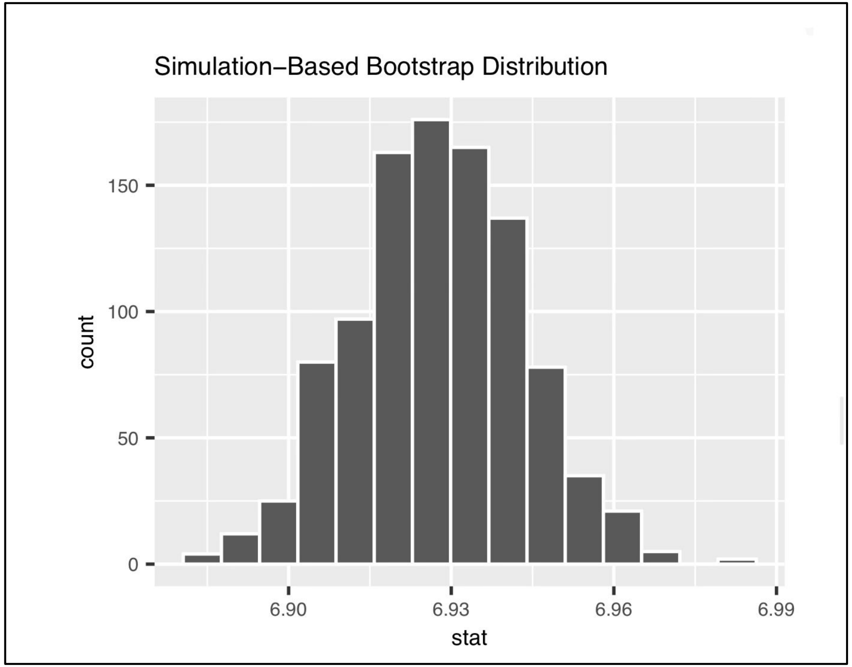
This is not always practical.

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: {250, 120, 120, 250}, $\bar{x}$ = 185

Bootstrap sample #2: {75, 120, 75, 250}, $\bar{x}$ = 130

Bootstrap sample #3: {75, 75, 120, 75}, $\bar{x}$ = 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**

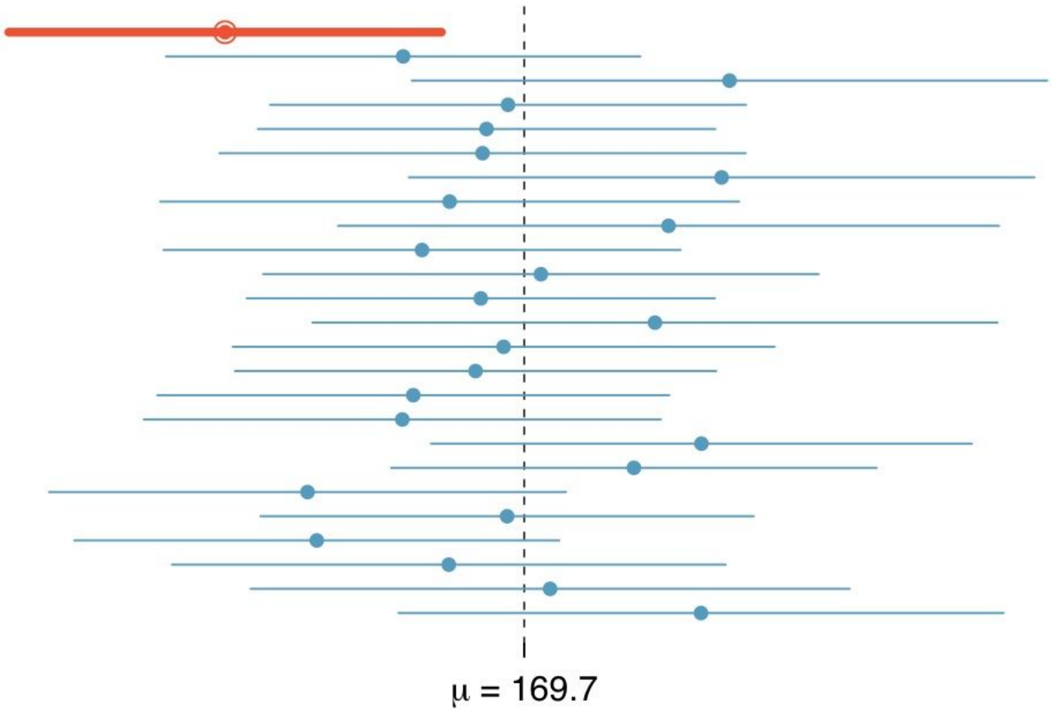
- **SE method: CI = statistic ± z* × ($\hat{SE}$)**
  - z* is critical value, $\hat{SE}$ is **standard deviation** of bootstrapped statistics (spread of **bootstrap distribution**)
  - *Ex: 95% CI = statistic ± 1.96($\hat{SE}$)*
- **Percentile method: CI = the middle (CL)% of the bootstrap distribution**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

Interpreting Confidence Intervals

- **"We are {<u>confidence level</u>}% confident that the true {<u>population parameter</u>} lies between {<u>lower bound</u>} and {<u>upper bound</u>}."**
  - **Confidence** is NOT **probability**
  - Either the **parameter** 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.



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/1Rmadk9HC-uP7UopgojoPQtXky81j6lx1/view?usp=sharing

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/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

- **Parameter vs. statistic**
- **Distributions**: Sampling, bootstrap
- **Confidence intervals**: Constructing, interpreting

Midterm Tips

- **PLEASE SET A TIMER!** 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
- 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://docs.google.com/document/d/1JDJq_JlwrJBLWlKcKWoAKYSw_wl4zEviBdQrpme3v8I/edit#heading=h.ewccgla427fy

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

https://docs.google.com/document/d/1ukYrohJJBqqYOBBe79gD5uW7h_LUcZ7QEgonaYK0x2g/edit?usp=sharing

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

https://docs.google.com/document/d/1B7N4EHJamlo5Z8BCDVP8Zr-mb2TiZYtQeKPz-ozy-18/edit?usp=sharing

P-Set 4

Have a great rest of your week!