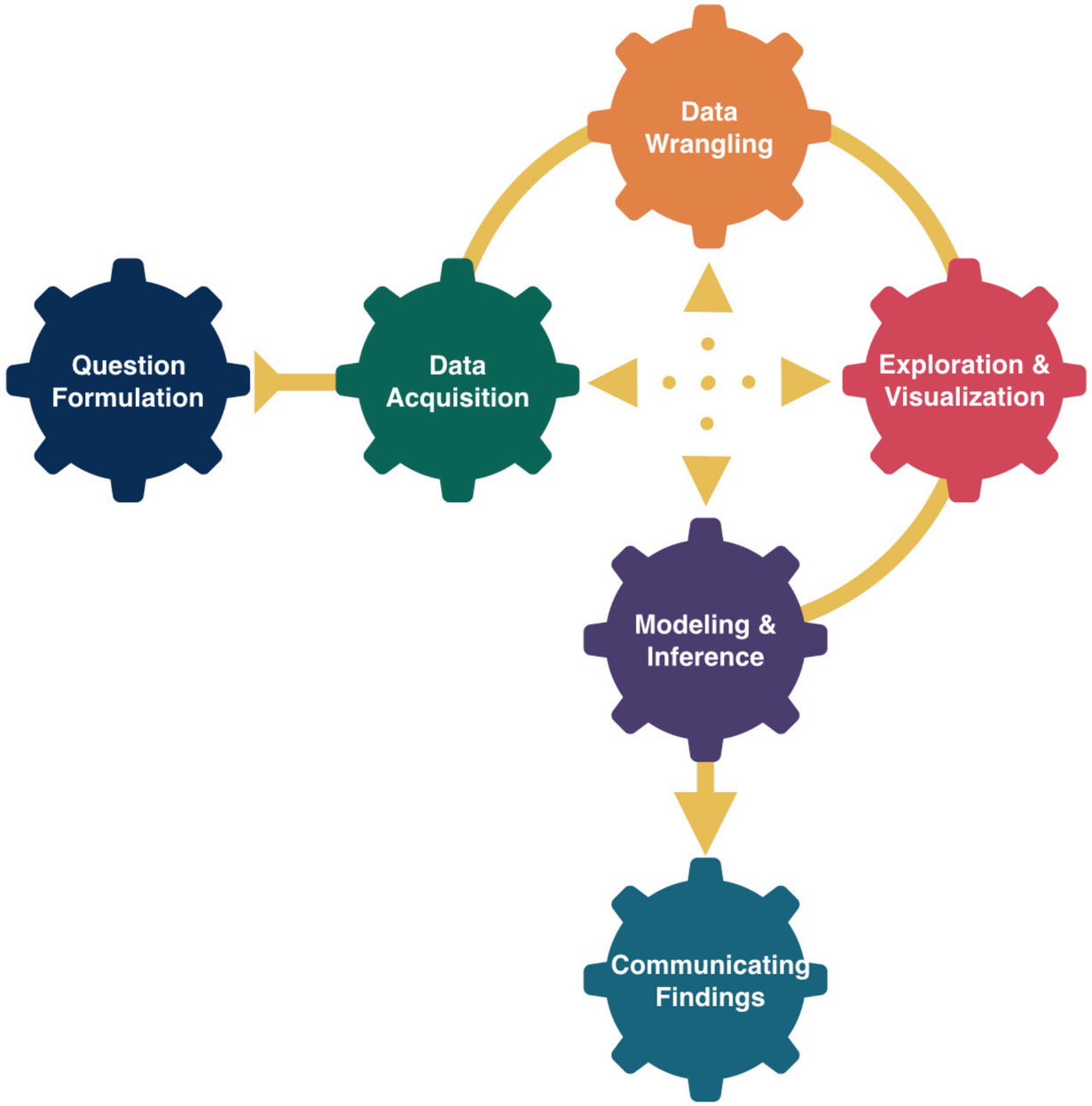
MATH 141: Combined Lecture Slides 1-7



Statistical Thinking

Megan Ayers

Math 141 | Spring 2026

Monday, Week 1

Before we get started…

Attendance/waitlist form check

Goals for Today

1. Getting started in Math 141
    i. Course structure and technologies
    ii. Where to find resources
    iii. Course expectations
2. Introduce statistical thinking
3. Introduce data frames

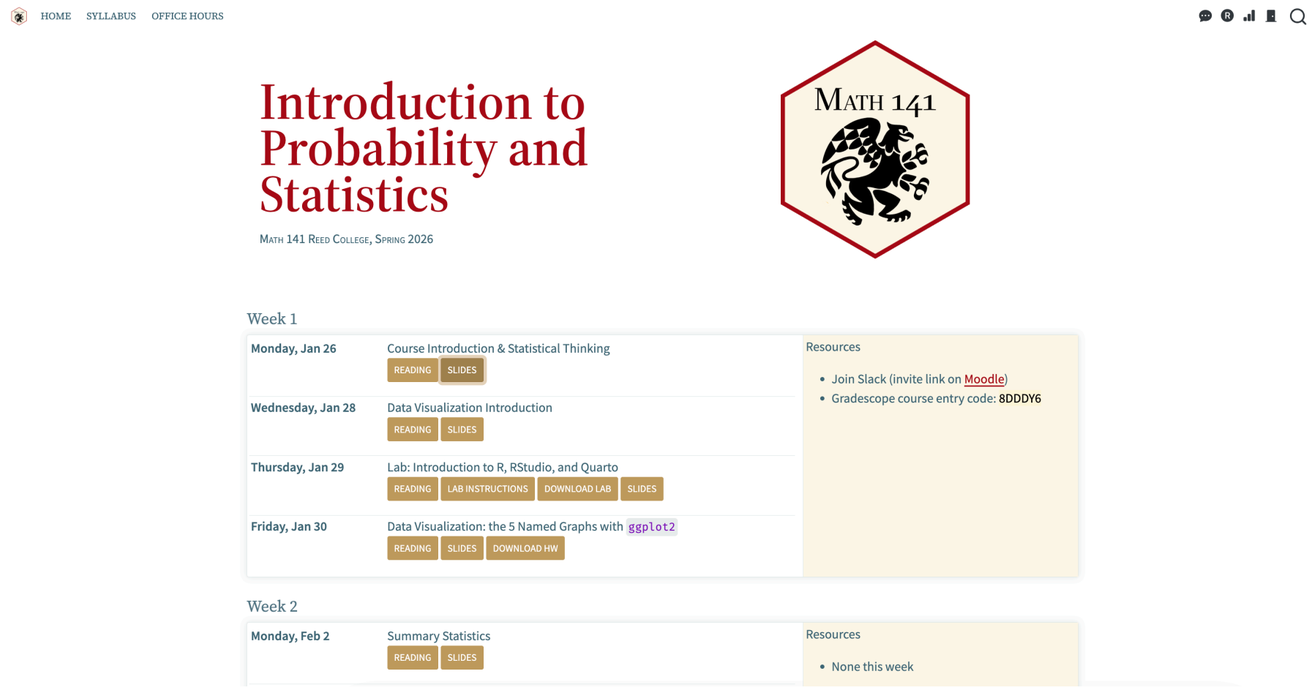
A bit about me

- Boise, ID → Lewis & Clark College (Portland) → Yale (New Haven) → back to Portland!
- My Research Interests:
  - Causal inference
  - Environmental policy evaluation
  - Text as data
  - R programming
- I’m looking forward to getting to know each other more in lab!

Getting Started in Math 141

Getting Started in Math 141

Course website



- The course website, megan-k-ayers.github.io/math-141-sp26, will be the central location for all our course materials.
- We'll use a few other resources to navigate Math 141, but everything will be linked/directed to from the course website.

Getting Started in Math 141

Other Resources

Reed's **RStudio Server**, for coursework,

A course-wide **Slack** workspace, for course communication,

Gradescope, for turning in assignments, and

Moodle, sparingly, for private course materials & info about meeting times/locations

Textbooks (all free and online)

- Statistical Inference via Data Science: A ModernDive into R and the tidyverse, Second Edition (**MD**)
- Introduction to Modern Statistics, Second Edition (**IMS**)
- OpenIntro Statistics, Fourth Edition (**OI**)

Getting Started in Math 141

Other Resources

You will need access to a computer for this course.

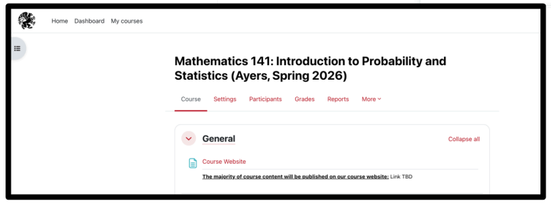
Please let me know ASAP after class or via email if you do not have access to a personal computer.

Getting Started in Math 141

Finding Resources

Links to resources

Lab and HW files



My info

Name: Megan Ayers (she/her), you can call me ‘Megan’

Email: meganayers@reed.edu

Office and office hours: Posted on Moodle, or by appointment

Learning support resources

- The Math 141 Teaching Team are excited to support your learning!
  - Course assistant office hours coming soon on Moodle
- Each student is entitled to one hour of individual tutoring per week

A typical week in Math 141

- **Monday**
  - Attend lecture
- **Tuesday**
  - Turn in lab assignment from previous week by 11:59pm
- **Wednesday**
  - Attend lecture
- **Thursday**
  - Attend lab
- **Friday**
  - Turn in homework assignment by 11:59pm (most weeks)
  - Attend lecture
  - Next homework assignment is released

Assignments and Exams

- **Lab assignments** (*weekly-ish*)
  - Assigned Thursday in lab
  - Due the next Tuesday, 11:59pm on **Gradescope**
  - Mostly R practice. Please collaborate!
- **Homework assignments** (*weekly-ish*)
  - Assigned on Friday (usually)
  - Due the next Friday (usually), 11:59pm on **Gradescope**
  - Mostly theory/conceptual, some R. Please collaborate!
- **In-class activities** (*semi-regular*)
  - Individual activities
  - Group activities
- **Exams** (*1 midterm, 1 final*)
  - Both are written and in-person
  - Midterm: 3/12
  - Final: Finals week

Expectations for submitted work

- Address questions completely. If you have a partial solution, say so, explain what you've tried.
- Respond in full sentences for written answers.
- Show your work.
- Write comments in your code explaining your work.

Late Work

- **Labs**: up to 4 extensions days can be used throughout the semester.
  - e.g., 1 additional day for 4 labs, 4 additional days for 1 lab, …
  - rounding days up
- **HW**: up to 4 extensions days can be used throughout the semester.
  - e.g., 1 additional day for 4 HWs, 4 additional days for 1 HW, …
  - rounding days up
- Lab extension days cannot be used for HW, and vice versa
- **Any other late work has a 50% grade deduction, no guarantees on grading timing/feedback**
- **In-class assignments**: cannot be made up.
- **Exams**: no late exams are accepted. Please arrive on time!

Engagement

- Being **actively present** is key. This requires **attendance**, and can look like:
  - Active listening
  - Contributing to discussion at the small group and/or classroom level
  - Contributing to discussions on Slack
- Missing more than 4 lectures or 1 lab will have a small, accumulating impact on your grade.
- During lecture and lab, remove distractions.
  - When we are on our computers, close email, social media, news, etc.
  - Hide your phone.

Engagement

- I have high expectations but know that all of you (regardless of your stats, math, or computing background) have the ability to meet them.
- We are all going to make mistakes, we will learn more because of them.

Course Climate

We expect everyone in this class to strive to foster a learning environment that is equitable, inclusive, and welcoming. If you experience any barriers to learning, please come to Professor Megan Ayers or a college administrator with your concerns.

Code of Conduct:

We expect all members of Math 141 to make participation a harassment-free experience for everyone, regardless of age, body size, visible or invisible disability, ethnicity, sex characteristics, gender identity and expression, level of experience, education, socio-economic status, nationality, personal appearance, race, religion, or sexual identity and orientation.

We expect everyone to act and interact in ways that contribute to an open, welcoming, inclusive, and healthy community of learners. You can contribute to a positive learning environment by demonstrating empathy and kindness, being respectful of differing viewpoints and experiences, and giving and gracefully accepting constructive feedback.[1]

Academic Accomodations

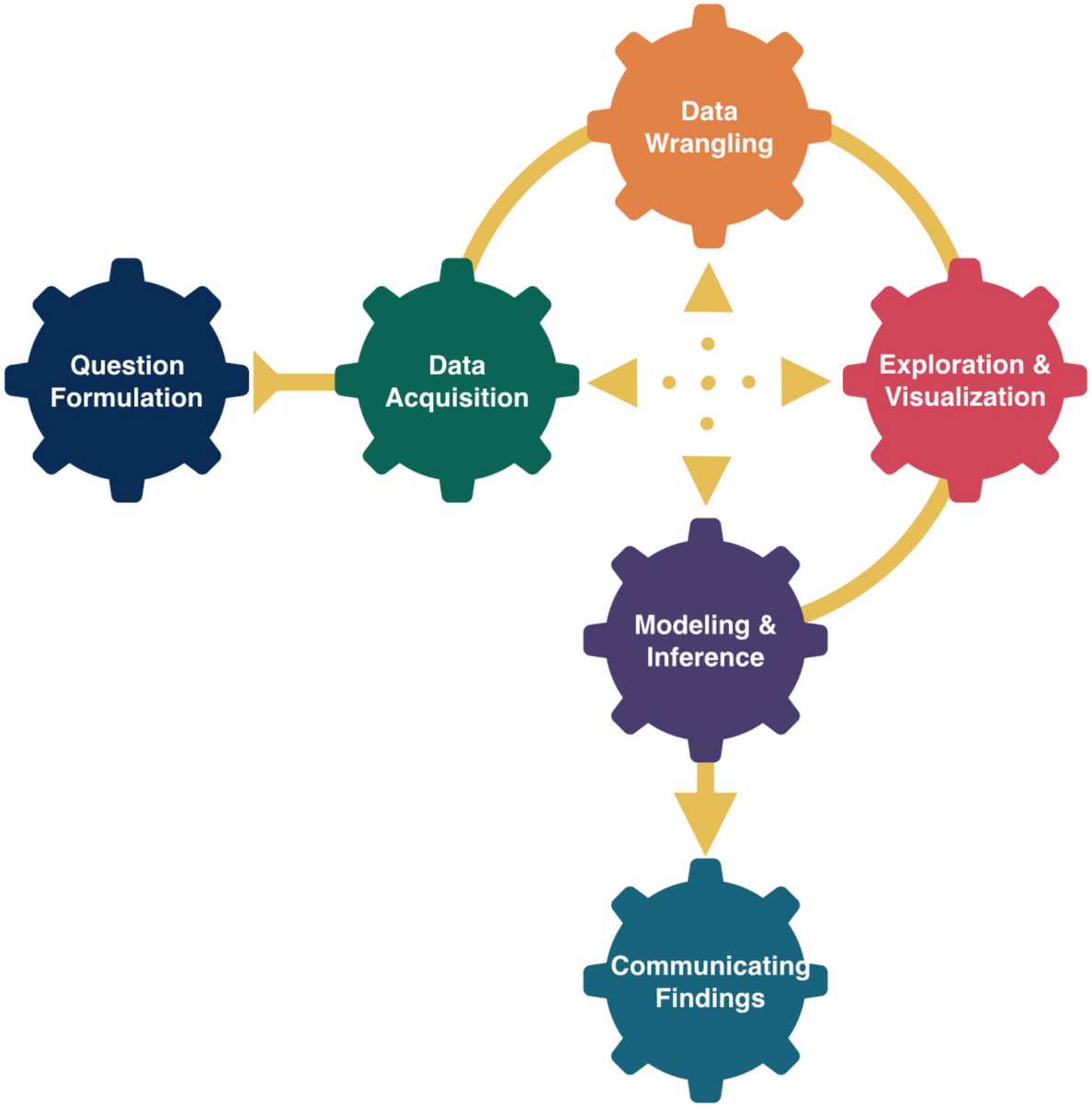
If you plan to request academic accommodations, please submit these **through the DAR student portal** as soon as possible.

Artificial Intelligence (AI) Policy

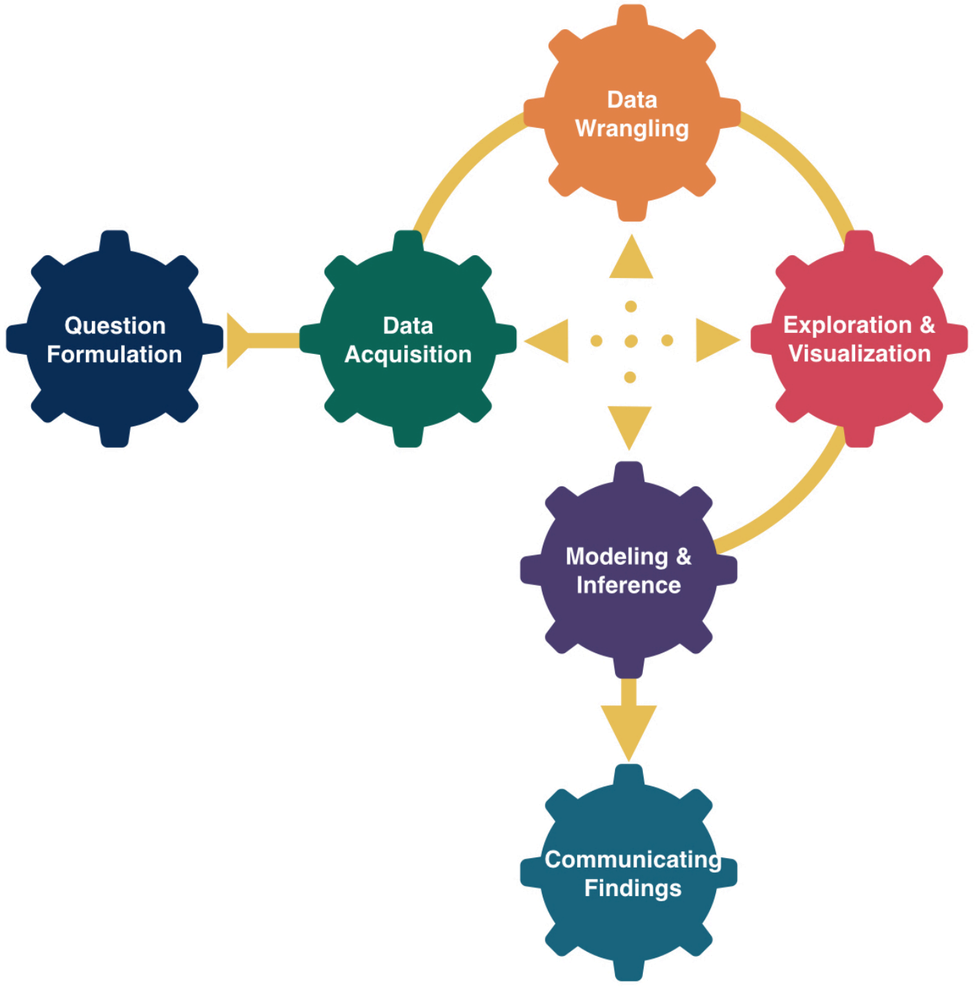
> *Artificial intelligence (AI) tools, such as ChatGPT, Claude, Gemini, and others are being used to generate code, analyze data, and much more. However, learning to think critically about a problem at hand, and engaging with your peers, tutors, and instructors when not understanding a concept or question are integral components of a liberal arts education. Further, a key goal of this course is for you to learn how to thoughtfully, ethically, and independently extract knowledge from data and engage in statistical reasoning. Therefore, the use of generative AI tools, such as ChatGPT and others, is strictly prohibited in any stage of the work process for this course. If you have questions about whether a tool is allowed for this course, ask the Instructor before using it.*

Course or syllabus questions?

Math 141: The whole game


Data Wrangling
Question Formulation
Data Acquisition
Exploration & Visualization
Modeling & Inference
Communicating Findings

Learning Outcomes



- In this course, you will learn how to think critically with data by engaging in the entire data analysis process.
- Most of our time will be spent in the **Exploration and Visualization**, **Data Wrangling**, and **Modeling and Inference** steps, but we will spend some time in each cog!
  - First ~3 weeks in **Exploration and Visualization**, **Data Wrangling**, and **Data Acquisition**
  - Next ~2 weeks in **Modeling**
  - Next ~6 weeks of the course in **Inference**
  - Final weeks combining **Modeling** and **Inference**

Statistical thinking

Math 141 is about developing our statistical thinking skills.

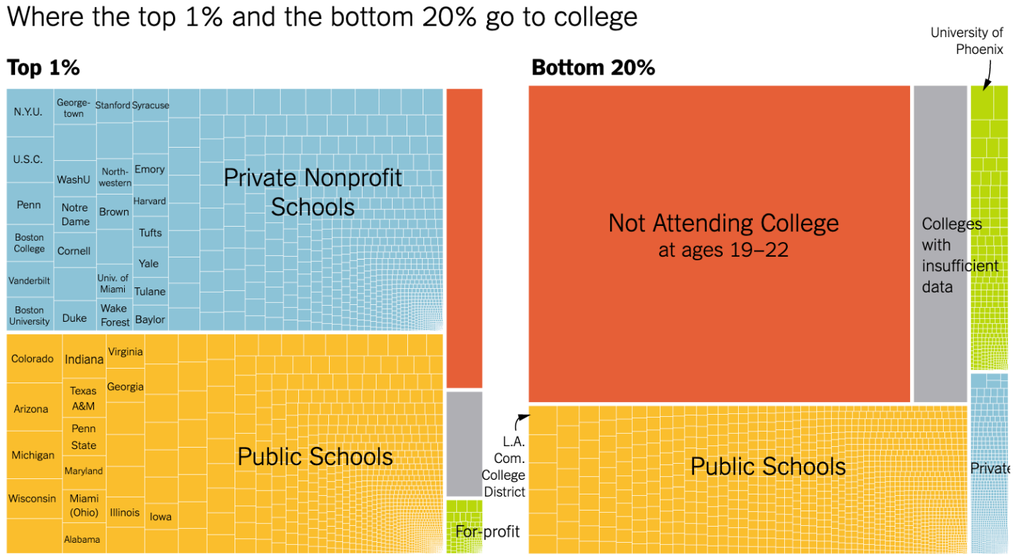
What is **statistical thinking**?

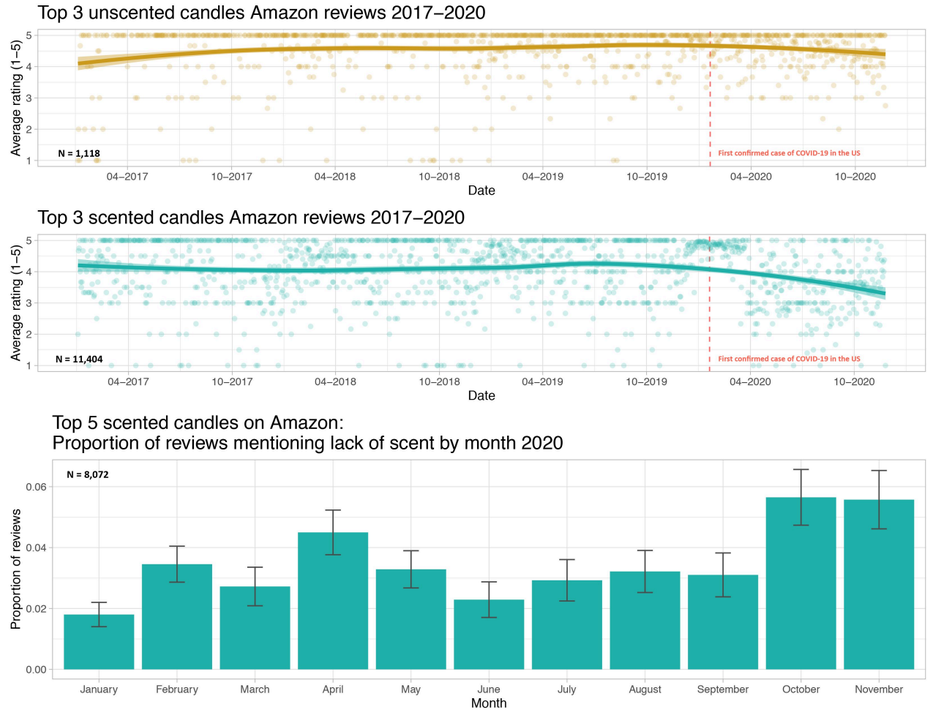
It is distinct from mathematical thinking.

Let’s discover what **statistical thinking** is through some examples.

Data in Math 141

Will use a wide-range of **real** and **relevant** data examples

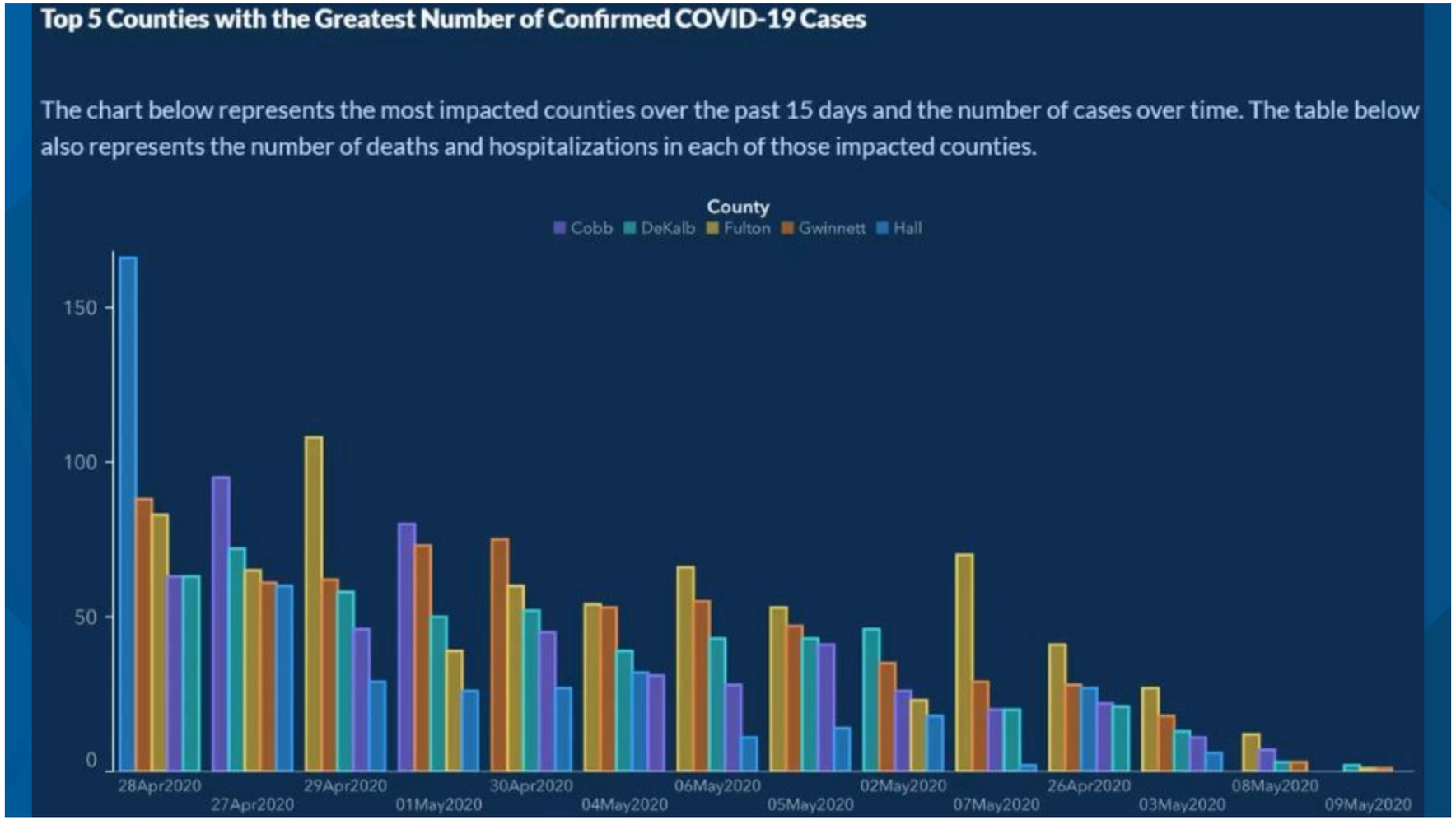




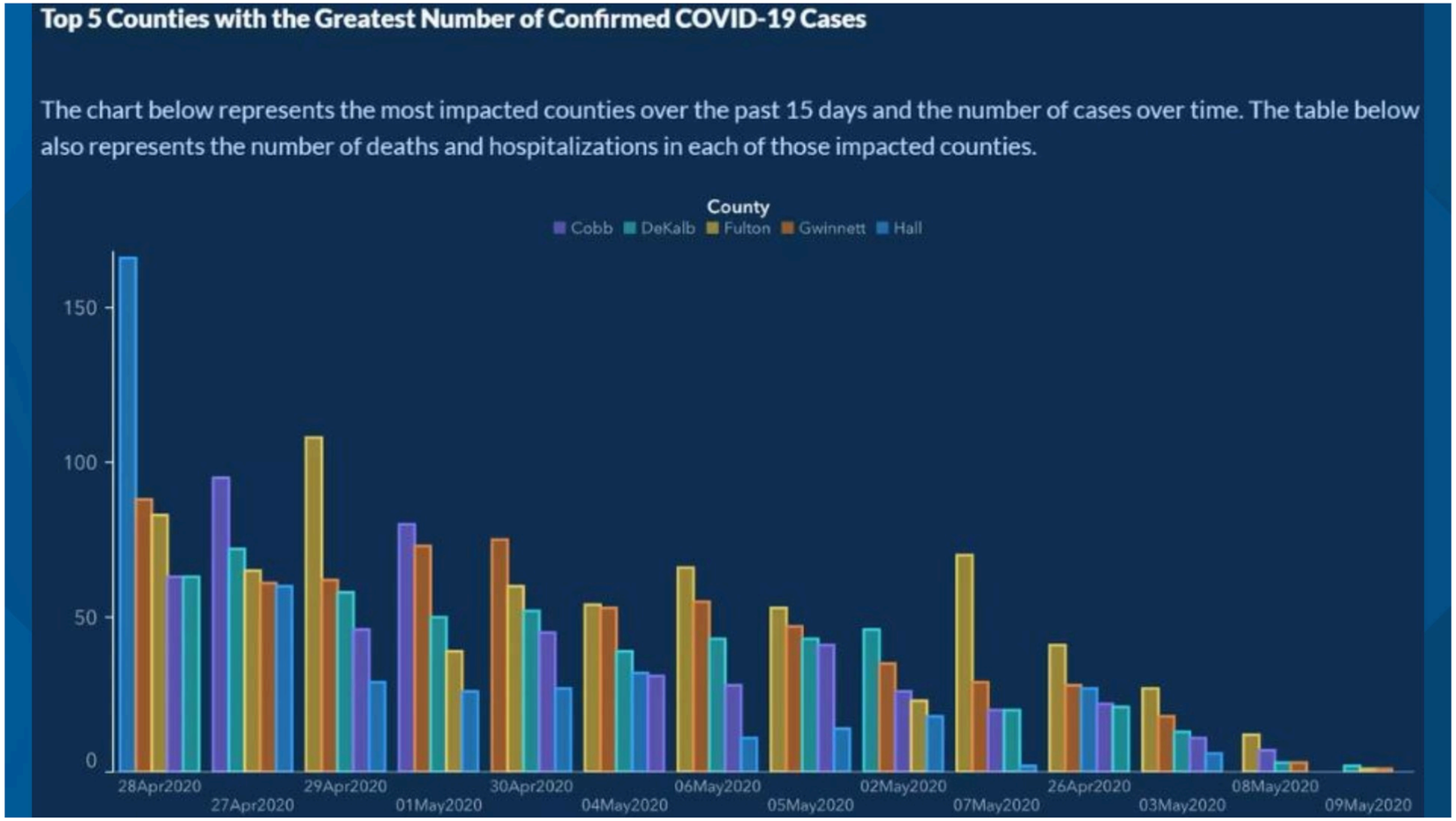
Example: Visualizing COVID Prevalence

Example: Visualizing COVID Prevalence

- In May of 2020, the Georgia Department of Public Health posted the following graph:

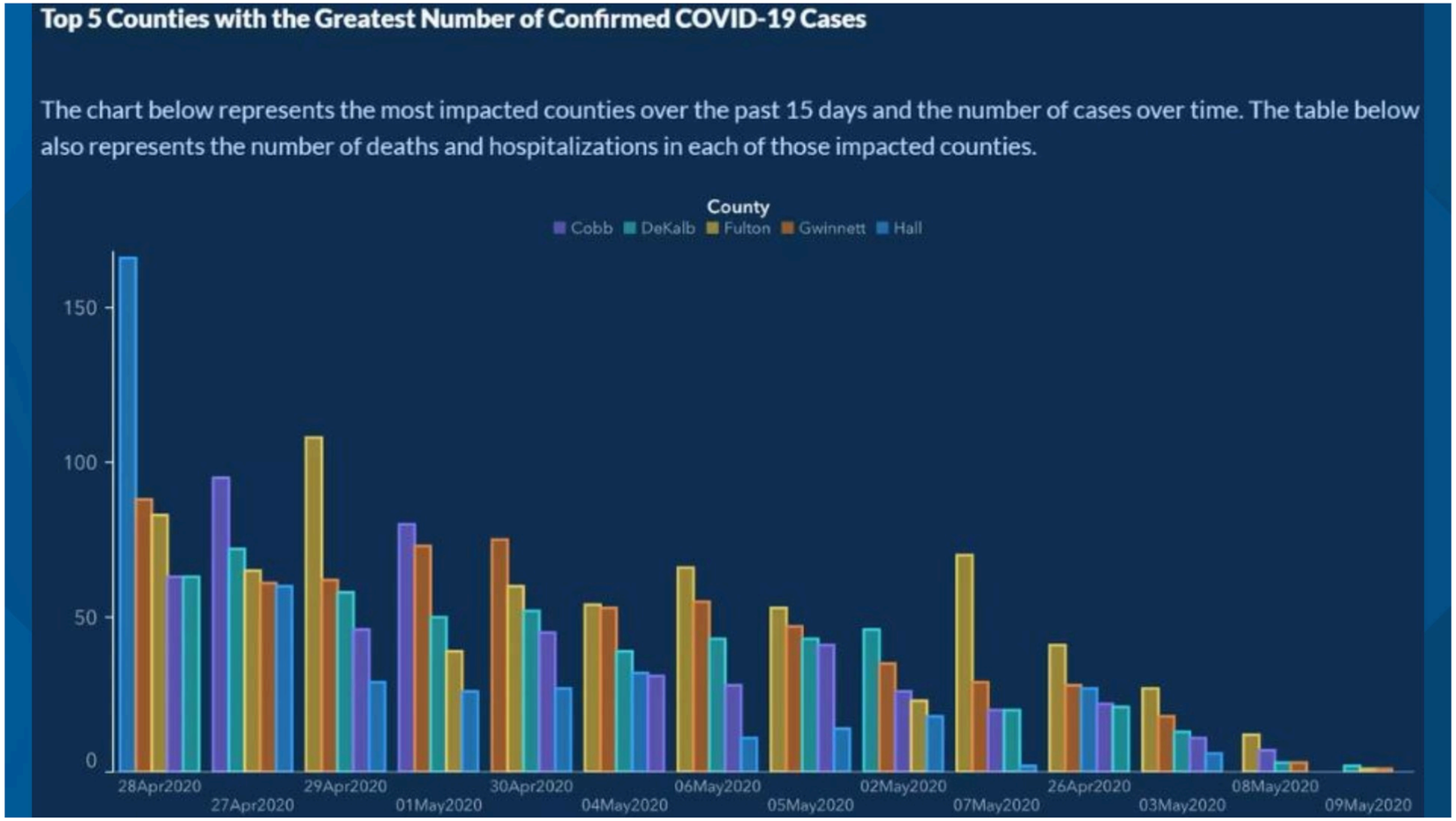


Example: Visualizing COVID Prevalence



- At a quick first glance, what story does the graph appear to be telling?

Example: Visualizing COVID Prevalence



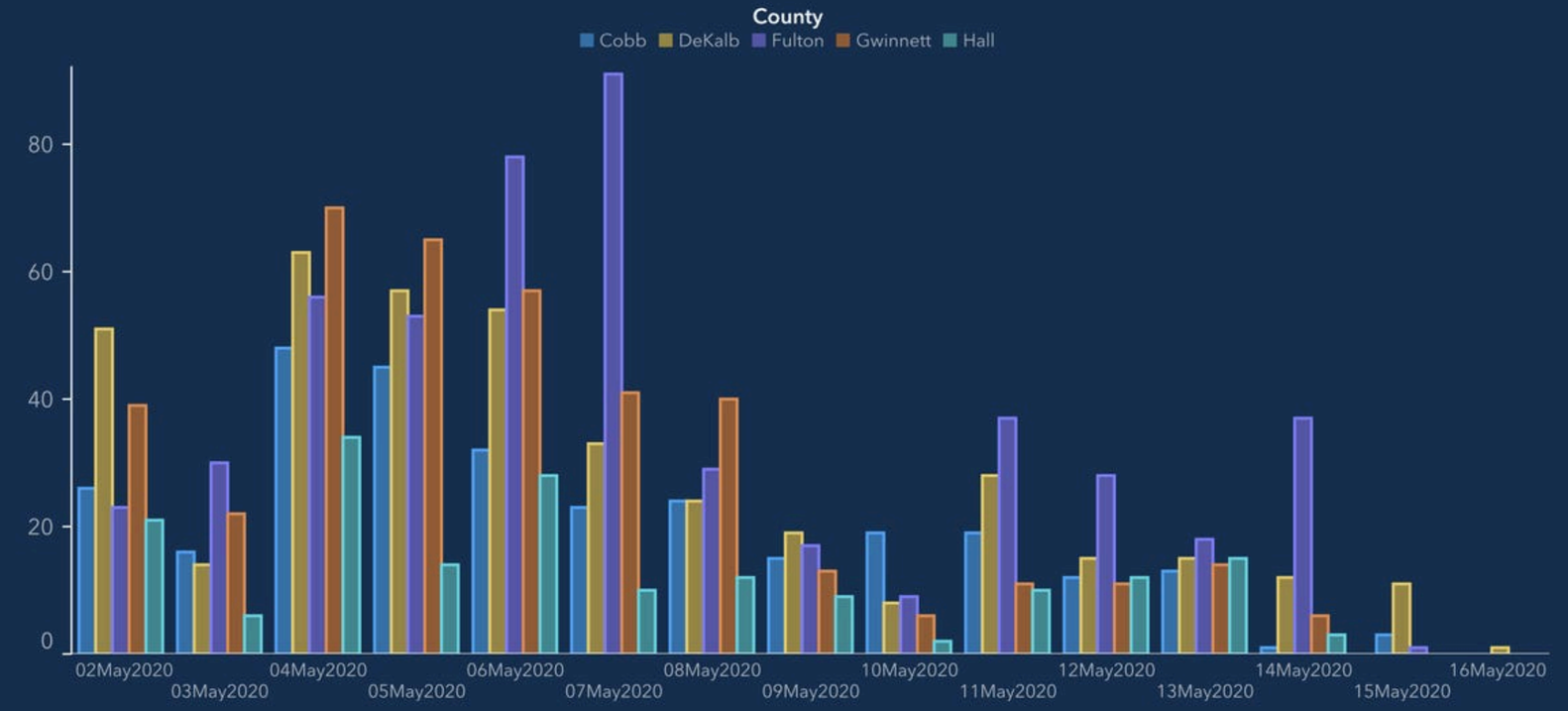
- What is misleading about the graph? How could we fix this issue?

Example: Visualizing COVID Prevalence

- After public outcry, the Georgia Department of Public Health updated the graph:

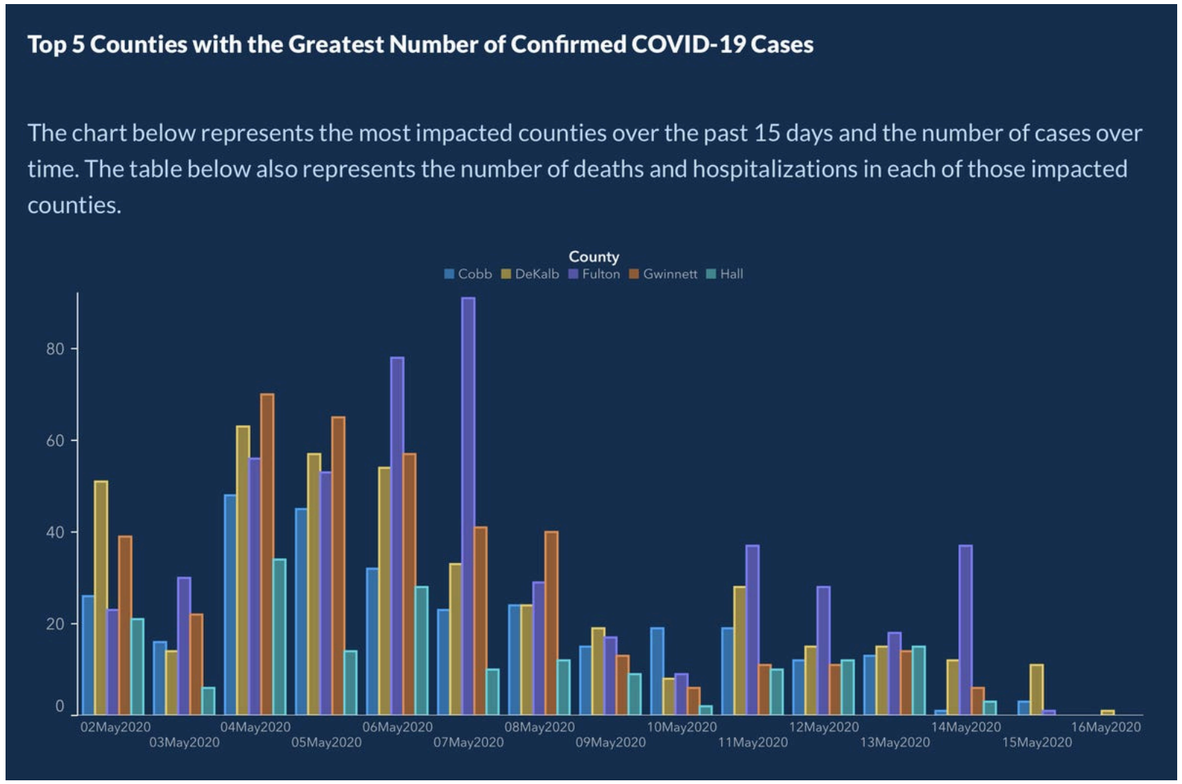
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



Example: Visualizing COVID Prevalence

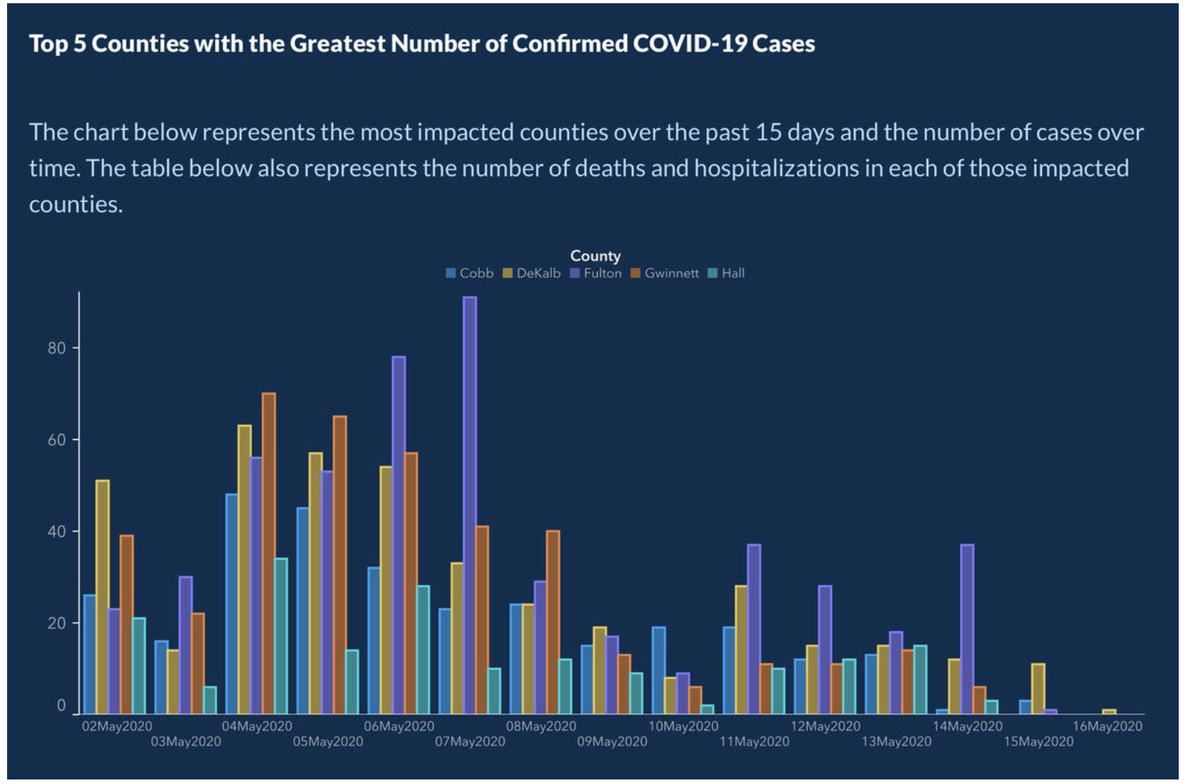
- After public outcry, the Georgia Department of Public Health updated the graph:

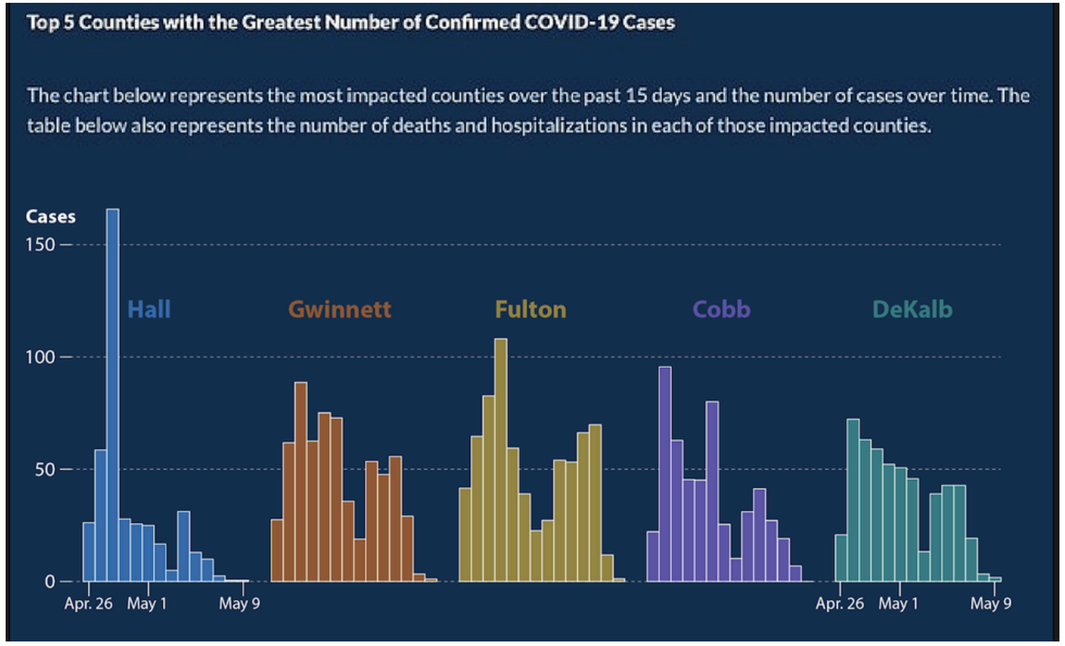


- How do your conclusions about COVID-19 cases in Georgia change when interpreting the new graph?

Example: Visualizing COVID Prevalence

Alberto Cairo, a journalist and designer, created the second graph of the Georgia COVID-19 data:





- A key principle of data visualization is to **"help the viewer make meaningful comparisons"**.
- What comparisons are made easy by the lefthand graph? What about by the righthand graph?

Statistical Thinking

- About developing **reasoning** (not just learning definitions and formulae).
- Statistical thinking requires **judgment** that takes time to develop.
  - Will see **examples** and **practice** applying statistical thinking throughout the course.
  - Numeric calculations only get us so far: it is critical to understand the **underlying data** and **assumptions**
- Developing our statistical thinking skills will allow us to soundly **extract knowledge from data**!

What are/is Data?

> *"'Raw data' is an oxymoron." – Lisa Gitelman*

> *"Data … is information made tractable." – Catherine D'Ignazio and Lauren Klein*

Data Frames

Data Frames

Data in **spreadsheet**-like format where:

- Rows = Observations/cases
- Columns = Variables

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

- Data from **GPT Detectors Are Biased Against Non-Native English Writers.** *Weixin Liang*, *Mert Yuksekgonul*, *Yining Mao*, *Eric Wu*, *James Zou.* CellPress Patterns and available in the R package `detectors`.

Data Frames

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

Rows = Observations/cases

What are the cases? What does each row represent?

Data Frames

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

Columns = Variables

Variables: Describe characteristics of the observations

- **Quantitative**: Numerical in nature
- **Categorical**: Values are categories
- **Identification**: Uniquely identify each case

Data Frames

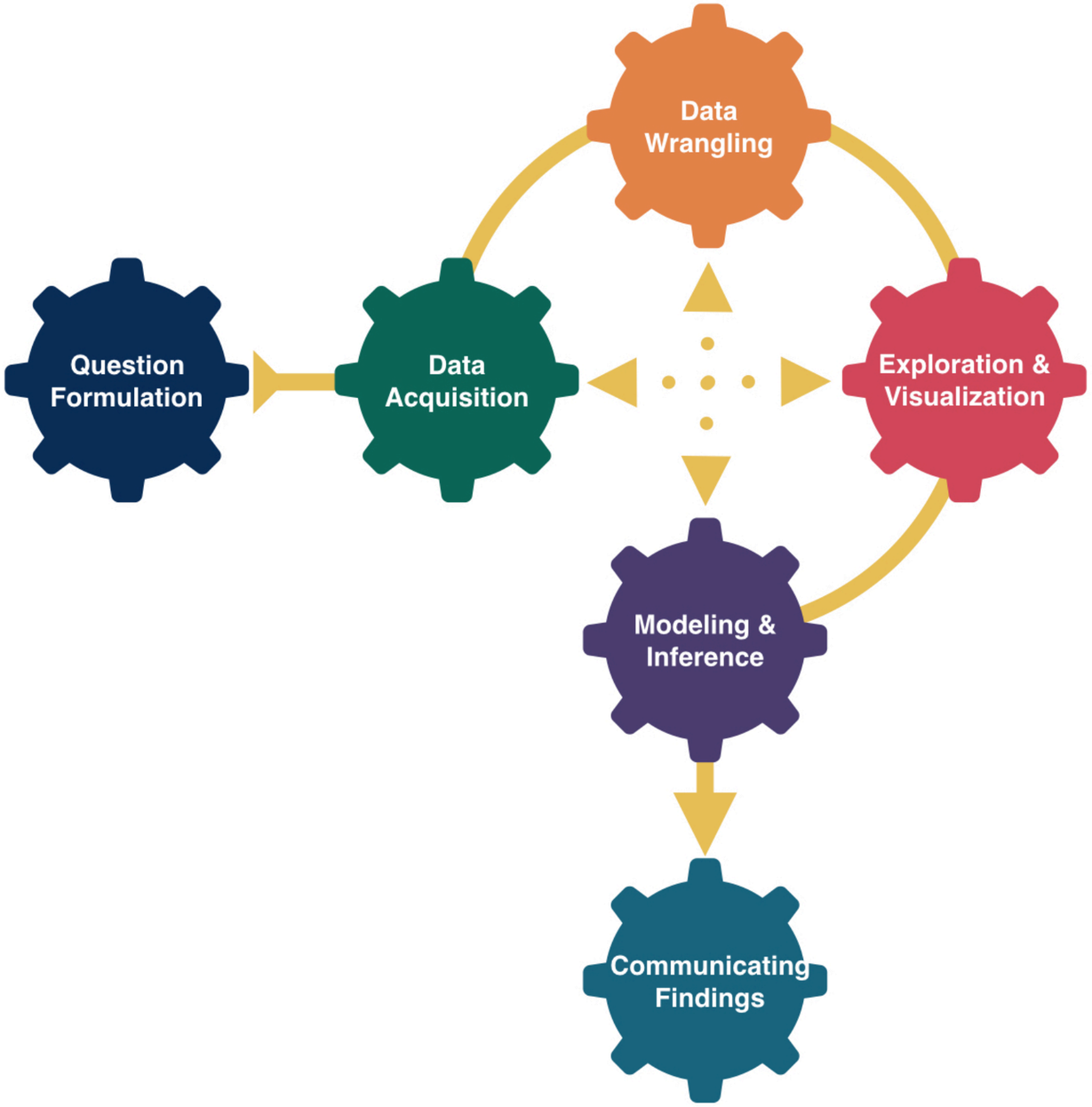
ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

Every time you get a new dataset, spend time exploring the variables.

Example questions:

Next time

- Introduction to data visualization!



Data Visualization

Megan Ayers

Math 141 | Spring 2026

Wednesday, Week 1

Announcements/reminders

- Waitlist movement - let me know if you're still on the waitlist and haven't heard from me
- Readings posted Fridays for the following week, slides right after lecture

Last Time

- Introductions
- Statistical thinking
- Introduced data frames

Goals for Today

- Review data frames
- Motivate data visualizations
- Develop **language** to talk about the components of a graphic
- Practice deconstructing graphics
- Discuss good graphical practices

Data Frames

Data in **spreadsheet**-like format where:

- Rows = Observations/cases
- Columns = Variables

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

- Data from **GPT Detectors Are Biased Against Non-Native English Writers.** *Weixin Liang*, *Mert Yuksekgonul*, *Yining Mao*, *Eric Wu*, *James Zou.* CellPress Patterns and available in the R package `detectors`.

Data Frames

ID	kind	.pred_AI	.pred_class	detector	native	name	model
1	Human	0.9999942	AI	Sapling	No	Real TOEFL	Human
2	Human	0.8281448	AI	Crossplag	No	Real TOEFL	Human
3	Human	0.0002137	Human	Crossplag	Yes	Real College Essays	Human
4	AI	0.0000000	Human	ZeroGPT	NA	Fake CS224N - GPT3	GPT3
5	AI	0.0017841	Human	OriginalityAI	NA	Fake CS224N - GPT3, PE	GPT4
6	Human	0.0001783	Human	HFOpenAI	Yes	Real CS224N	Human

Columns = Variables

Variables: Describe characteristics of the observations

- **Quantitative**: Numerical in nature
- **Categorical**: Values are categories
- **Identification**: Uniquely identify each case

Types of variables

- **Identification**: Uniquely identify each case

- **Quantitative**: Numerical in nature
  - Those that can take a range of values are called **continuous** (e.g., age, income)
  - Those that only take particular (often whole number) values are called **discrete** (e.g., result of a dice roll)
  - Not every variable involving numbers is quantitative!

- **Categorical**: Values represent categories
  - Usually take *non-numeric* values (e.g., Race/Ethnicity or Gender)
  - But can take numeric values! (e.g., zip code)
  - The values that a categorical variable can take are called its **levels**
  - **Ordinal** categorical variables can be ordered (e.g., level of education, variables collected with a Likert scale)
  - Categorical variables that cannot be ordered are called **nominal** (e.g., Race/Ethnicity)

Why construct a graph?

To **explore** the data.

To **summarize** the data.

To showcase **trends** and make **comparisons**.

To tell a compelling **story**.

Doing any of this by only looking at a data frame (even a small one) would be hard!

Challenger

- On January 27th, 1986, engineers from Morton Thiokol recommended NASA delay launch of space shuttle *Challenger* due to cold weather.
  - Believed cold weather impacted the o-rings that held the rockets together.
  - Used 13 charts in their argument.
- After a two hour conference call, the engineer's recommendation was overruled due to lack of persuasive evidence and the launch proceeded.
- The Challenger exploded 73 seconds into launch.

Challenger

Here's one of those charts.

BLOW BY HISTORY

SRM-15 WORST BLOW-BY

- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY

- 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A

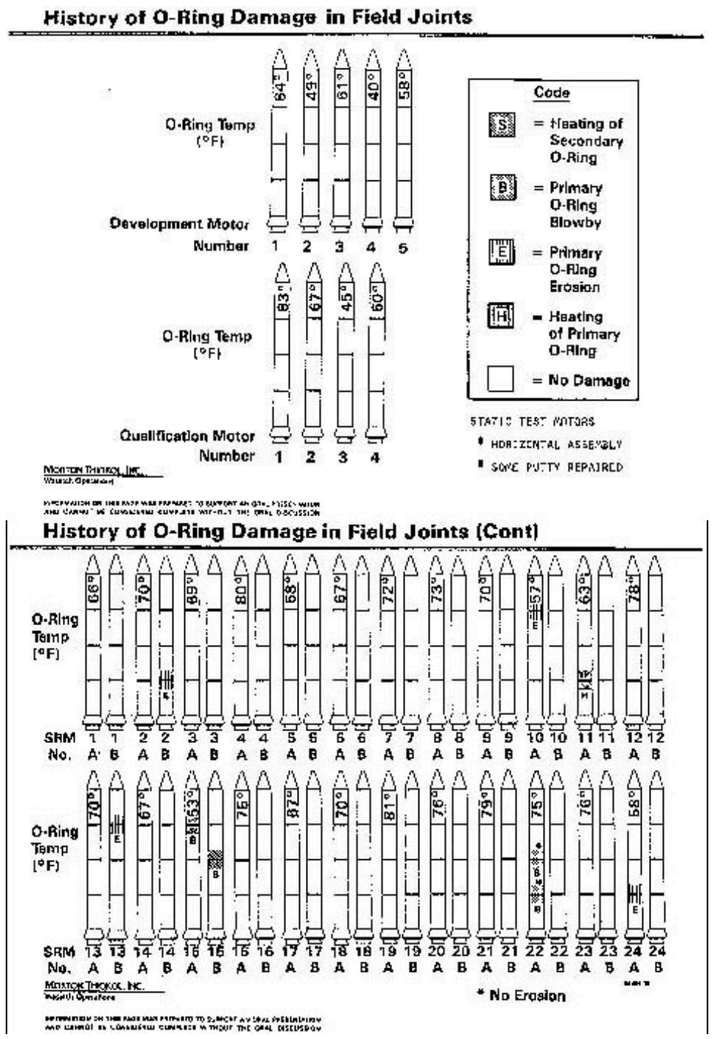
- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES
(DEGREES-F)

MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

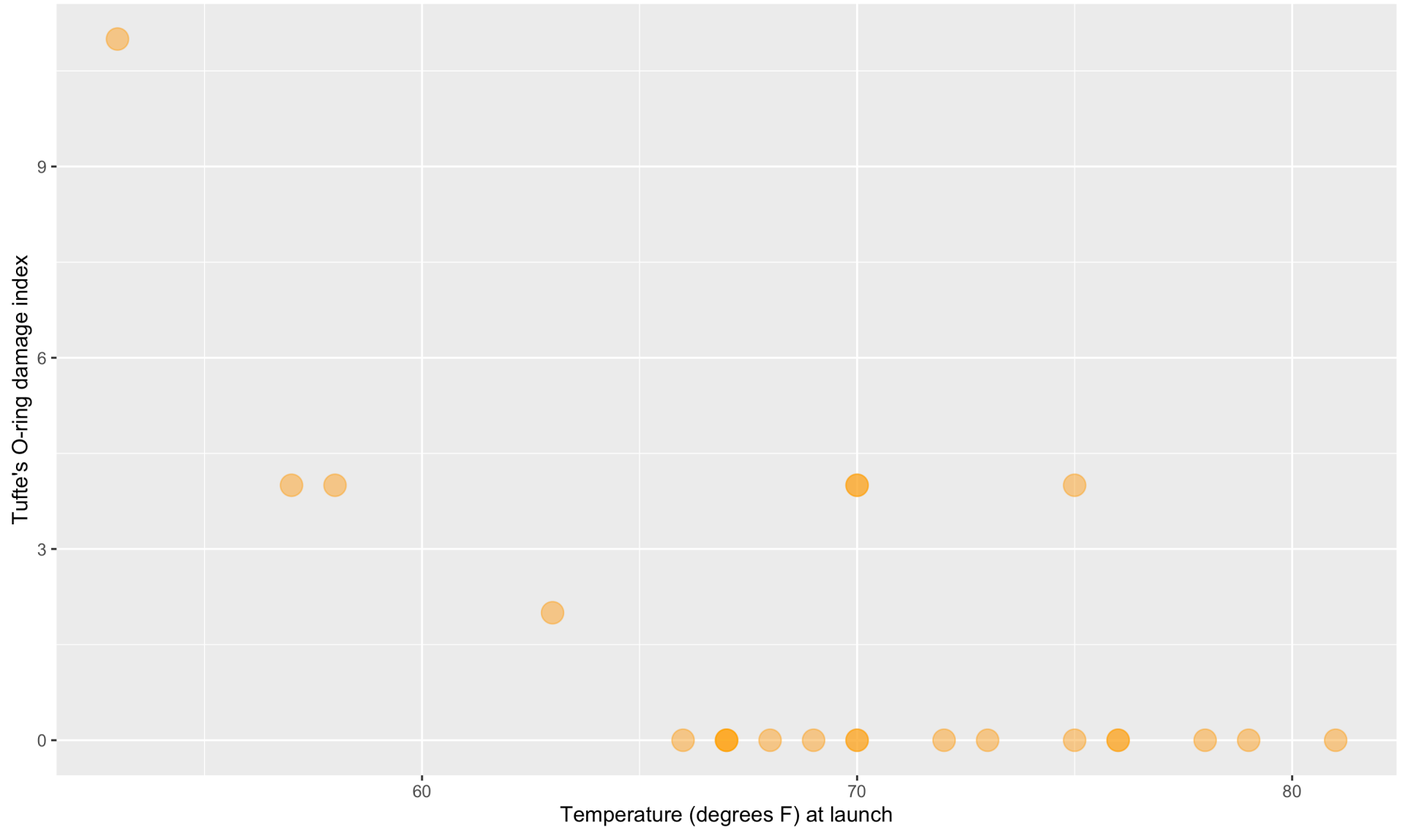
Challenger

Here’s another one of those charts.



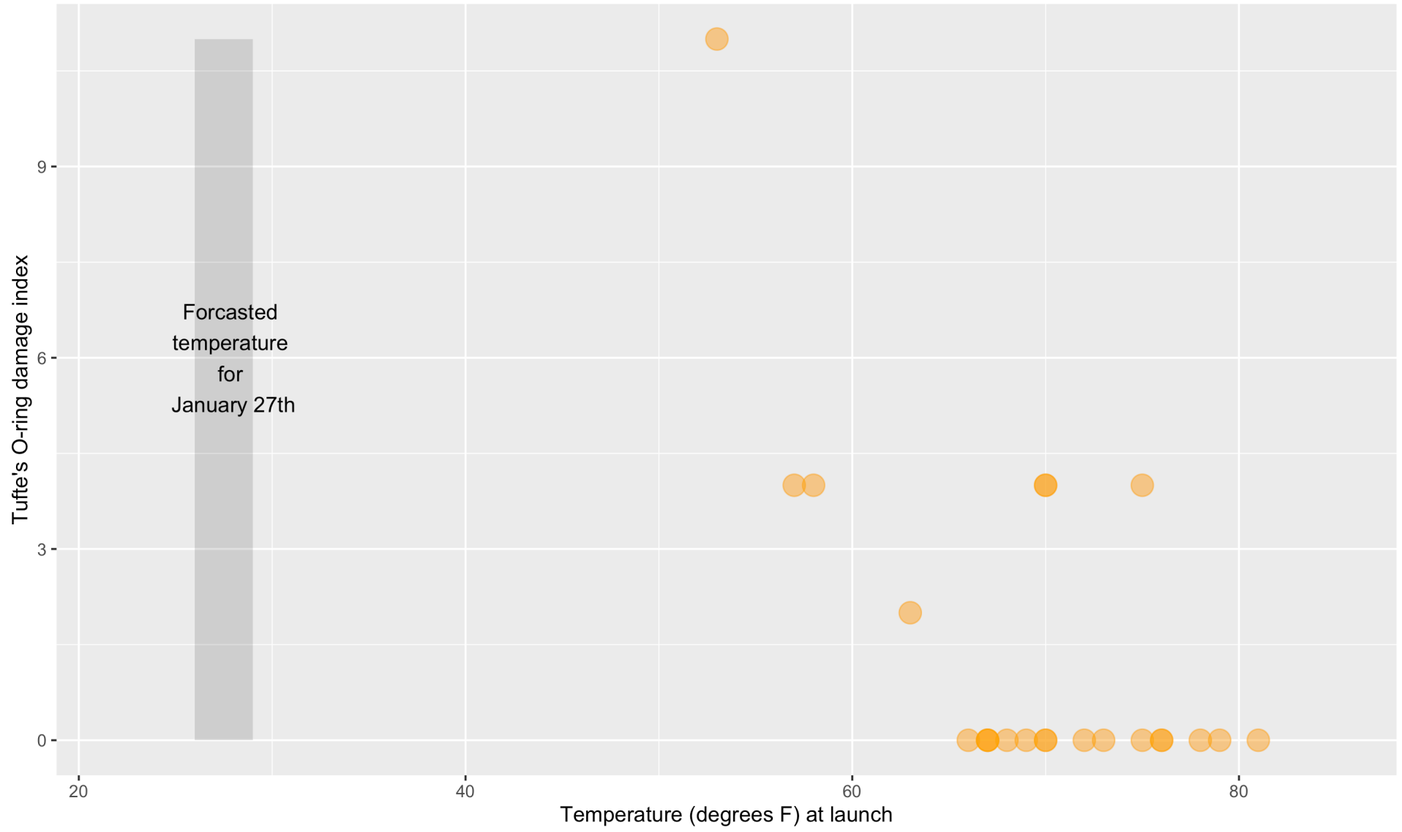
Challenger

Here's a graphic created in R from Statistician Edward Tufte's data.



Challenger

This adaptation is a recreation of Edward Tufte's graphic.



Now let's learn the Grammar of Graphics.

We will use this grammar to:

Decompose and understand existing graphs.

Create our own graphs with the R package `ggplot2`.

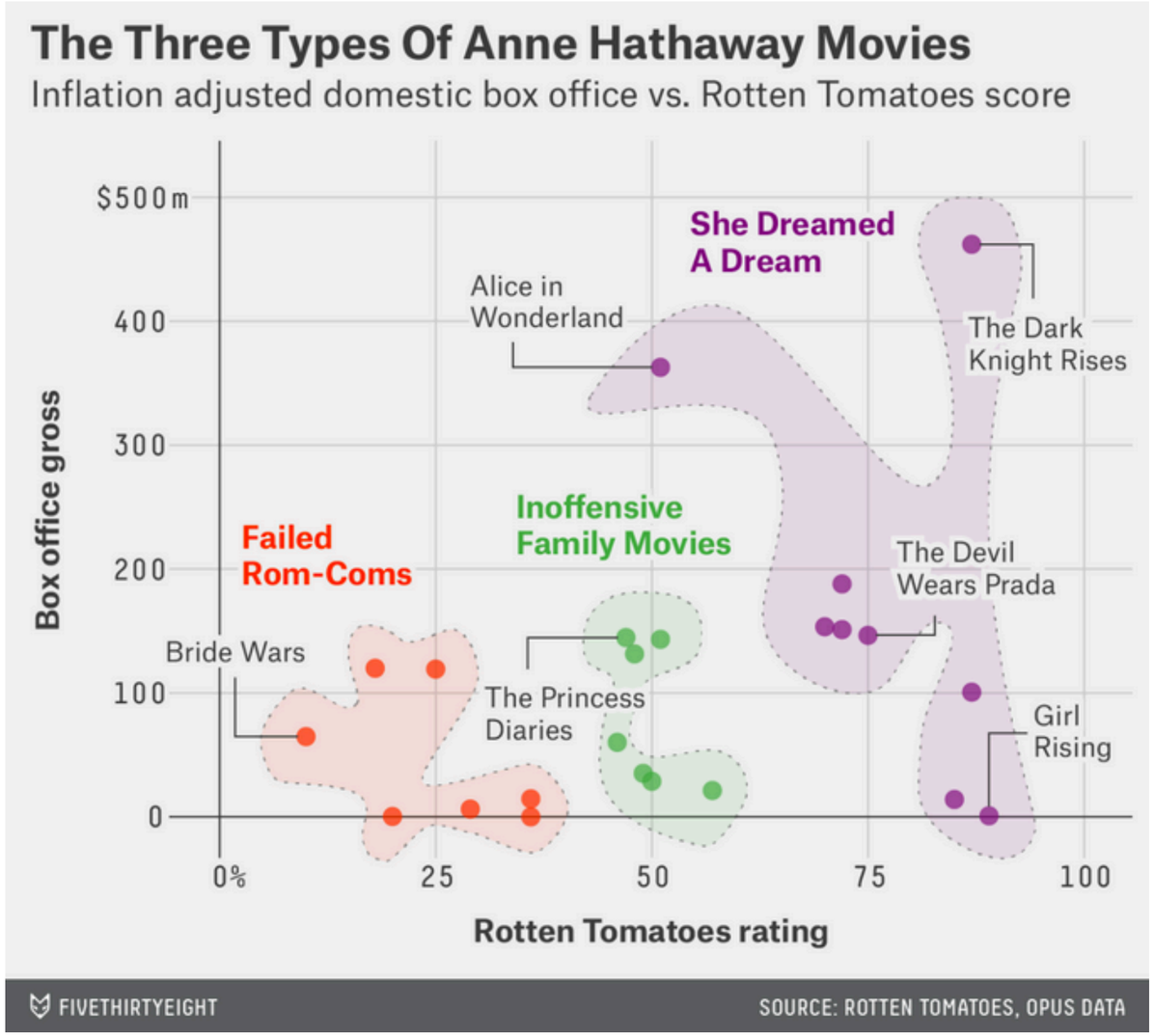
Grammar of Graphics

- **data**: Data frame that contains the raw data
  - Columns are variables used in the graph
- **geom**: Geometric **form** that the data are mapped to.
  - EX: Point, line, bar, text, …
- **aesthetic**: Visual properties of the **geom**
  - EX: X (horizontal) position, y (vertical) position, color, fill, shape
- **scale**: Controls how data are mapped to the visual values of the aesthetic.
  - EX: particular colors, log scale
- **guide**: Legend/key to help user convert visual display back to the data

For right now, we won't focus on the **names** of particular types of graphs (e.g., scatterplot) but on the **elements** of graphs.

Example 1: Think-pair-share

- What are the variables?
- What **geom** (i.e. shape, form) are the variables mapped to?
- What are the **aesthetic**s (visual properties) of the **geom**?
- How is each variable mapped to an **aesthetic**?
- What additional context is provided? Is any missing?
- What story is the graph telling?



Example 2: Think-pair-share

- What are the variables?
- What **geom** are the variables mapped to?
- What are the **aesthetic**s of the **geom**?
- How is each variable mapped to an **aesthetic**?
- What additional context is provided? Is any missing?
- What story is the graph telling?

Sexual harassment charges, by industry

Among charges filed by women, fiscal years 2005-2015

INDUSTRY	CHARGES FILED
Accommodation and food services	4,801
Retail trade	4,380
Health care and social assistance	3,898
Manufacturing	3,741
Office administration and waste management	2,350
Public administration	2,239
Professional, scientific and technical services	1,944
Transportation and warehousing	1,601
Finance and insurance	1,380
Educational services	1,340
Other services (except public administration)	1,003
Information	962
Construction	774
Wholesale trade	752
Real estate rental and leasing	611
Arts, entertainment and recreation	537
Agriculture, forestry, fishing and hunting	276
Management of companies and enterprises	213
Utilities	211
Mining	157

Not including 35,304 charges filed without a specified industry

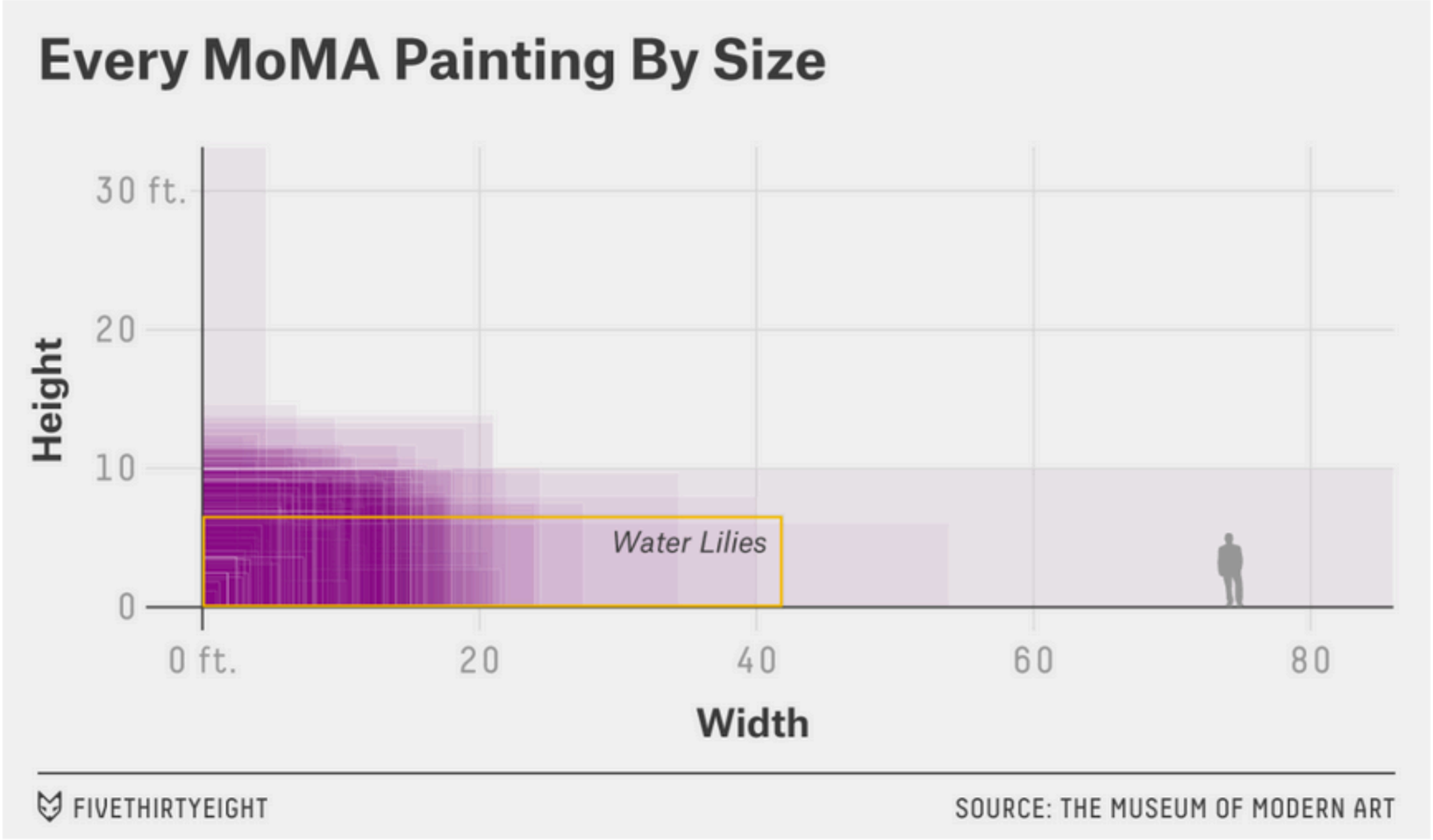
SOURCE: EQUAL EMPLOYMENT OPPORTUNITY COMMISSION

Visualization Considerations

What additional context should my graphs have?

- For context, at a minimum include
  - Axis labels (with units reported).
  - Legends.
  - Data source.
- Think about the **stories/questions** your visualization answers.
- Determine what **context/background information** your viewer needs.
- Visualizing data involves **editorial choices**.
  - What to highlight.
  - What comparisons to make easy to see.
  - What scales to use.

Context Example



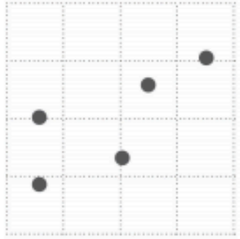
Water Lilies

What visual cues are easier to compare?

Visual cues

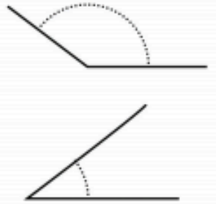
When you visualize data, you encode values to shapes, sizes, and colors.

Position
Where in space the data is

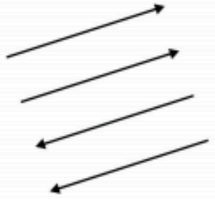


Length
How long the shapes are

Angle
Rotation between vectors



Direction
Slope of a vector in space



Shapes
Symbols as categories

Area
How much 2-D space

Volume
How much 3-D space

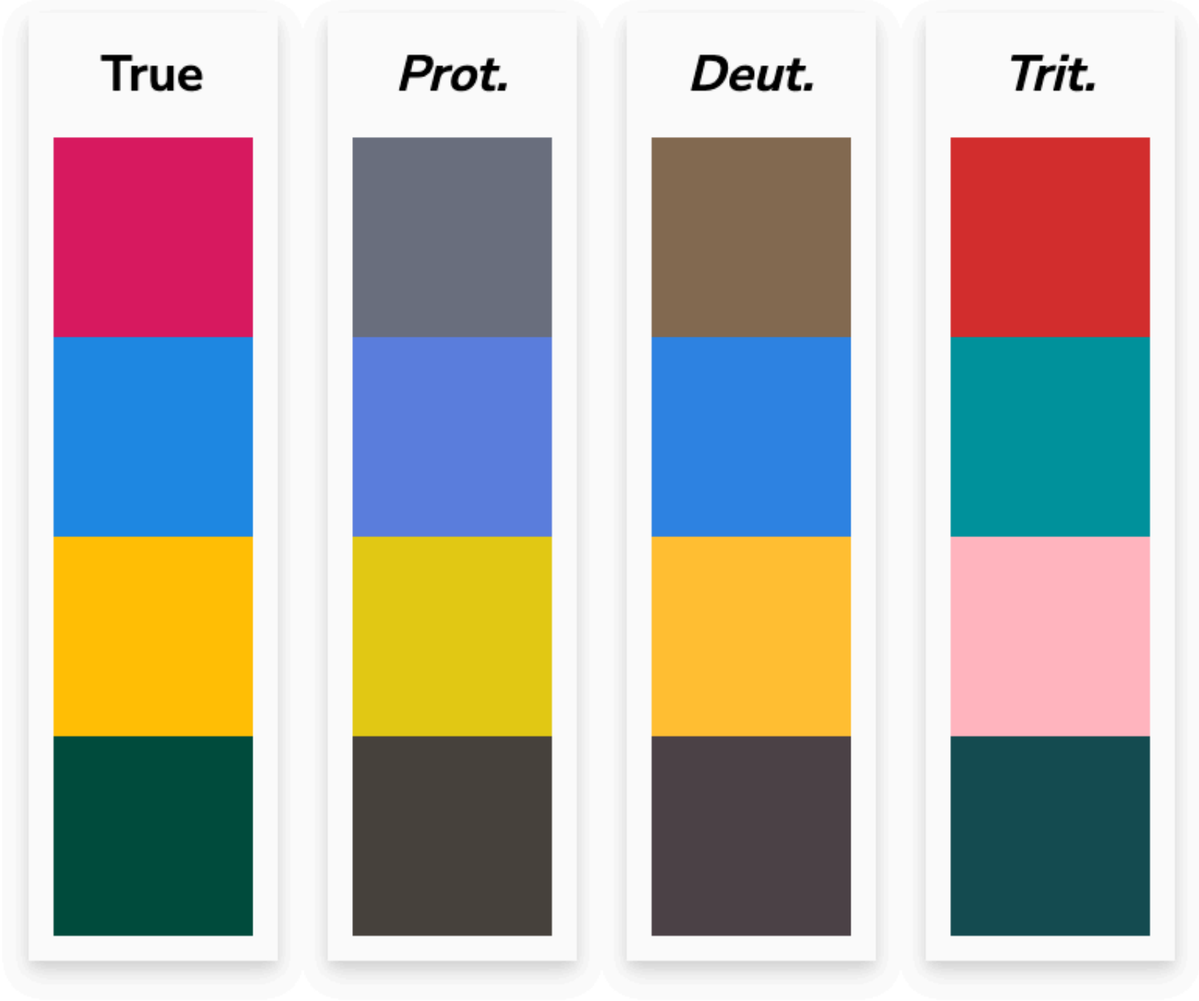
Color saturation
Intensity of a color hue

Color hue
Usually referred to as color

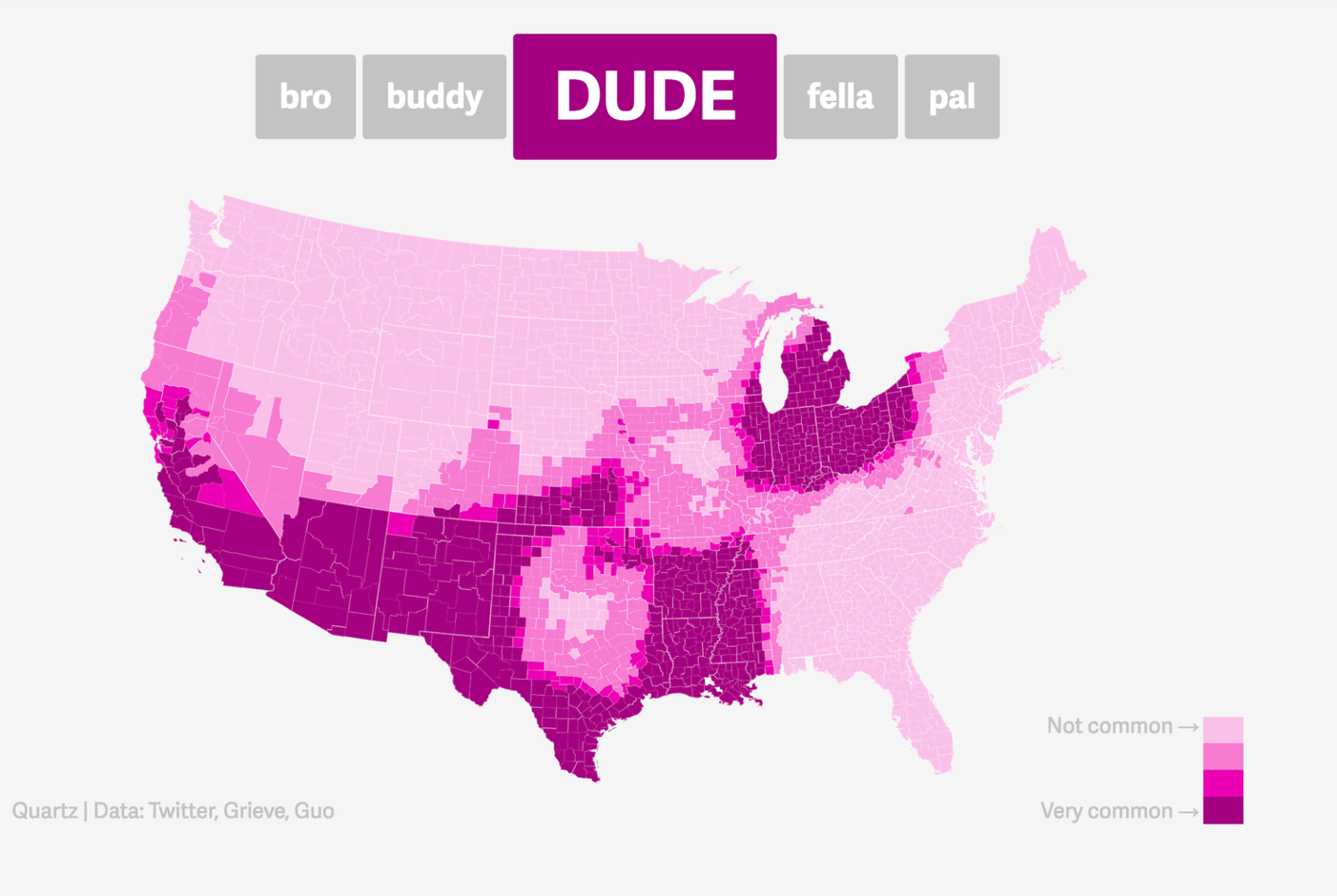
FIGURE 3-3 *Visual cues*

What to consider with color?

Consider color blindness (click here for a helpful tool for picking colors).

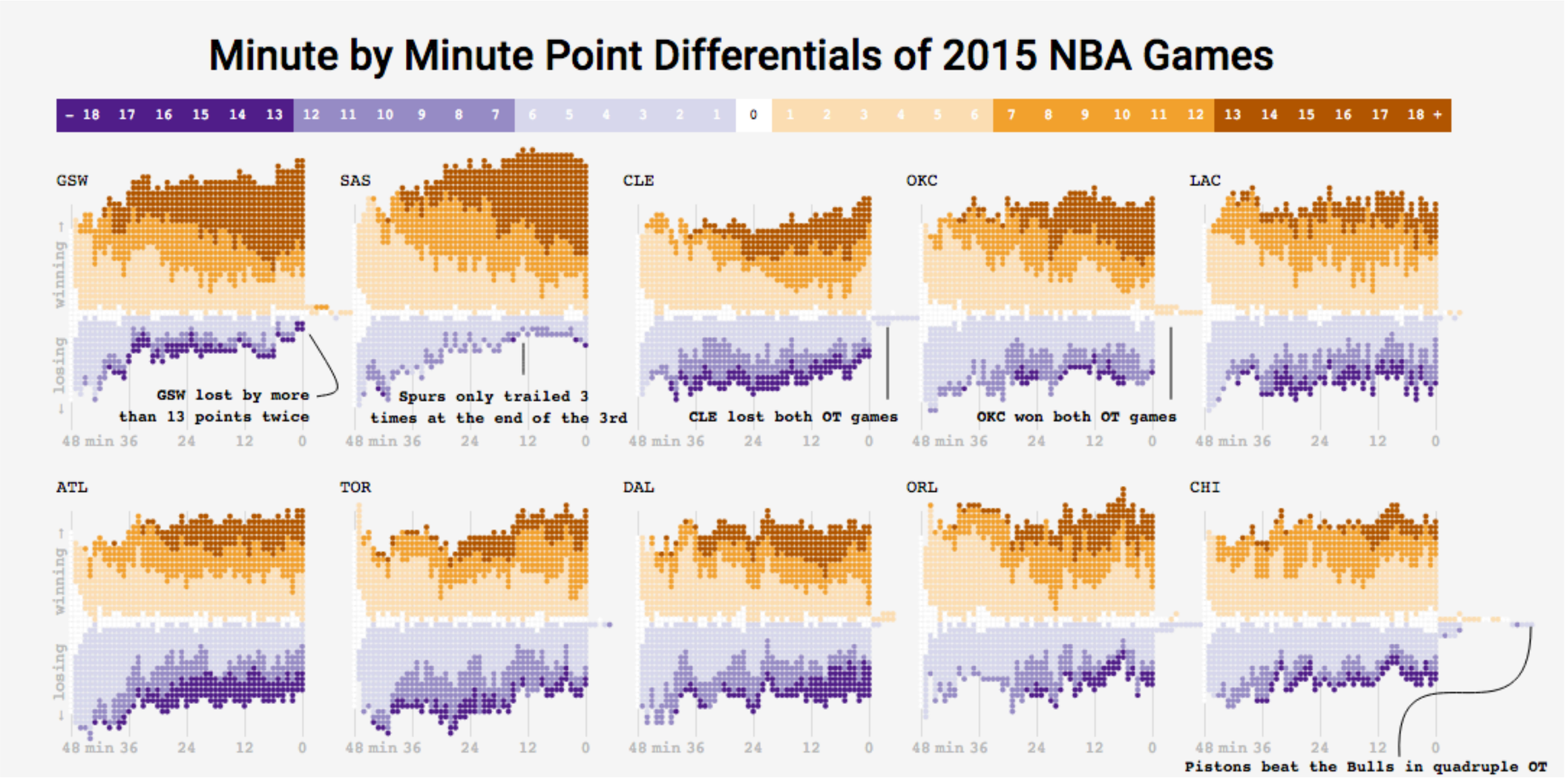


Color Palettes – Sequential



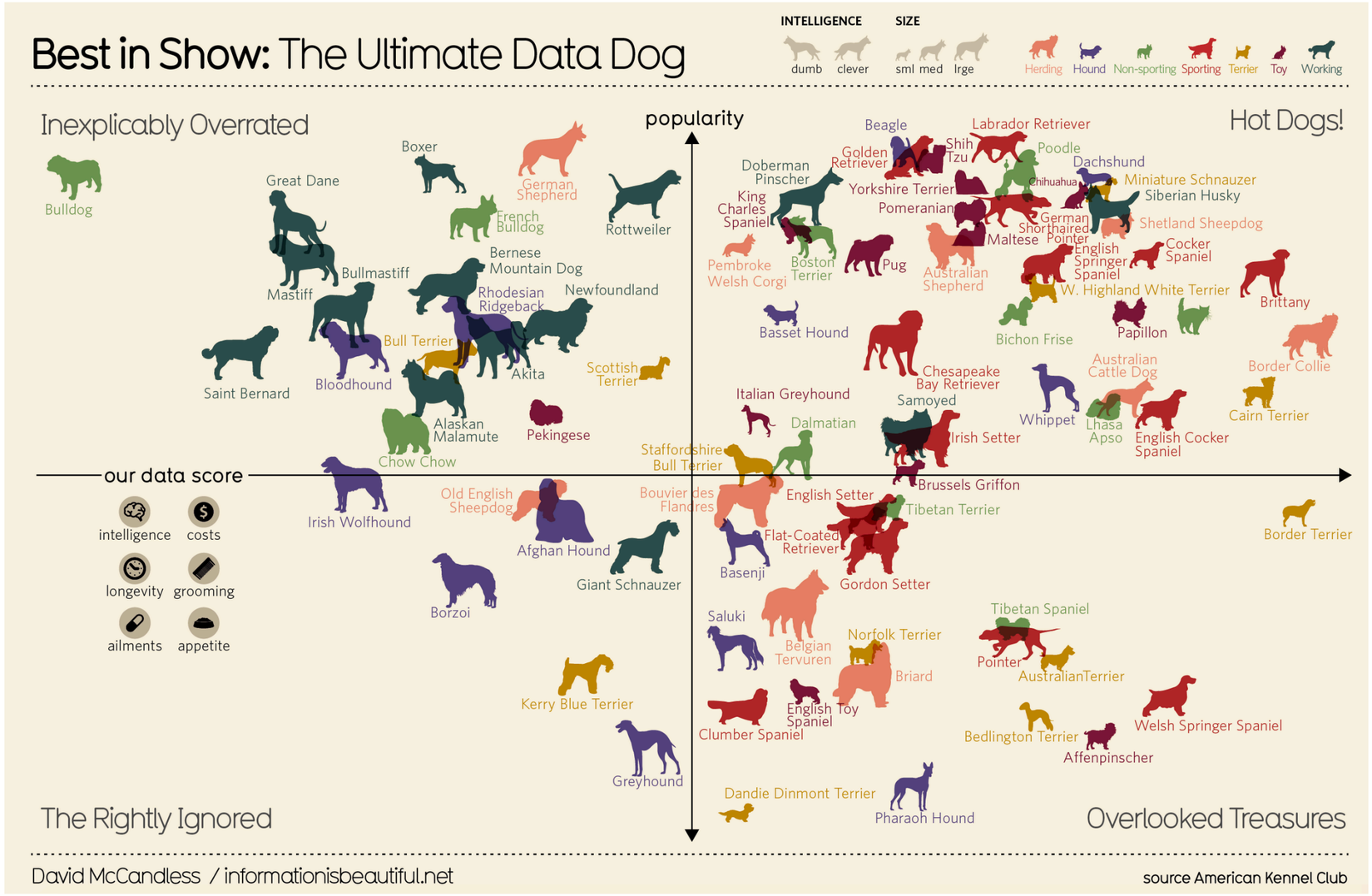
Maps, like the Dude map are also a great way to provide context!

Color Palettes – Diverging



Adam Pearce's 2015 NBA Games

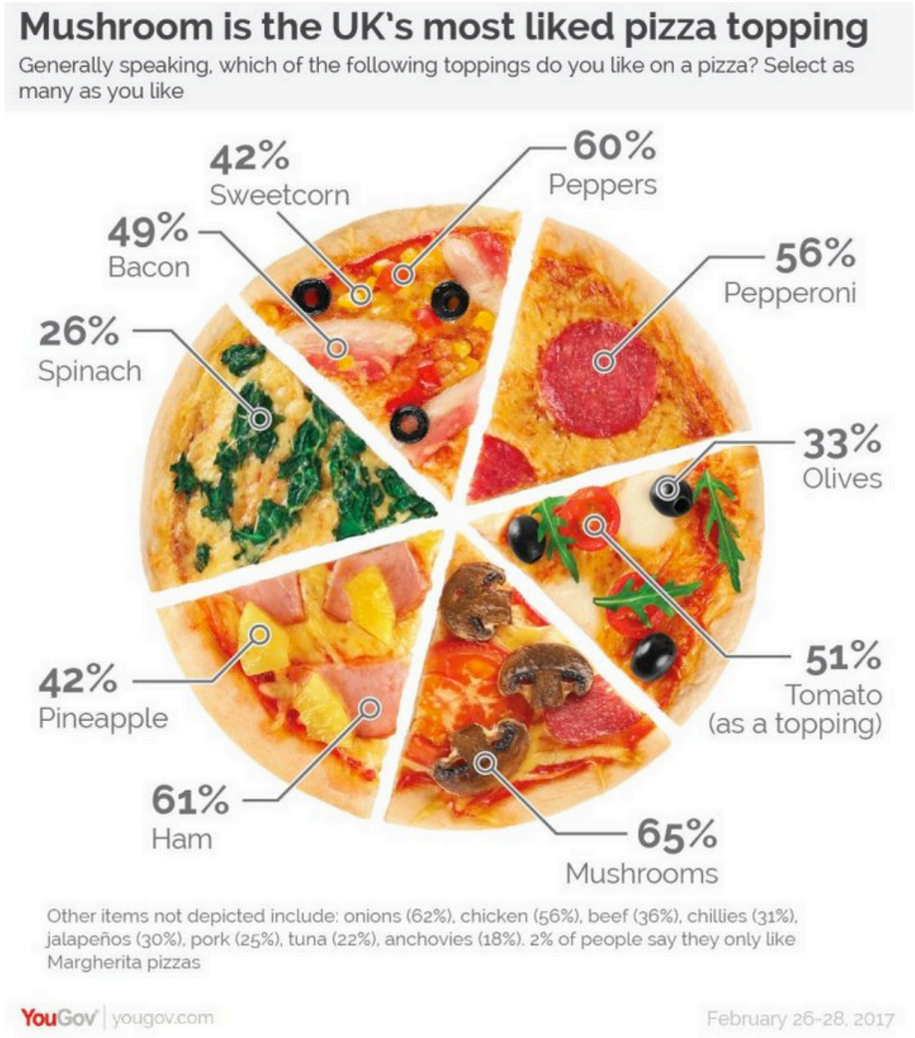
Color Palettes – Qualitative



information is beautiful's Best in Show

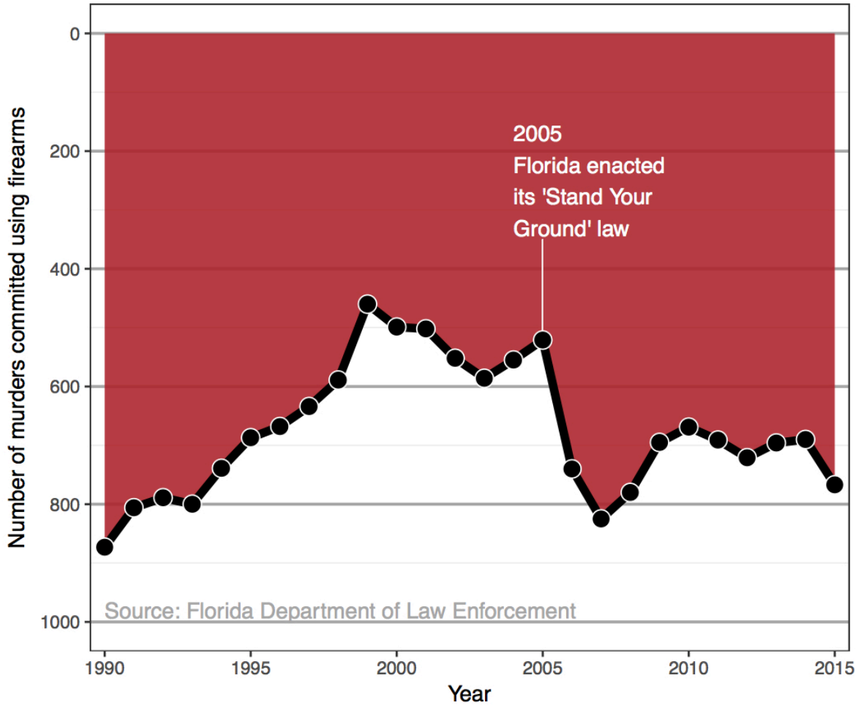
Bad Graphics

Because of all the design choices, it is much easier to make a bad graph than a good graph.



Misleading Graphics

Be careful that your design choices don't cause your viewer to draw incorrect conclusions about the data:



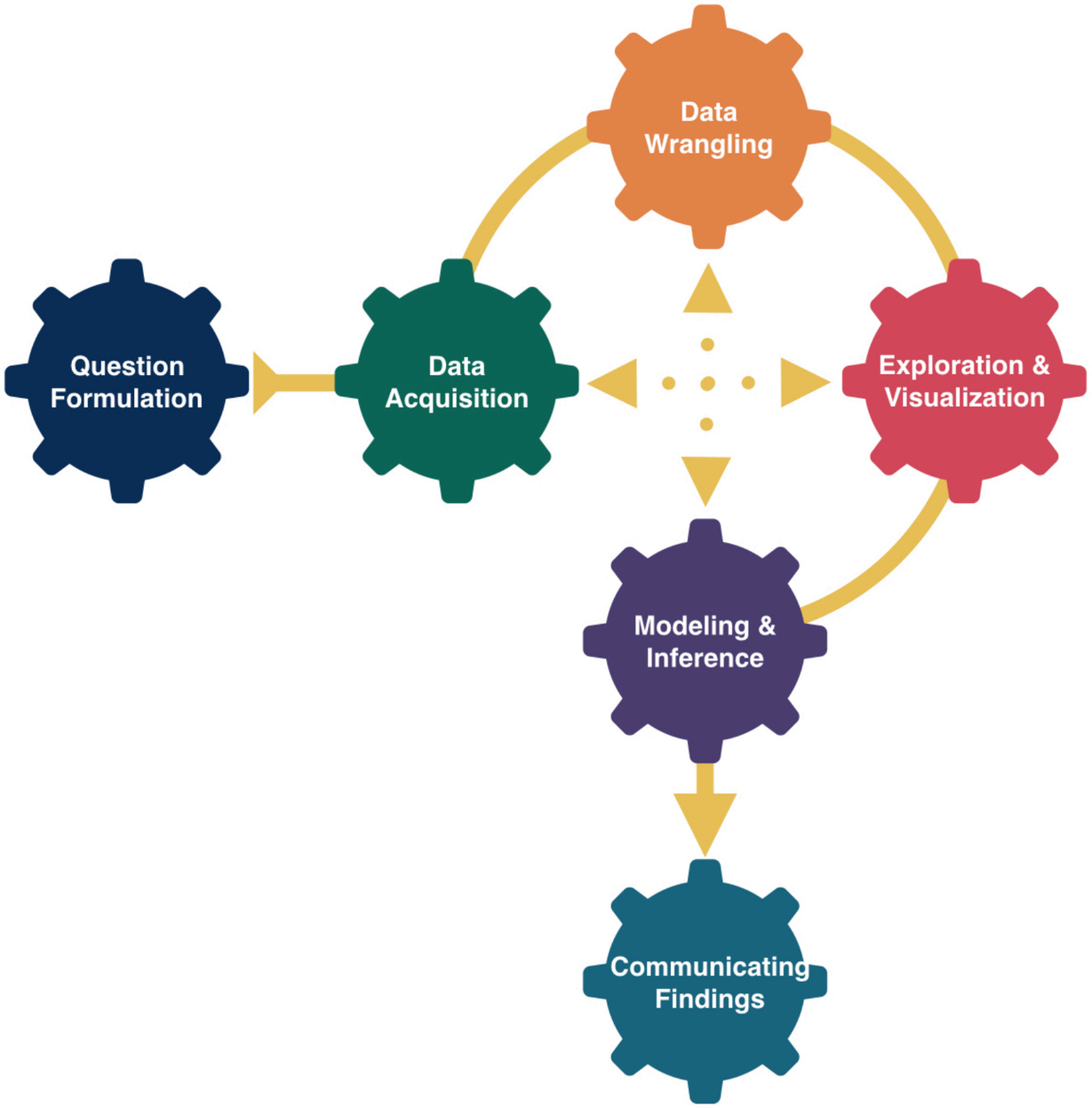
- Just letting the software make all the design choices can still lead to misleading graphs (recall the Georgia COVID graph).

Summary Thoughts on Graphical Considerations

- Good graphics are one's where the findings and insights are **obvious** to the viewer.
  - Add information and key **context**.
- Facilitate the **comparisons** that correspond to the research question.
  - Recall the three Georgia COVID counts graphs from Day 1!
- Data visualizations are **not neutral**.
- It is easier to see the differences and similarities between different types of graphics if we learn the **grammar of graphics**.
- Practicing **decomposing** graphics should make it easier for us to **compose** our own graphics.

Next time

- Tomorrow in lab: intro to using RStudio to code and explore data frames
- Friday: We'll learn about the `ggplot2` package so that we can use the grammar of graphics to create beautiful visualizations!



Data Visualization: the 5 Named Graphs with `ggplot2`

Reminders/Announcements

- If you plan to request academic accommodations, please submit these through the **DAR student portal**
- Course assistant office hours are now on Moodle
- The DataLab @ Reed
- First homework assignment is posted!

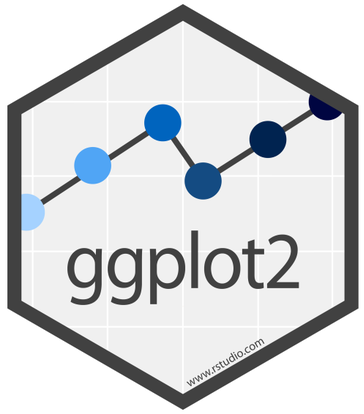
Last Time

- Data frames
- Motivation for data visualizations
- “Grammar” of graphics and good graphical practices

Goals for Today

- Recall our motivation for good graphics
- Learn the general structure of `ggplot2`
- Learn five standard graphs for numerical/quantitative data:
  - **Histogram**: one numerical variable
  - **Boxplot**: one numerical variable
  - **Barplot**: one numerical variable and at least one categorical variable
  - **Scatterplot** and **Linegraph**: two numerical variables

Load Necessary Packages

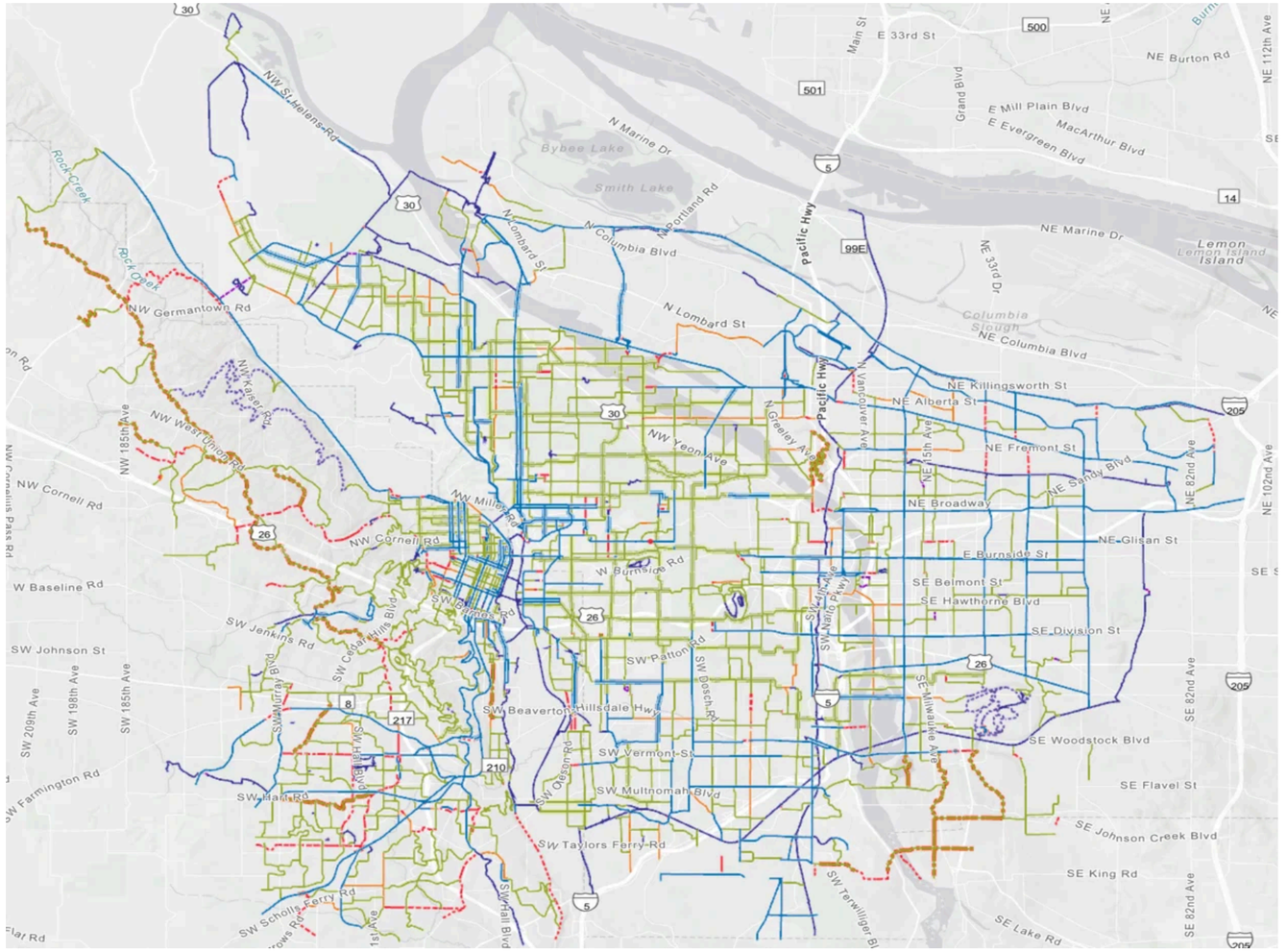


`ggplot2` is part of this collection of data science packages.

```
1 # Load necessary packages
2 library(tidyverse)
```

Also, above is an example of a code comment: `# Load necessary packages`

Data Setting: Portland Bikeshare Data



Import the Data

```r
biketown <- read.csv("data/biketown.csv")

# Inspect the data
glimpse(biketown)
```

```
Rows: 9,999
Columns: 19
$ RouteID          <int> 4074085, 3719219, 3789757, 3576798, 3459987, 3947695,…
$ PaymentPlan      <chr> "Subscriber", "Casual", "Casual", "Subscriber", "Casu…
$ StartHub         <chr> "SE Elliott at Division", "SW Yamhill at Director Par…
$ StartLatitude    <dbl> 45.50513, 45.51898, 45.52990, 45.52389, 45.53028, 45.…
$ StartLongitude   <dbl> -122.6534, -122.6813, -122.6628, -122.6722, -122.6547…
$ StartDate        <chr> "8/17/2017", "7/22/2017", "7/27/2017", "7/12/2017", "…
$ StartTime        <chr> "10:44:00", "14:49:00", "14:13:00", "13:23:00", "19:3…
$ EndHub           <chr> "Blues Fest - SW Waterfront at Clay - Disabled", "SW …
$ EndLatitude      <dbl> 45.51287, 45.52142, 45.55902, 45.53409, 45.52990, 45.…
$ EndLongitude     <dbl> -122.6749, -122.6726, -122.6355, -122.6949, -122.6628…
$ EndDate          <chr> "8/17/2017", "7/22/2017", "7/27/2017", "7/12/2017", "…
$ EndTime          <chr> "10:56:00", "15:00:00", "14:42:00", "13:38:00", "20:3…
```

Inspect the Data

```
1 # Look at first few rows
2 head(biketown)
```

```
  RouteID PaymentPlan                    StartHub StartLatitude StartLongitude
1 4074085  Subscriber      SE Elliott at Division      45.50513      -122.6534
2 3719219      Casual SW Yamhill at Director Park      45.51898      -122.6813
3 3789757      Casual          NE Holladay at MLK      45.52990      -122.6628
4 3576798  Subscriber             NW Couch at 2nd      45.52389      -122.6722
5 3459987      Casual    NE 11th at Holladay Park      45.53028      -122.6547
6 3947695      Casual          SW Moody at Thomas      45.49429      -122.6719
  StartDate StartTime                                      EndHub EndLatitude
1 8/17/2017  10:44:00 Blues Fest - SW Waterfront at Clay - Disabled    45.51287
2 7/22/2017  14:49:00                                SW 2nd at Pine    45.52142
3 7/27/2017  14:13:00 NE Alberta at NE 29th/30th - Community Corral    45.55902
4 7/12/2017  13:23:00                            NW Raleigh at 21st    45.53409
5  7/3/2017  19:30:00                            NE Holladay at MLK    45.52990
6  8/8/2017  10:01:00                              SW 3rd at Ankeny    45.52248
```

What does a row represent here?

Inspect the Data

```
# Determine type
# To access one variable: dataset$variable
class(biketown$BikeName)
```

```
[1] "character"
```

```
class(biketown$Distance_Miles)
```

```
[1] "numeric"
```

```
class(biketown)
```

```
[1] "data.frame"
```

Grammar of Graphics

- **data**: Data frame that contains the raw data
  - Variables used in the graph
- **geom**: Geometric **shape** that the data are mapped to.
  - EX: Point, line, bar, text, …
- **aesthetic**: Visual properties of the **geom**
  - EX: X (horizontal) position, y (vertical) position, color, fill, shape
- **scale**: Controls how data are mapped to the visual values of the aesthetic.
  - EX: particular colors, log scale
- **guide**: Legend/key to help user convert visual display back to the data

`ggplot2` example code

Guiding Principle: We will map variables from the **data** to the **aes**thetic attributes (e.g. location, size, shape, color) of **geom**etric objects (e.g. points, lines, bars).

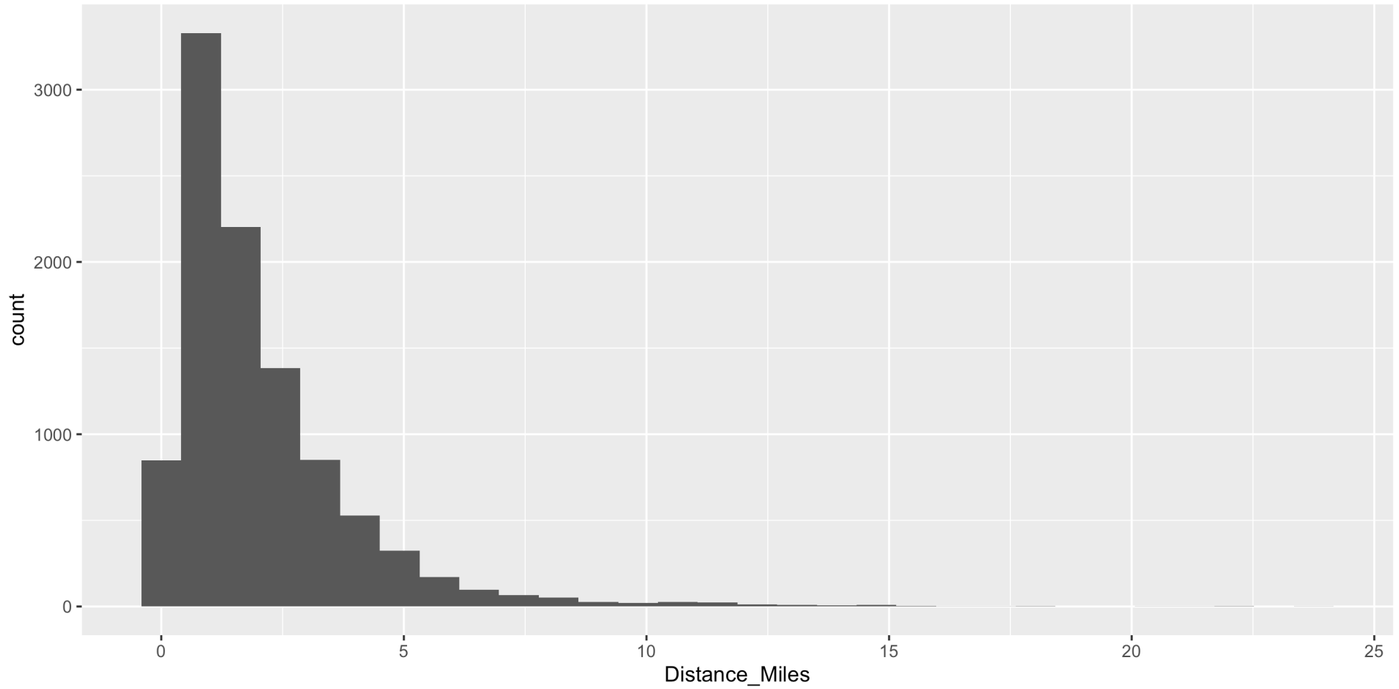
```
ggplot(data = ---, mapping = aes(---)) +
  geom_---(---)
```

- There are other layers, such as `scale_---_---()` and `labs()`, but we will wait on those.

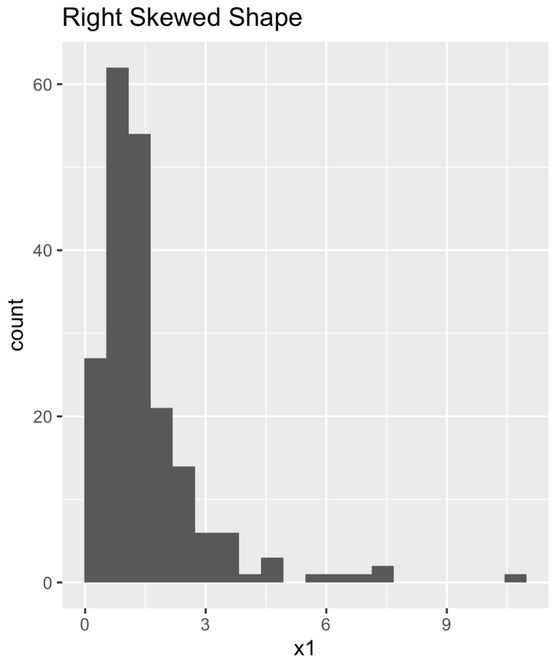
- We are about to touch on many details of graphs - focus on recognizing this general pattern, revisit the slides to refresh on the coding specifics

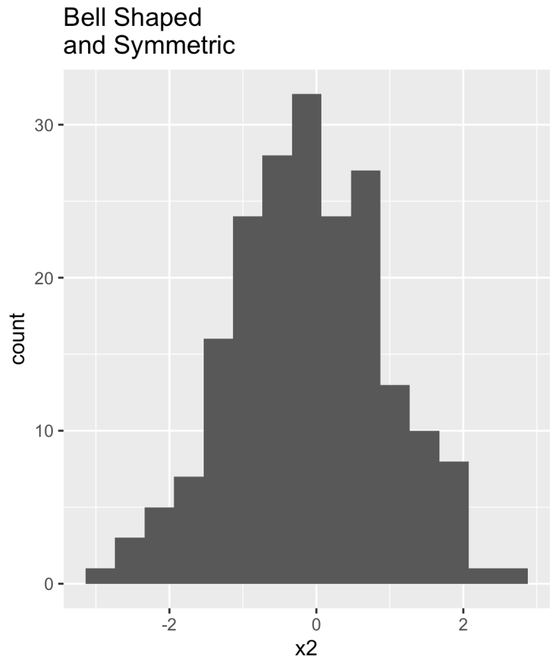
Histograms

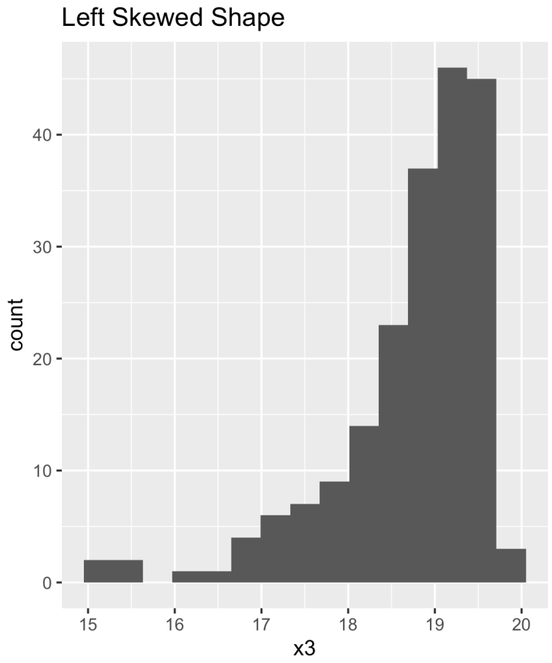
- Binned counts of data.
- Great for assessing data distribution and shape.
- **Question**: are histograms used for *quantitative* or *categorical* variables?
- **Answer**: *Quantitative*.



Data Shapes

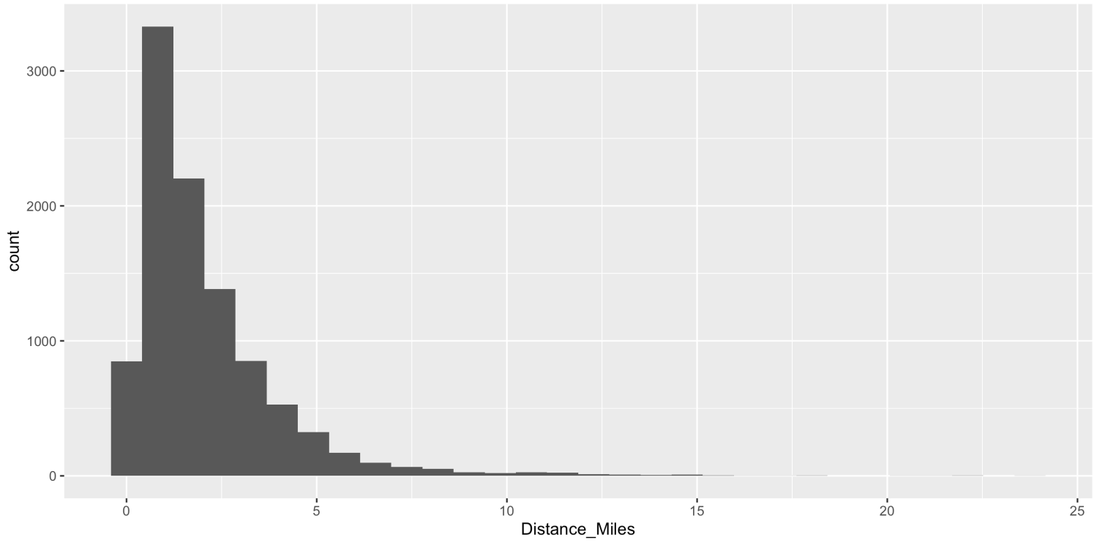

Right Skewed Shape
count
x1
0
20
40
60
0
3
6
9


Bell Shaped
and Symmetric
count
x2
0
10
20
30
-2
0
2


Left Skewed Shape
count
x3
0
10
20
30
40
15
16
17
18
19
20

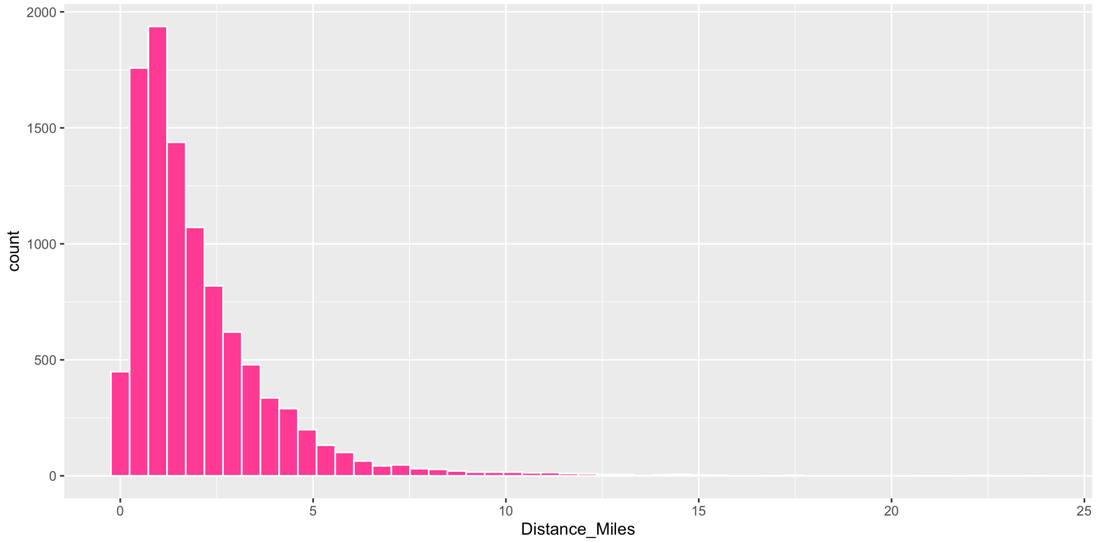
Histograms

```
# Create histogram
ggplot(data = biketown,
       mapping = aes(x = Distance_Miles)) +
  geom_histogram()
```



Histograms

```
# Create histogram
ggplot(data = biketown,
       mapping = aes(x = Distance_Miles)) +
  geom_histogram(color = "white",
                 fill = "violetred1",
                 bins = 50)
```

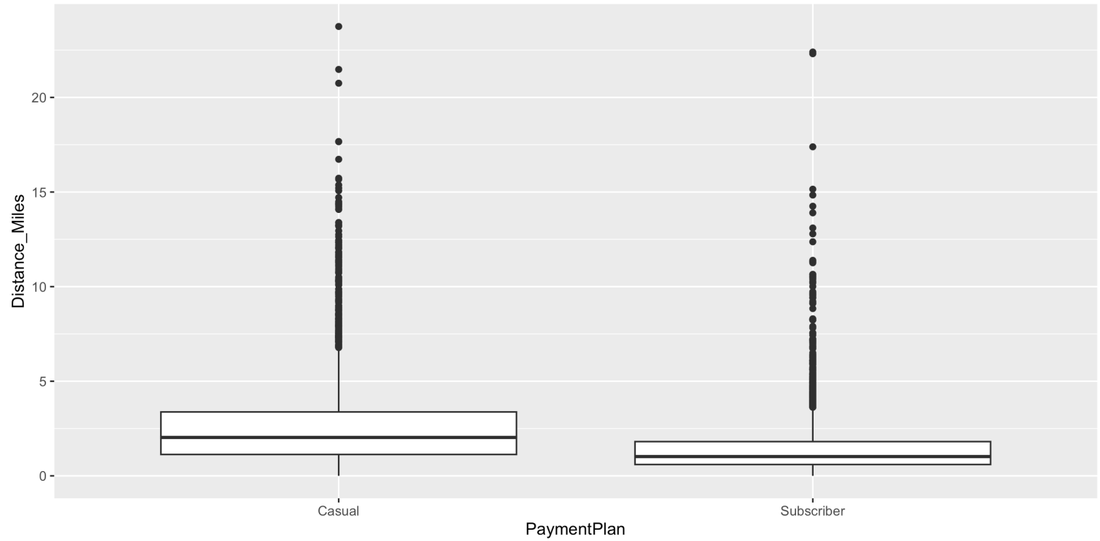


- **mapping** to a variable goes in `aes()`
- **setting** to a specific, *constant*, value goes in the `geom_---()`

- Does the right **tail** of this distribution make sense?

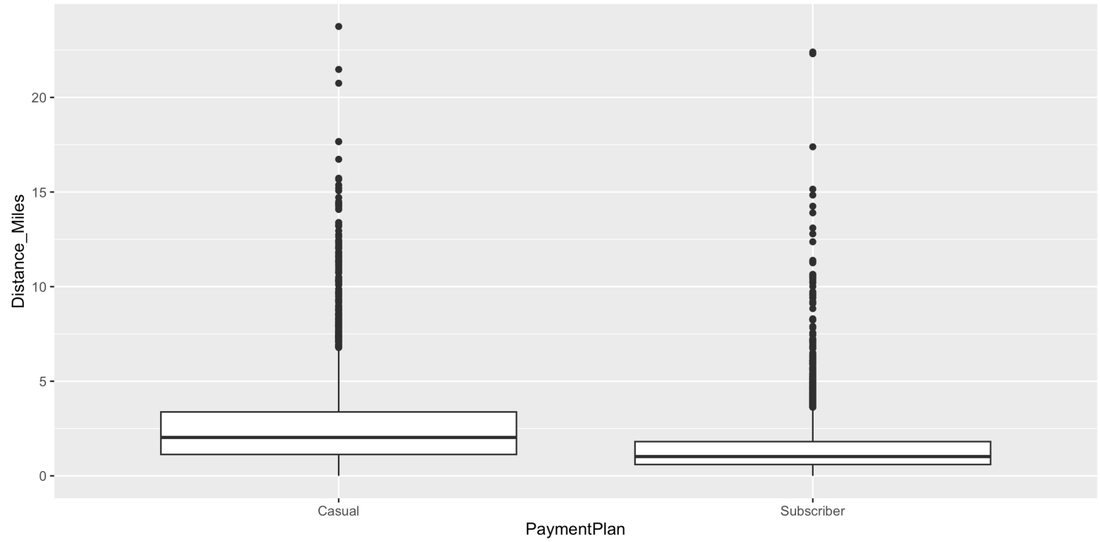
Boxplots

- **Five number summary**:
  - Minimum
  - First quartile (Q1)
  - Median
  - Third quartile (Q3)
  - Maximum
- Interquartile range (IQR) = Q3 − Q1
- Outliers: **unusual** points
  - Boxplot defines unusual as being beyond $1.5 * IQR$ from $Q1$ or $Q3$.
- Whiskers: reach out to the furthest point that is NOT an outlier



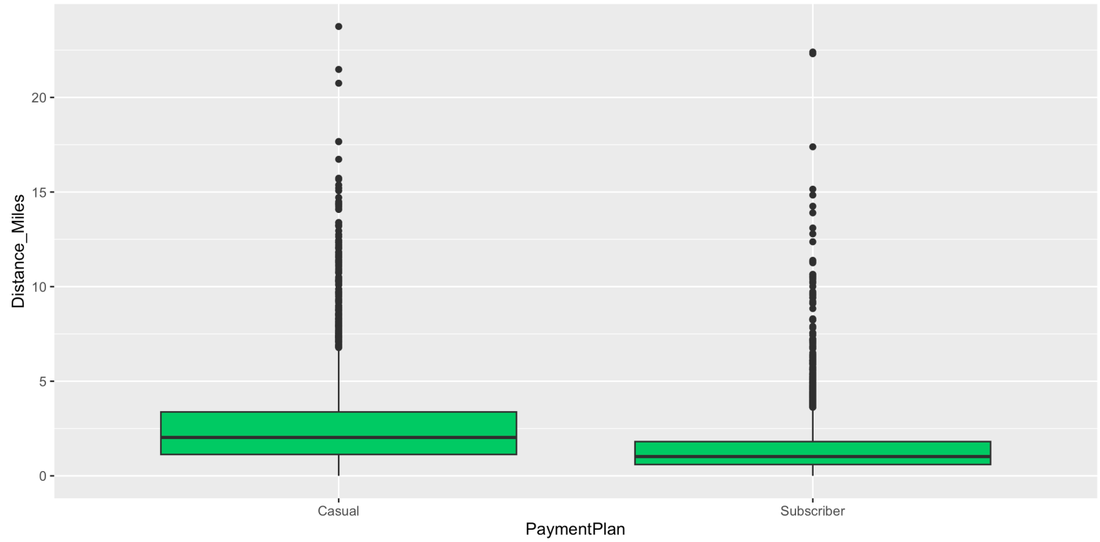
Boxplots

```
# Create boxplot
ggplot(data = biketown,
       mapping = aes(x = PaymentPlan,
                     y = Distance_Miles)) +
  geom_boxplot()
```



Boxplots

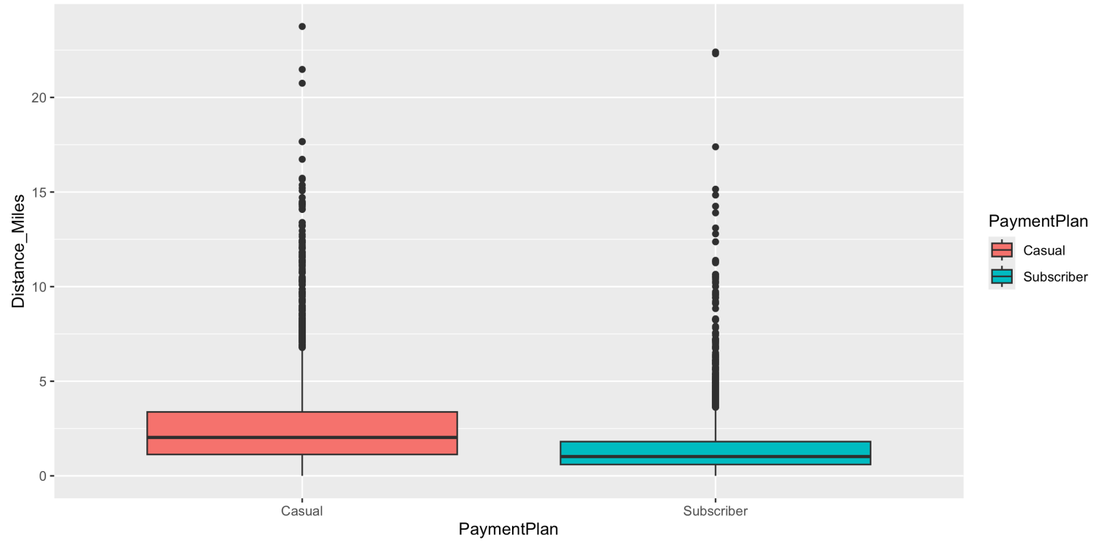
```
ggplot(data = biketown,
       mapping = aes(x = PaymentPlan,
                     y = Distance_Miles)) +
  geom_boxplot(fill = "springgreen3")
```



- Is this `fill` an `aes`thetic mapping?

Boxplots

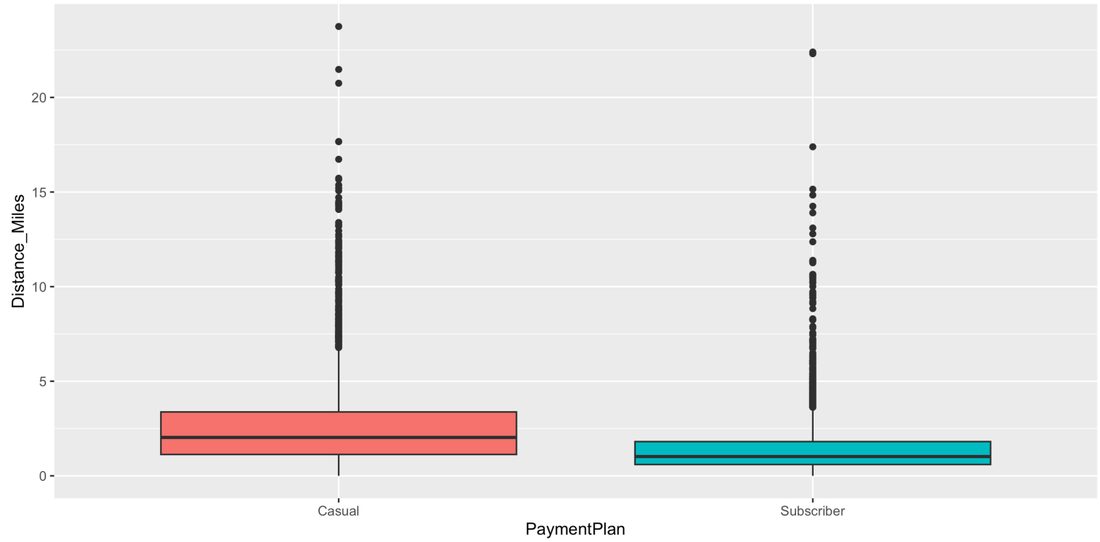
```
ggplot(data = biketown,
       mapping = aes(x = PaymentPlan,
                     y = Distance_Miles,
                     fill = PaymentPlan)) +
  geom_boxplot()
```



- Is this `fill` an `aes`thetic mapping?
- What variable is mapped to `fill`?

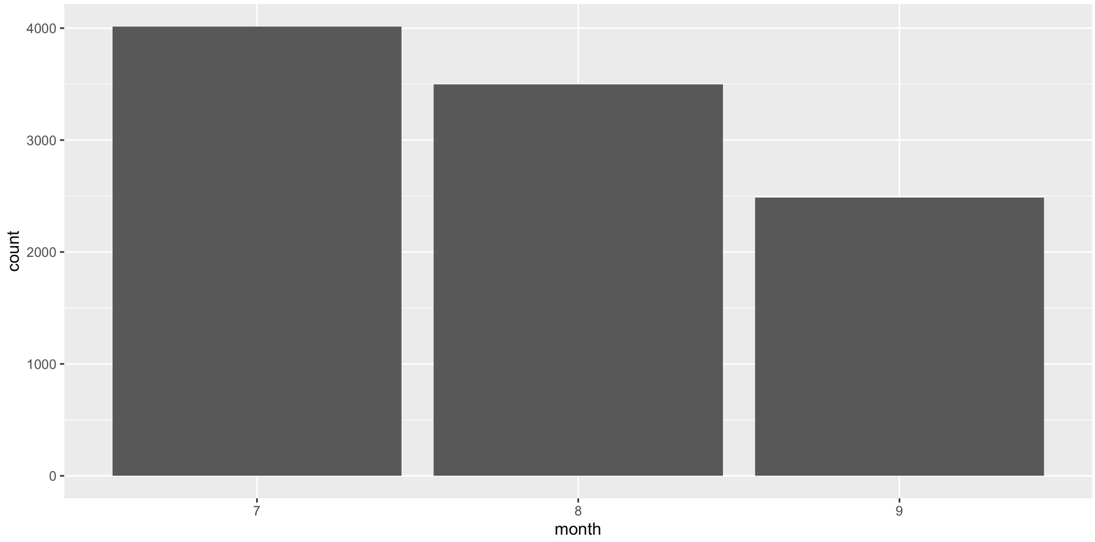
Boxplots

```
ggplot(data = biketown,
       mapping = aes(x = PaymentPlan,
                     y = Distance_Miles,
                     fill = PaymentPlan)) +
  geom_boxplot() +
  guides(fill = "none")
```



Barplots

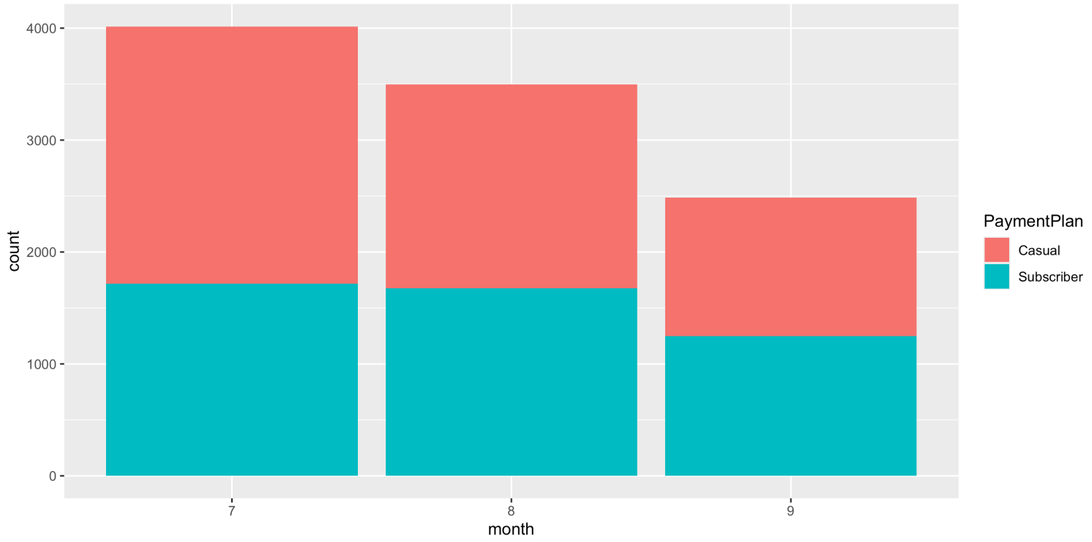
```
ggplot(data = biketown,
       mapping = aes(x = month)) +
  geom_bar()
```



- Boxplots and histograms show the overall distribution of *quantitative* variables
- Barplots show the distribution of quantitative variables within distinct levels defined by a *categorical* variable

Barplots

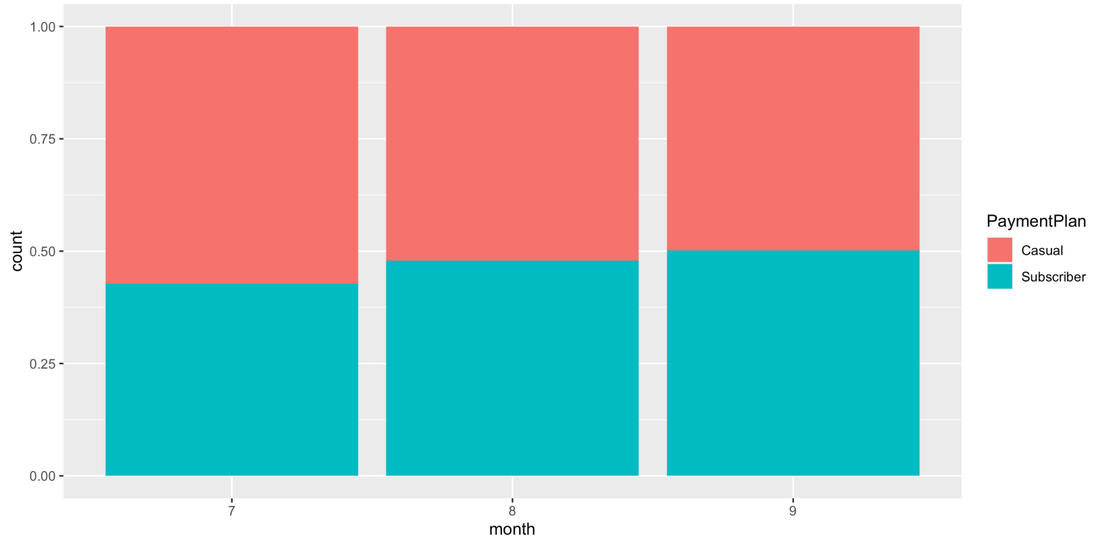
```
1 ggplot(data = biketown,
2        mapping = aes(x = month,
3                      fill = PaymentPlan)) +
4   geom_bar()
```



- Barplots can also show the joint distribution of *two* categorical variables via the color or fill **aes**thetic.
- Here, each bar is divided into separate counts with respect to the `Payment Plan` variable.

Barplots

```
ggplot(data = biketown,
       mapping = aes(x = month,
                     fill = PaymentPlan)) +
  geom_bar(position = "fill")
```



- Alternatively, we can consider the y-axis to represent proportion, making direct comparison easier.

New Data Context: `pdxTrees`

- The `pdxTrees` R package contains data on all the trees in the Portland Metro Area.
- Today, we'll look at the Maple, Oak, Pine, Cedar, and Douglas-fir trees in a few parks near Reed.
- Let's load the data

New Data Context: `pdxTrees`

- The `pdxTrees` R package contains data on all the trees in the Portland Metro Area.
- Today, we'll look at the Maple, Oak, Pine, Cedar, and Douglas-fir trees in a few parks near Reed.
- Let's load the data
- Don't worry, we haven't learned the below code yet.

```
library(pdxTrees)
near_Reed <- get_pdxTrees_parks(park = c("Woodstock Park", "Sellwood Riverfront Park", "Kenilworth Park"))
near_Reed <- near_Reed[near_Reed$Genus %in% c("Acer", "Quercus", "Pinus", "Thuja", "Pseudotsuga"), ]
```

Inspect the data

```
1 glimpse(near_Reed)
```

```
Rows: 323
Columns: 34
$ Longitude                 <dbl> -122.6304, -122.6301, -122.6301, -122.6299,…
$ Latitude                  <dbl> 45.49201, 45.49080, 45.49081, 45.49094, 45.…
$ UserID                    <chr> "7670", "7671", "7672", "7902", "7903", "79…
$ Genus                     <chr> "Quercus", "Pseudotsuga", "Pseudotsuga", "Q…
$ Family                    <chr> "Fagaceae", "Pinaceae", "Pinaceae", "Fagace…
$ DBH                       <dbl> 3.3, 43.1, 48.2, 2.4, 11.7, 33.5, 23.5, 37.…
$ Inventory_Date            <dttm> 2018-07-26, 2018-07-26, 2018-07-26, 2018-0…
$ Species                   <chr> "QURU", "PSME", "PSME", "QURU", "PSME", "PS…
$ Common_Name               <chr> "Northern Red Oak", "Douglas-Fir", "Douglas…
$ Condition                 <chr> "Fair", "Fair", "Fair", "Fair", "Good", "Fa…
$ Tree_Height               <dbl> 16, 148, 148, 16, 64, 118, 121, 105, 24, 12…
$ Crown_Width_NS            <dbl> 14, 61, 52, 9, 29, 32, 37, 43, 38, 44, 31, …
```

Inspect the data

```
1 head(near_Reed)
```

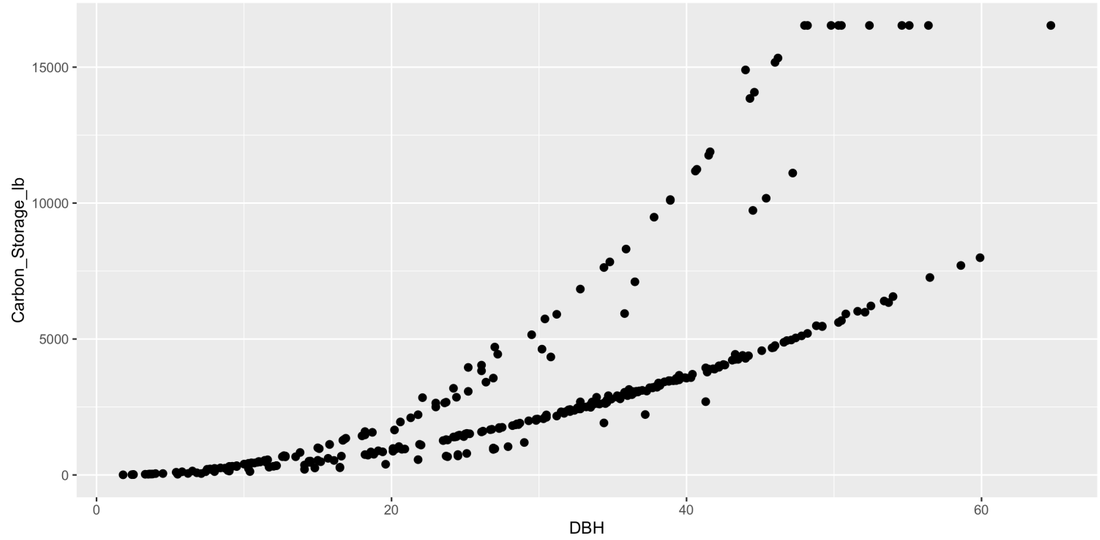
```
# A tibble: 6 × 34
  Longitude Latitude UserID Genus       Family   DBH Inventory_Date      Species
      <dbl>    <dbl> <chr>  <chr>       <chr>  <dbl> <dttm>              <chr>
1     -123.     45.5 7670   Quercus     Fagac…   3.3 2018-07-26 00:00:00 QURU
2     -123.     45.5 7671   Pseudotsuga Pinac…  43.1 2018-07-26 00:00:00 PSME
3     -123.     45.5 7672   Pseudotsuga Pinac…  48.2 2018-07-26 00:00:00 PSME
4     -123.     45.5 7902   Quercus     Fagac…   2.4 2018-07-26 00:00:00 QURU
5     -123.     45.5 7903   Pseudotsuga Pinac…  11.7 2018-07-26 00:00:00 PSME
6     -123.     45.5 7905   Pseudotsuga Pinac…  33.5 2018-07-26 00:00:00 PSME
# ℹ 26 more variables: Common_Name <chr>, Condition <chr>, Tree_Height <dbl>,
#   Crown_Width_NS <dbl>, Crown_Width_EW <dbl>, Crown_Base_Height <dbl>,
#   Collected_By <chr>, Park <chr>, Scientific_Name <chr>,
#   Functional_Type <chr>, Mature_Size <fct>, Native <chr>, Edible <chr>,
#   Nuisance <chr>, Structural_Value <dbl>, Carbon_Storage_lb <dbl>,
```

What does a row represent here?

Scatterplots

- Explore relationships between numerical variables.
  - We will be especially interested in **linear** relationships.

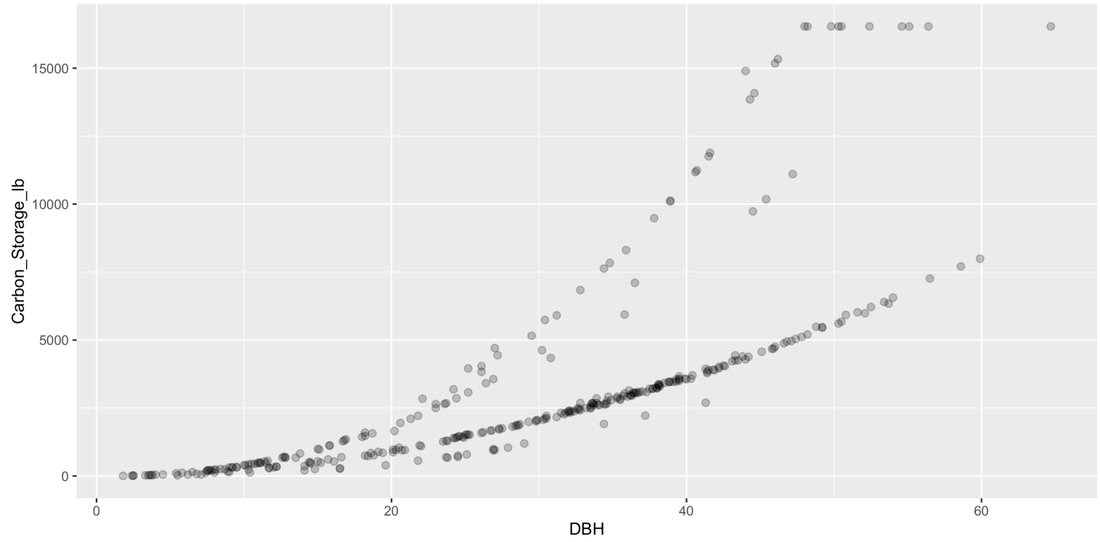
```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb)) +
  geom_point(size = 2)
```



Is there something visually off with the points in this graph?

Scatterplots

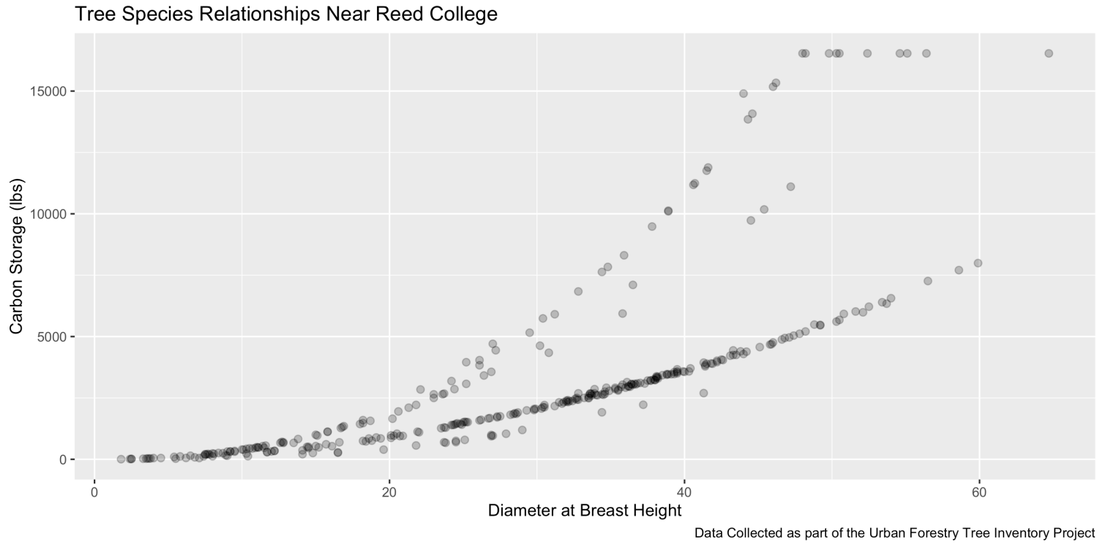
```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb)) +
  geom_point(size = 2, alpha = 0.25)
```



- Fix over-plotting (using **alpha**)
- What's going on in this graph?

Scatterplots

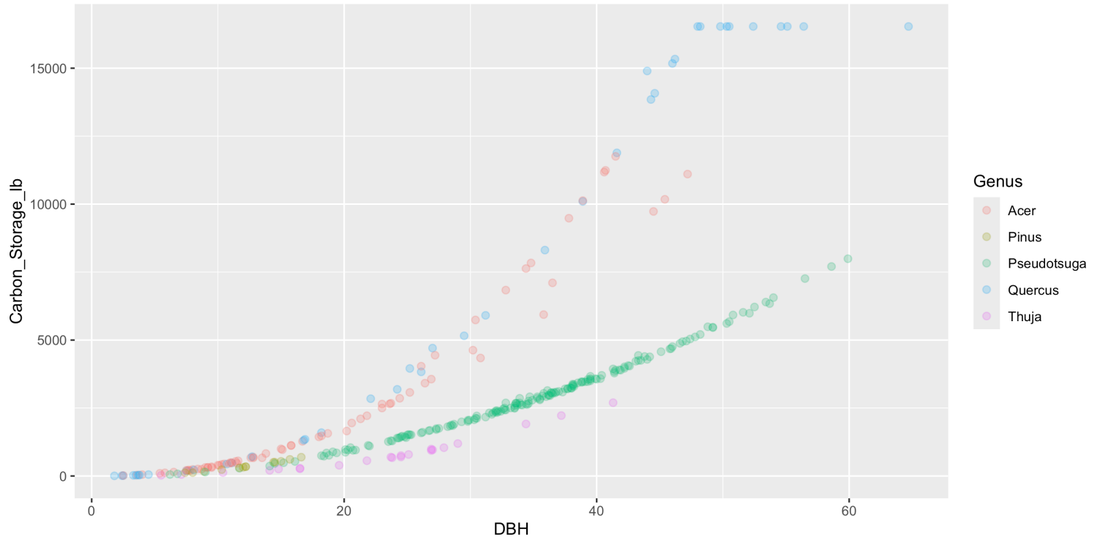
```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb)) +
  geom_point(size = 2, alpha = 0.25) +
  labs(x = "Diameter at Breast Height",
       y = "Carbon Storage (lbs)",
       caption = "Data Collected as part of the Urban F
       title = "Tree Species Relationships Near Reed Co
```



- Fix over-plotting (using **alpha**)
- What's going on in this graph? (labels help add context)

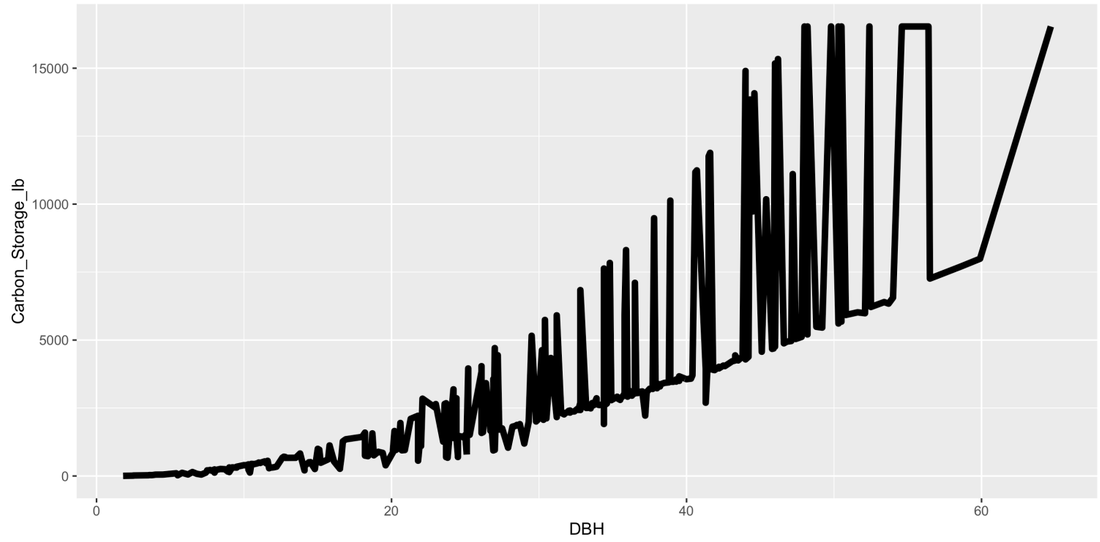
Scatterplots

```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb,
                     color = Genus)) +
  geom_point(size = 2, alpha = 0.25)
```



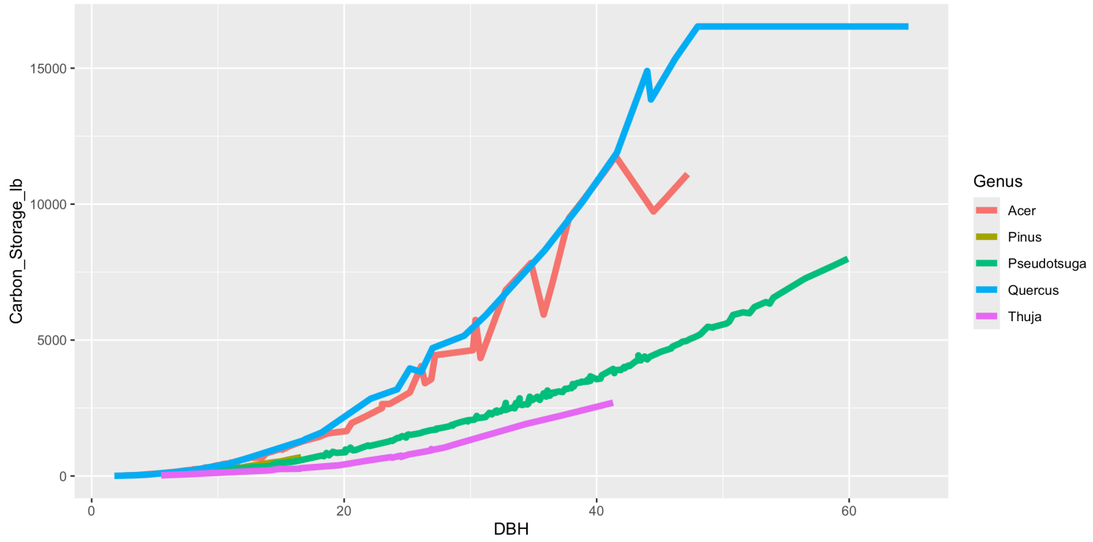
Linegraphs

```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb)) +
  geom_line(linewidth = 2)
```

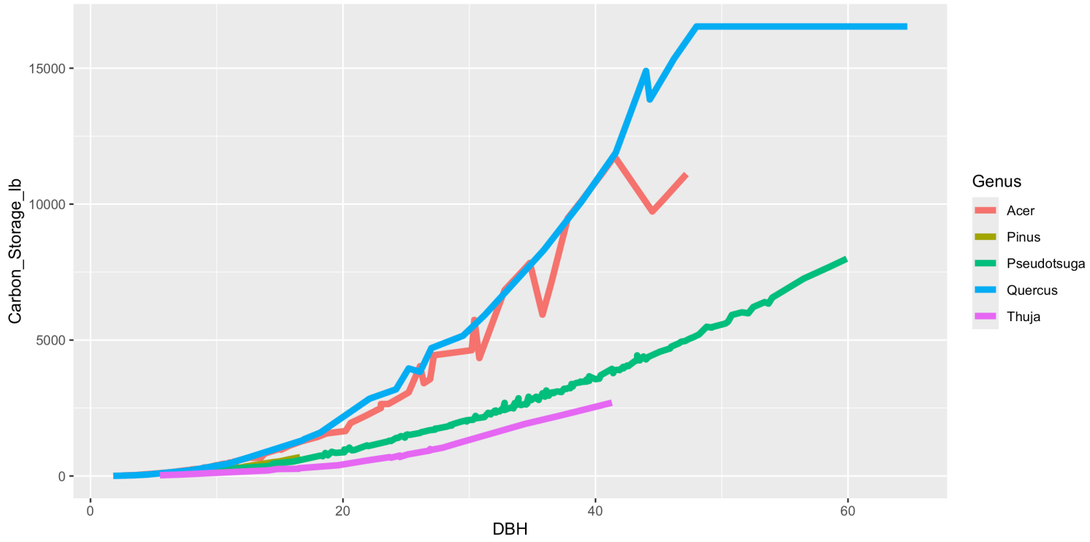


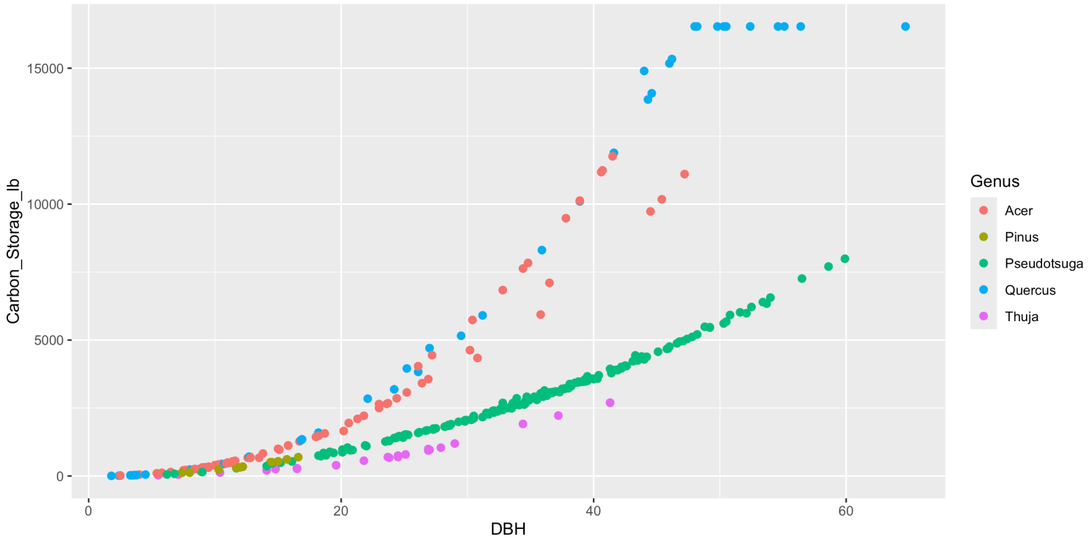
Linegraphs

```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb,
                     color = Genus)) +
  geom_line(linewidth = 2)
```



Linegraphs vs scatterplots





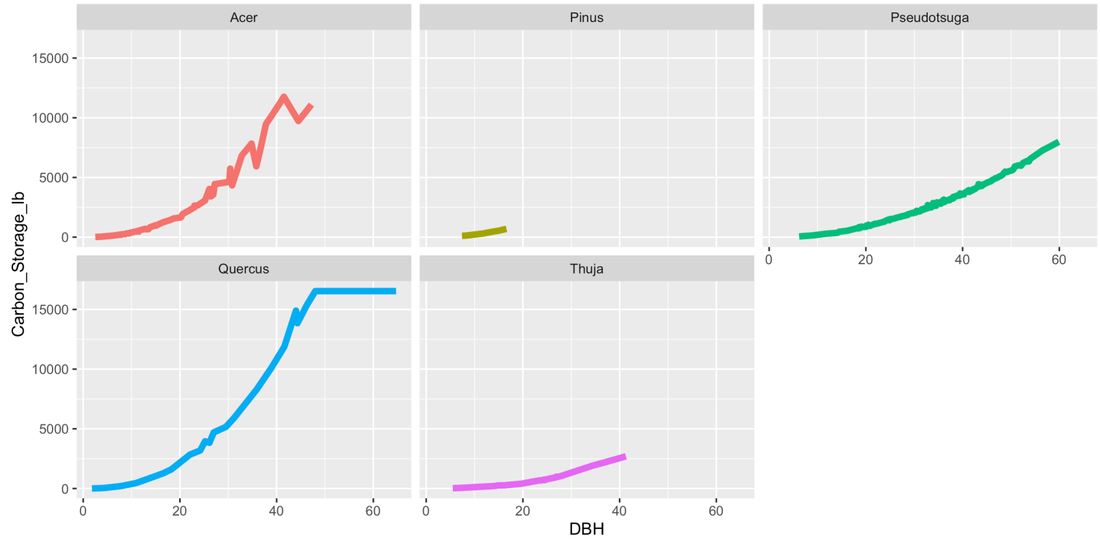
- Which do you prefer?
- Does it depend on context?

A speedy peek at advanced techniques!

If we run out of time, we'll pick this back up during lab next Thursday.

Faceting

```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb,
                     color = Genus)) +
  geom_line(linewidth = 2) +
  facet_wrap(~Genus) +
  guides(color = "none")
```



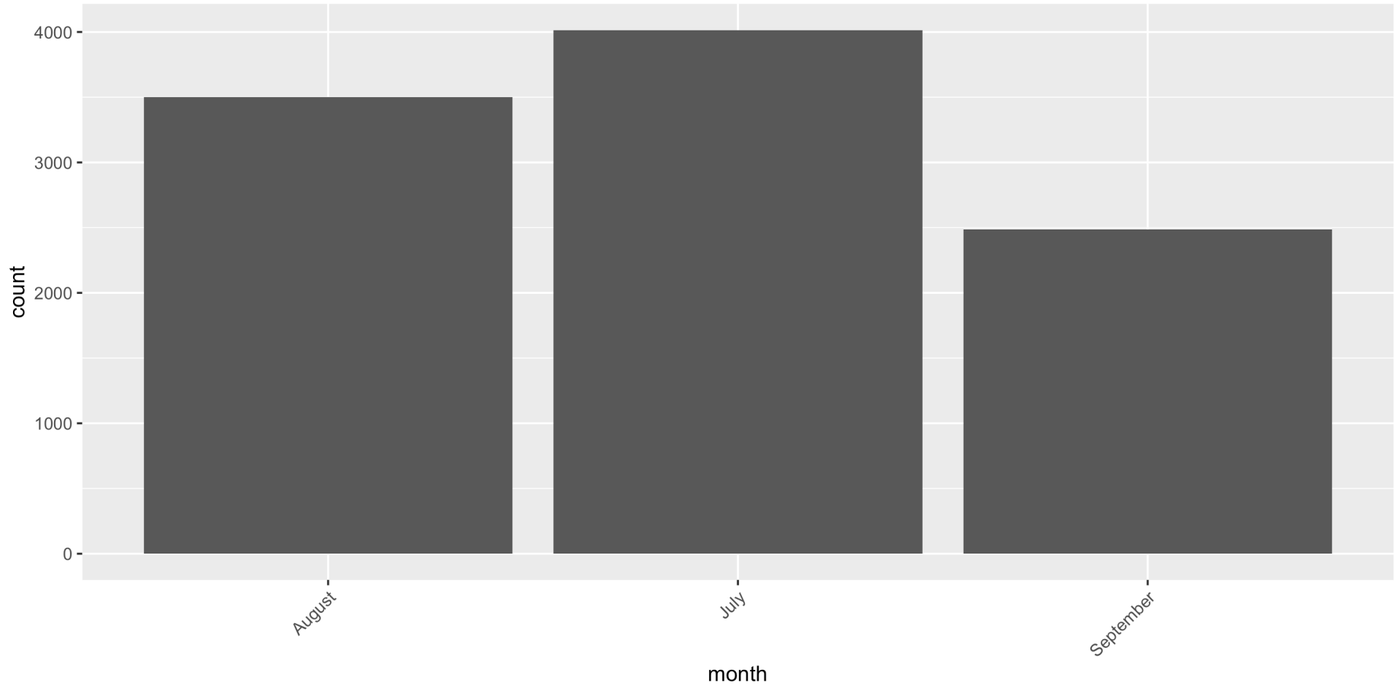
- Faceting is used to split one graphic into several smaller ones, based on the values of a categorical variable

Customizing your `ggplot2` Plots

- There are so **many** ways you can customize the look of your `ggplot2` plots
- Let's look quickly at some common changes:
  - Fussing with labels
  - Zooming in
  - Using multiple `geoms`
  - Color!
  - Themes

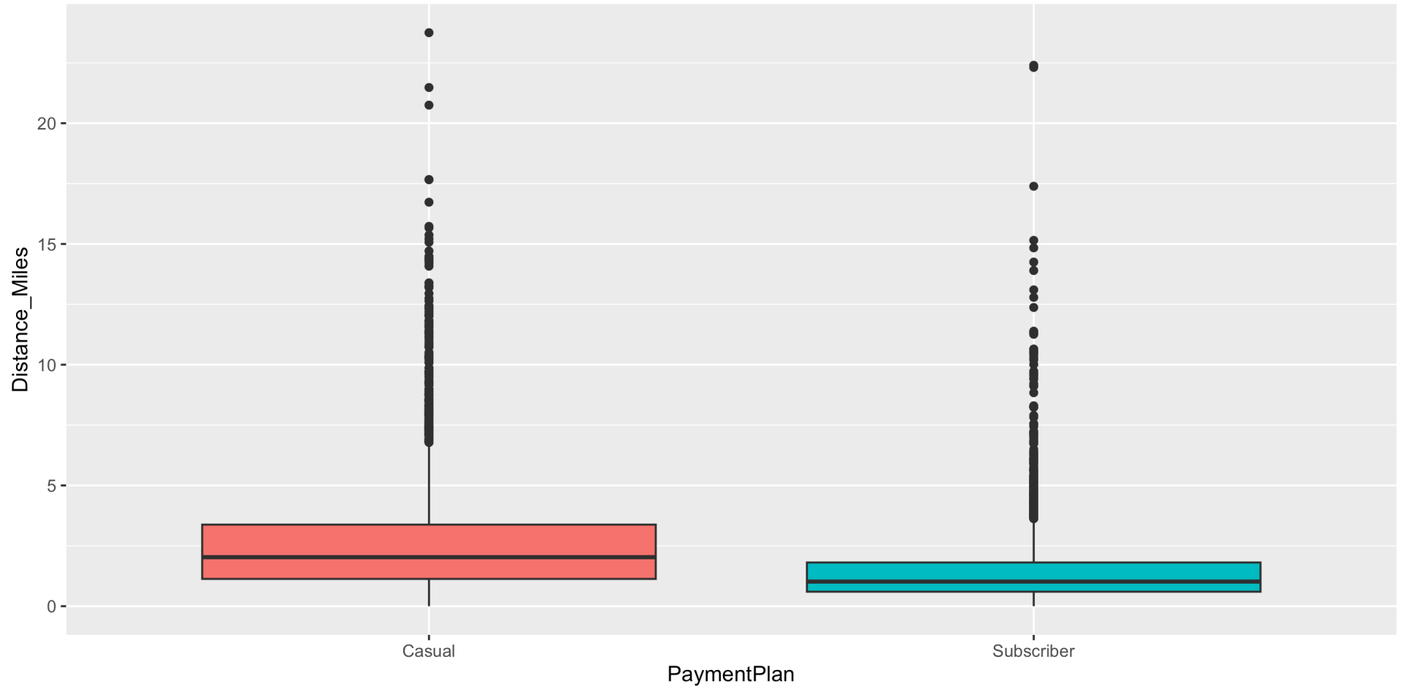
Fussing with Labels: Rotate

```
biketown$month <- ifelse(biketown$month == 7, "July",
                         ifelse(biketown$month == 8, "August", "September"))
ggplot(data = biketown,
       mapping = aes(x = month)) +
  geom_bar() +
    theme(axis.text.x =
          element_text(angle = 45,
                       vjust = 1,
                       hjust = 1))
```



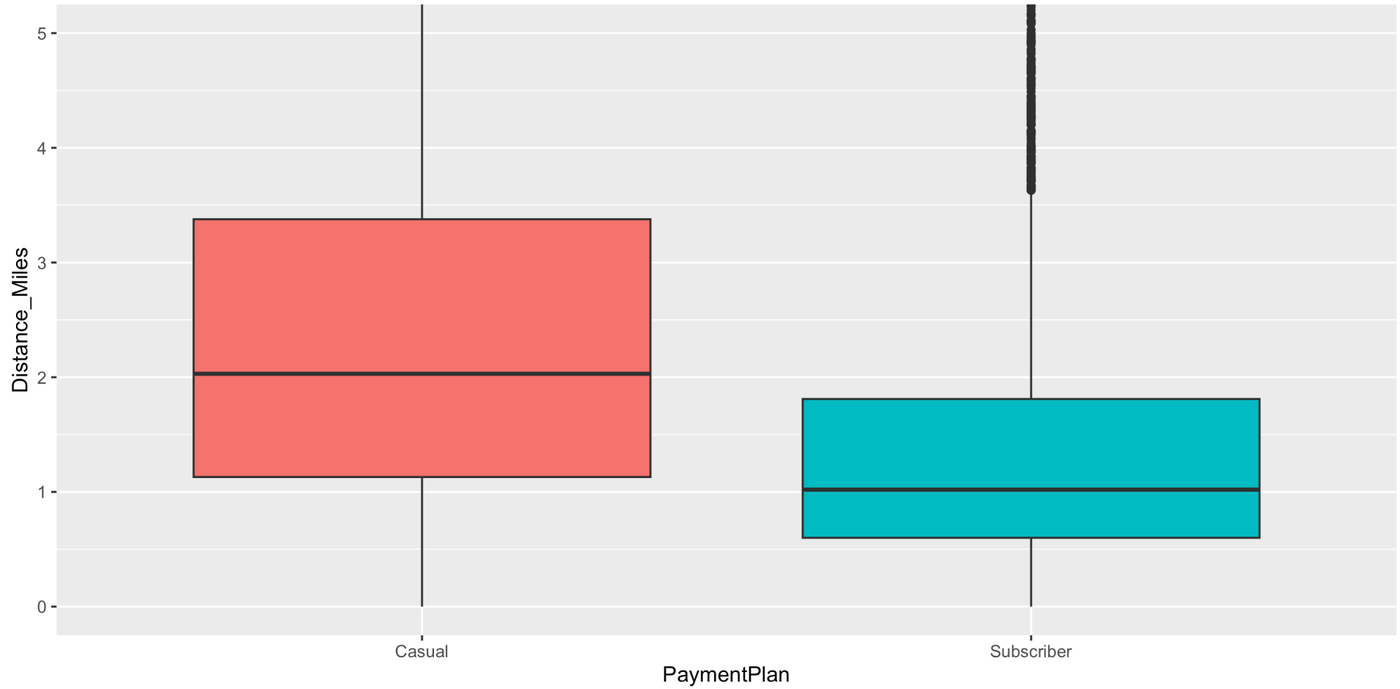
Zooming In

```
ggplot(data = biketown,
       mapping = aes(x = PaymentPlan,
                     y = Distance_Miles,
                     fill = PaymentPlan)) +
  geom_boxplot() +
  guides(fill = "none")
```



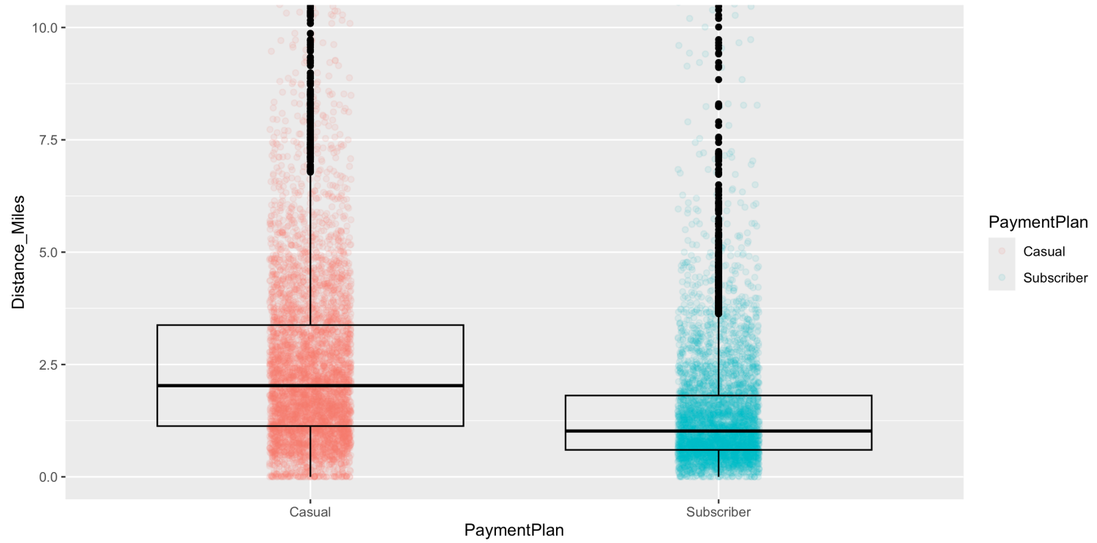
Zooming In

```
ggplot(data = biketown,
       mapping = aes(x = PaymentPlan,
                     y = Distance_Miles,
                     fill = PaymentPlan)) +
  geom_boxplot() +
  guides(fill = "none") +
  coord_cartesian(ylim = c(0, 5))
```



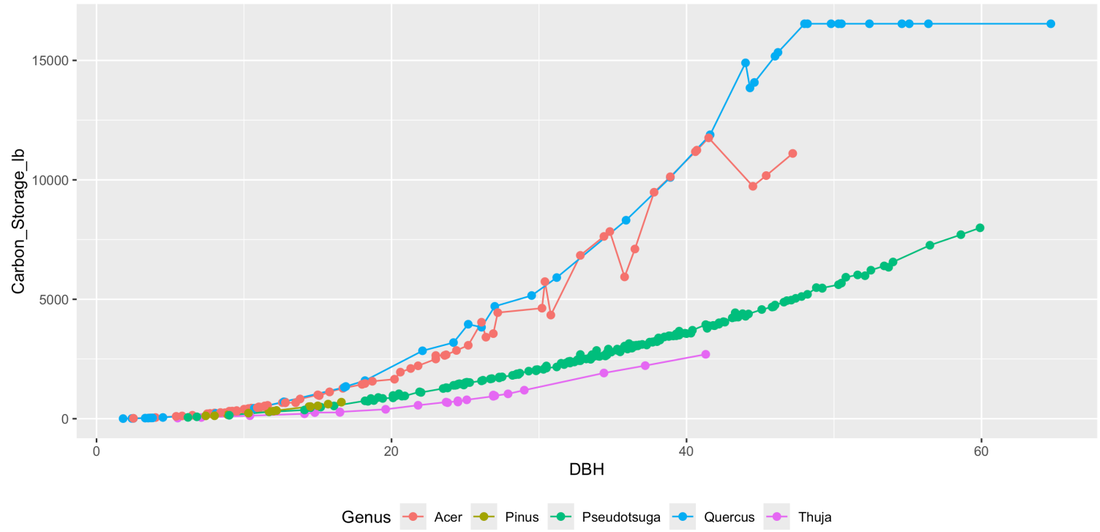
Multiple geoms

```
ggplot(data = biketown,
       mapping = aes(x = PaymentPlan,
                     y = Distance_Miles,
                     color = PaymentPlan)) +
  guides(fill = "none") +
  coord_cartesian(ylim = c(0, 10)) +
  geom_jitter(width = 0.1,
              height = 0,
              alpha = 0.1) +
    geom_boxplot(fill = NA, color = "black")
```



Multiple geoms

```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb,
                     color = Genus)) +
  geom_line() +
  theme(legend.position = "bottom") +
  geom_point(size = 2)
```



Change the Color

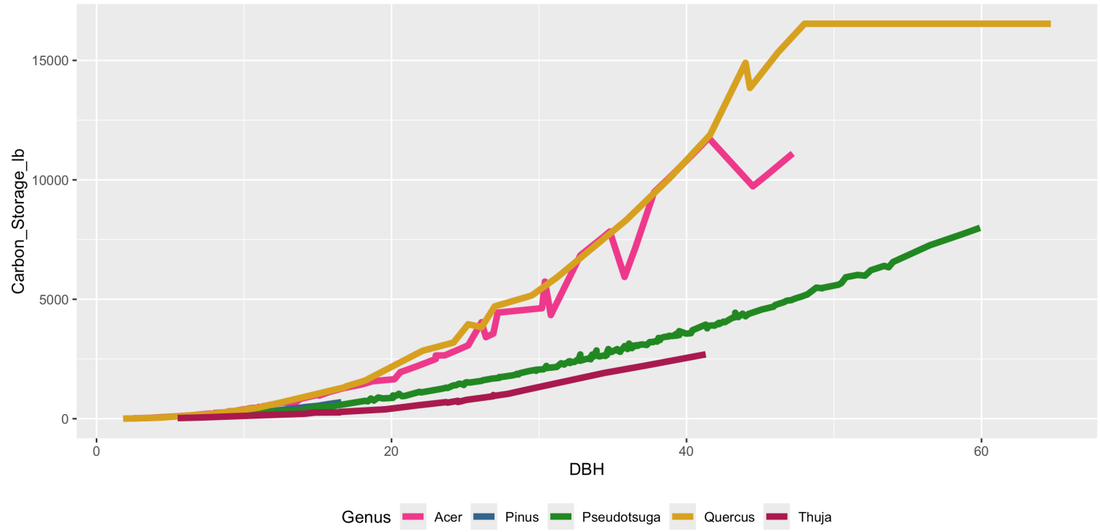
```
colors()
```

```
  [1] "white"                "aliceblue"            "antiquewhite"
  [4] "antiquewhite1"        "antiquewhite2"        "antiquewhite3"
  [7] "antiquewhite4"        "aquamarine"           "aquamarine1"
 [10] "aquamarine2"          "aquamarine3"          "aquamarine4"
 [13] "azure"                "azure1"               "azure2"
 [16] "azure3"               "azure4"               "beige"
 [19] "bisque"               "bisque1"              "bisque2"
 [22] "bisque3"              "bisque4"              "black"
 [25] "blanchedalmond"       "blue"                 "blue1"
 [28] "blue2"                "blue3"                "blue4"
 [31] "blueviolet"           "brown"                "brown1"
 [34] "brown2"               "brown3"               "brown4"
 [37] "burlywood"            "burlywood1"           "burlywood2"
 [40] "burlywood3"           "burlywood4"           "cadetblue"
```

You can also use hex color codes to fully customize colors.

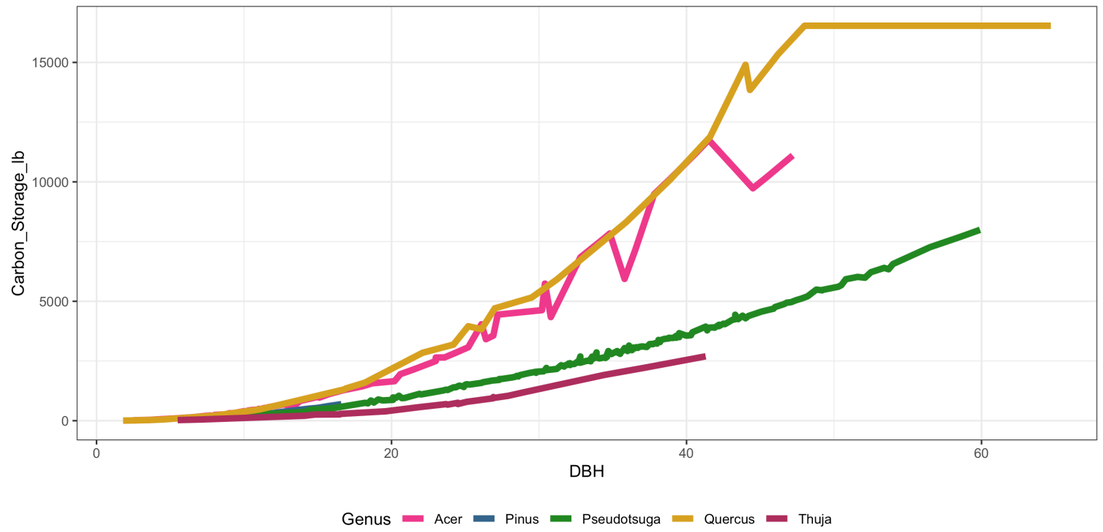
Change the Color

```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb,
                     color = Genus)) +
  geom_line(size = 2) +
  theme(legend.position = "bottom") +
  scale_color_manual(values = c("violetred2",
                                "steelblue4",
                                "forestgreen",
                                "goldenrod",
                                "#aa0951"))
```



Use a Different Theme

```
ggplot(data = near_Reed,
       mapping = aes(x = DBH,
                     y = Carbon_Storage_lb,
                     color = Genus)) +
  geom_line(size = 2) +
  scale_color_manual(values = c("violetred2",
                                "steelblue4",
                                "forestgreen",
                                "goldenrod",
                                "maroon")) +
  theme_bw() +
  theme(legend.position = "bottom")
```



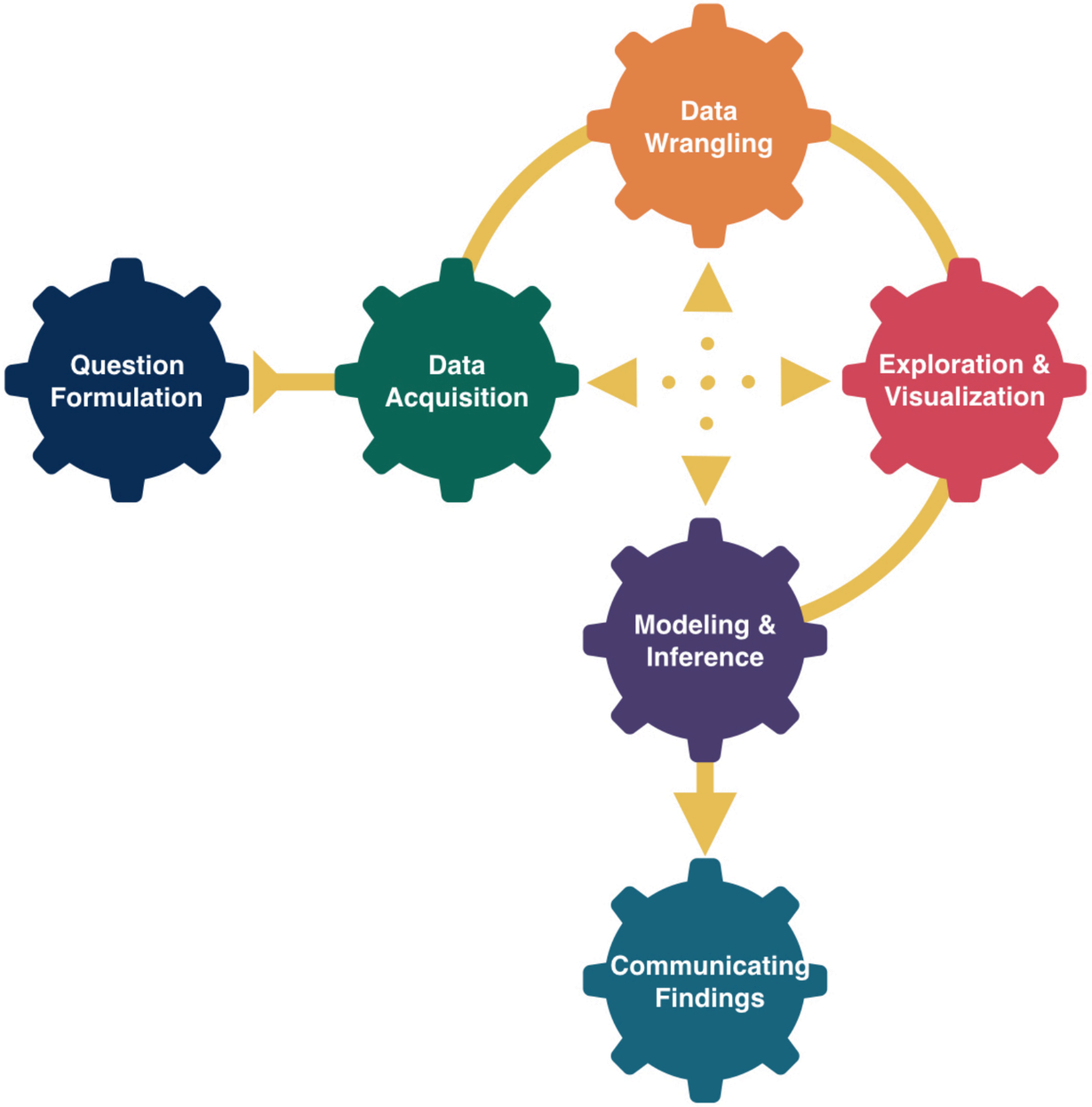
Recap: ggplot2

```
library(tidyverse)
ggplot(data = ---, mapping = aes(---)) +
  geom_---(---)
```

- This is the major concept to remember
- We covered a lot of ground - it takes lots of practice to become fluent in the details!

Next time:

- Data wrangling!



Summary Statistics

Megan Ayers

Math 141 | Spring 2026

Monday, Week 2

Announcements

- The teaching team would love to see you in office hours!
  - **Megan's office hours**: For individual or small group help on problems or concepts
  - **Course assistant office hours**: To work on assignments with your peers and get help from the course assistants when you are stuck
- Reminder: If you haven't already, make sure your access to Gradescope is set up.
- Please select pages for each problem part on Gradescope

Last Time

- Learned about the structure of `ggplot2`
- Learn five standard graphs for numerical/quantitative data: histograms, boxplots, barplots, scatterplots, and linegraphs

Goals for Today

- Consider measures for **summarizing** quantitative data
  - Center
  - Spread/variability
- Consider measures for **summarizing** categorical data
  - Contingency tables

Import the Data

```r
biketown <- read.csv("data/biketown.csv")

# Inspect the data
str(biketown)
```

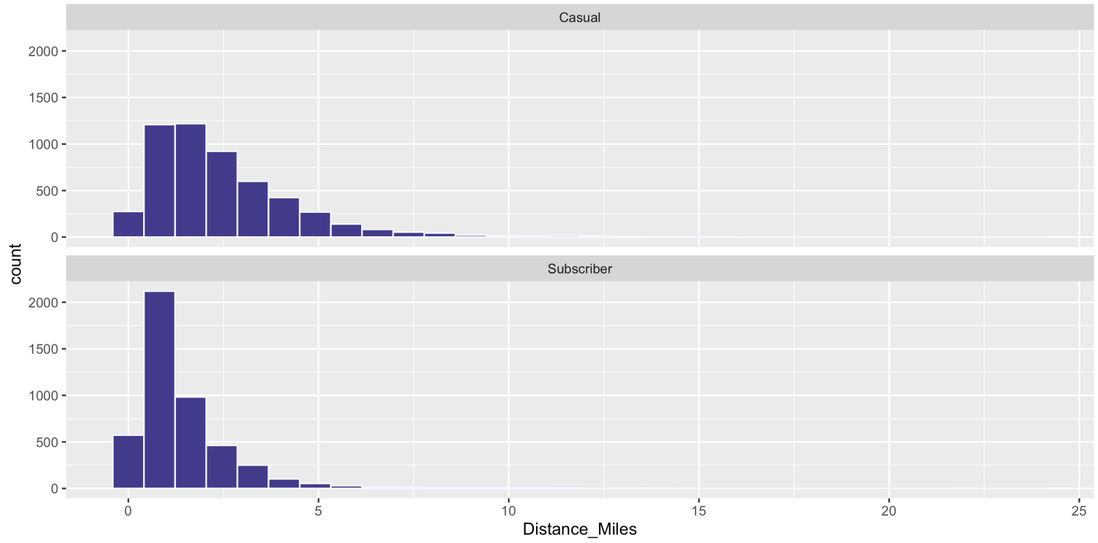
```
'data.frame':   9999 obs. of  19 variables:
 $ RouteID         : int  4074085 3719219 3789757 3576798 3459987 3947695 3549550 4411957 4098004 4096862 ...
 $ PaymentPlan     : chr  "Subscriber" "Casual" "Casual" "Subscriber" ...
 $ StartHub        : chr  "SE Elliott at Division" "SW Yamhill at Director Park" "NE Holladay at MLK" "NW Couch at 2nd"
...
 $ StartLatitude   : num  45.5 45.5 45.5 45.5 45.5 ...
 $ StartLongitude  : num  -123 -123 -123 -123 -123 ...
 $ StartDate       : chr  "8/17/2017" "7/22/2017" "7/27/2017" "7/12/2017" ...
 $ StartTime       : chr  "10:44:00" "14:49:00" "14:13:00" "13:23:00" ...
 $ EndHub          : chr  "Blues Fest - SW Waterfront at Clay - Disabled" "SW 2nd at Pine" "NE Alberta at NE 29th/30th -
Community Corral" "NW Raleigh at 21st" ...
 $ EndLatitude     : num  45.5 45.5 45.6 45.5 45.5 ...
 $ EndLongitude    : num  -123 -123 -123 -123 -123 ...
 $ EndDate         : chr  "8/17/2017" "7/22/2017" "7/27/2017" "7/12/2017" ...
```

Summarizing Data

	RouteID	PaymentPlan	StartHub	Distance_Miles
25	3596434	Subscriber	NW 18th at Flanders	0.59
26	3607170	Subscriber	NW Raleigh at 21st	0.71
27	3631639	Casual	SW 2nd at Pine	3.15
28	3912181	Casual	SE Water at Taylor	5.89
29	4031739	Casual	SE Clay at Water	1.34
30	3859969	Casual	SW Naito at Morrison	1.56
31	4315016	Casual	NW Everett at 22nd	3.50
32	4252609	Casual	NW Flanders at 14th	1.81
33	3809564	Casual	NA	2.74

- Hard to do by eyeballing a spreadsheet with many rows!

Summarizing Data Visually



For a quantitative variable, often want to answer:

- What is an **average** value?
- What is the **trend/shape** of the variable?
- How much **variation** is there from case to case?

Need to learn key **summary statistics**: Numerical values computed based on the observed cases.

Measures of Center

Mean: Average of all the observations

- n = Number of cases (sample size)
- x_i = value of the i-th observation
- Denote by $\bar{x}$

Measures of Center

Mean: Average of all the observations

- n = Number of cases (sample size)
- x_i = value of the i-th observation
- Denote by $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Measures of Center

Mean: Average of all the observations

- n = Number of cases (sample size)
- x_i = value of the i-th observation
- Denote by $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```r
# Test out on first 6 values
head(biketown$Distance_Miles)
```

```
[1] 1.91 0.72 3.42 1.81 4.51 5.54
```

```r
(1.91 + 0.72 + 3.42 + 1.81 + 4.51 + 5.54) / 6
```

```
[1] 2.985
```

```r
# Compute in R for all values
mean(biketown$Distance_Miles)
```

```
[1] 2.044768
```

Measures of Center

Median: Middle value

- Half of the data falls below the median
- Denote by m
- If n is even, then it is the average of the middle two values

Measures of Center

Median: Middle value

- Half of the data falls below the median
- Denote by m
- If n is even, then it is the average of the middle two values

```
# Test out on first 6 values
head(biketown$Distance_Miles)
```

```
[1] 1.91 0.72 3.42 1.81 4.51 5.54
```

```
sort(head(biketown$Distance_Miles))
```

```
[1] 0.72 1.81 1.91 3.42 4.51 5.54
```

```
(3.42 + 1.91) / 2
```

```
[1] 2.665
```

Compute in R:

```
median(biketown$Distance_Miles)
```

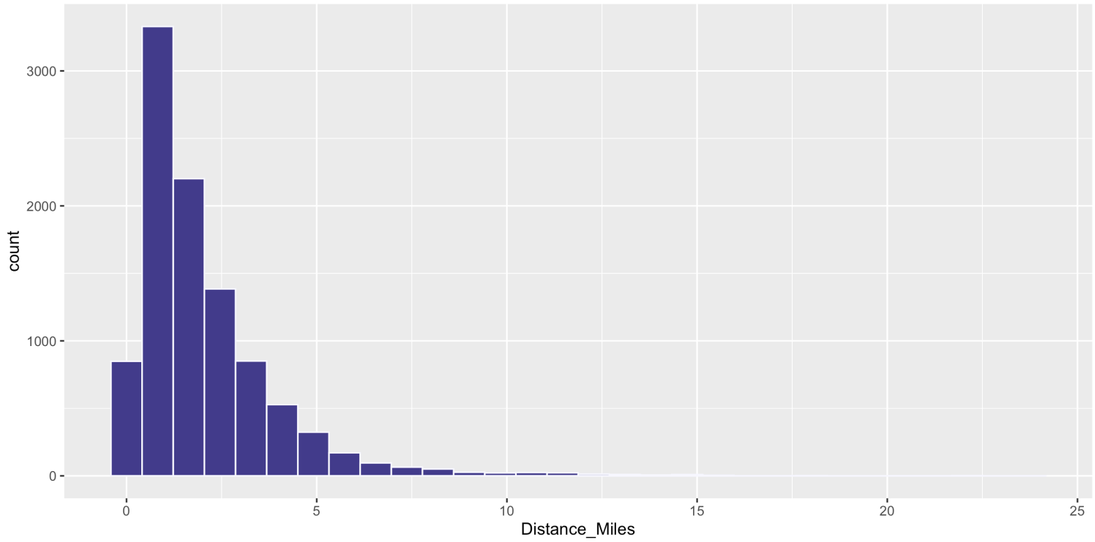
```
[1] 1.48
```

Measures of Center

Q: Why is the mean larger than the median?

```
1 data.frame(mean_miles = mean(biketown$Distance_Miles),
2            median_miles = median(biketown$Distance_Mile
```

```
  mean_miles median_miles
1   2.044768         1.48
```



Answer: the distribution is **skewed**. In skewed distributions, the mean is pulled farther in the direction of skew than the median.

Measures of Center

Note: The mean is very sensitive to outliers, while the median is not.

```r
my_data <- c(1, 2, 5, 7, 8, 10)
my_data_with_outlier <- c(1, 2, 5, 7, 8, 100)
```

```r
mean(my_data)
```

```
[1] 5.5
```

```r
mean(my_data_with_outlier)
```

```
[1] 20.5
```

```r
median(my_data)
```

```
[1] 6
```

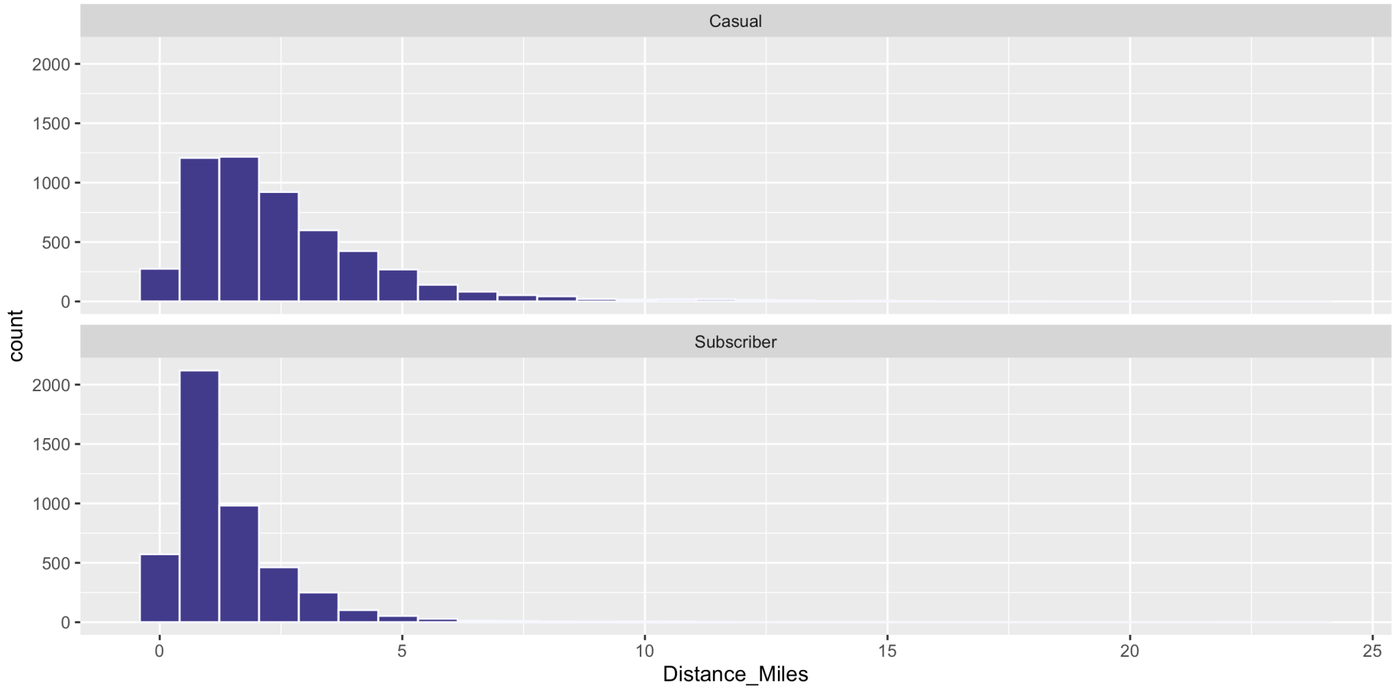
```r
median(my_data_with_outlier)
```

```
[1] 6
```

We call the median a **robust** statistic.

Computing Measures of Center by Groups

Question: Who travels further, on average? Casual biketown users or payment plan subscribers?

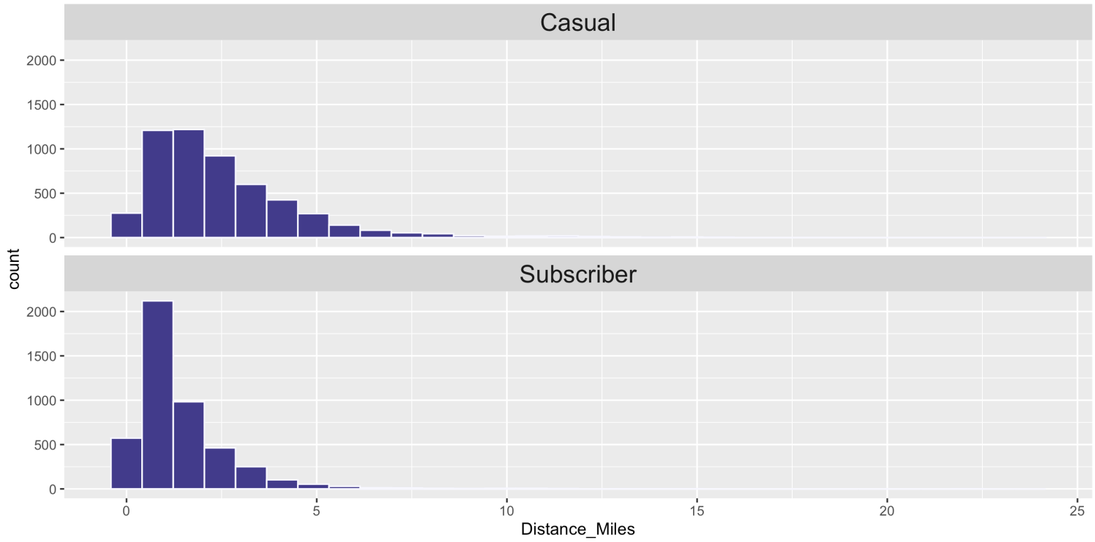


Computing Measures of Center by Groups

Handy `dplyr` functions that we'll learn: `group_by()` and `summarize()`.

```
 1 library(dplyr)
 2
 3 # Calculate group *MEANS*
 4 biketown %>%
 5   group_by(PaymentPlan) %>%
 6   summarize(mean_dist = mean(Distance_Miles))
```

```
# A tibble: 2 × 2
  PaymentPlan mean_dist
  <chr>           <dbl>
1 Casual           2.56
2 Subscriber       1.45
```




```
 1 # Calculate group *MEDIANS*
 2 biketown %>%
 3   group_by(PaymentPlan) %>%
 4   summarize(median_dist = median(Distance_Miles))
```

```
# A tibble: 2 × 2
  PaymentPlan median_dist
  <chr>             <dbl>
1 Casual             2.03
2 Subscriber         1.02
```

Measures of Variability

- Want a statistic that captures how much observations **deviate** from the mean

- Find how much each observation deviates from the mean.
- Idea: Compute the average of the deviations

Measures of Variability

- Want a statistic that captures how much observations **deviate** from the mean

- Find how much each observation deviates from the mean.
- Idea: Compute the average of the deviations

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$$

Measures of Variability

- Want a statistic that captures how much observations **deviate** from the mean

- Find how much each observation deviates from the mean.
- Idea: Compute the average of the deviations

```
# Test out on first 6 values
head(biketown$Distance_Miles)
```

```
[1] 1.91 0.72 3.42 1.81 4.51 5.54
```

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$$

Problem?

```
# Calculate mean by hand
(1.91 + 0.72 + 3.42 + 1.81 + 4.51 + 5.54) / 6
```

```
[1] 2.985
```

```
# Calculate average deviations by hand
((1.91 - 2.985) + (0.72 - 2.985) +
  (3.42 - 2.985) + (1.81 - 2.985) +
  (4.51 - 2.985) + (5.54 - 2.985)) / 6
```

```
[1] 0
```

Measures of Variability

- Want a statistic that captures how much observations **deviate** from the mean

NEW proposal:

- Find how much each observation deviates from the mean.
- Compute the average of the **squared** deviations.

Measures of Variability

- Want a statistic that captures how much observations **deviate** from the mean

NEW proposal:

- Find how much each observation deviates from the mean.
- Compute the average of the **squared** deviations.

```
# Test out on first 6 values
head(biketown$Distance_Miles)
```

```
[1] 1.91 0.72 3.42 1.81 4.51 5.54
```

```
# Calculate mean by hand
(1.91 + 0.72 + 3.42 + 1.81 + 4.51 + 5.54) / 6
```

```
[1] 2.985
```

```
# Calculate average squared deviations by hand
((1.91 - 2.985)^2 + (0.72 - 2.985)^2 +
  (3.42 - 2.985)^2 + (1.81 - 2.985)^2 +
  (4.51 - 2.985)^2 + (5.54 - 2.985)^2) / 6
```

```
[1] 2.784892
```

Measures of Variability

- Want a statistic that captures how much observations **deviate** from the mean

NEW proposal, formula:

- Find how much each observation deviates from the mean.
- Compute the (nearly) average of the **squared** deviations.
- Called **sample variance** s^2.

Measures of Variability

Measures of Variability

Measures of Variability

- Want a statistic that captures how much observations **deviate** from the mean

- Find how much each observation deviates from the mean.
- Compute the (nearly) average of the **squared** deviations (s^2).
- Because observations are squared, units differ from original data.

- The square root of the sample variance is called the **sample standard deviation** s.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Compute in R:

```r
var(biketown$Distance_Miles)  # Variance
```

```
[1] 3.805638
```

```r
sd(biketown$Distance_Miles)   # Standard deviation
```

```
[1] 1.950804
```

```r
sqrt(var(biketown$Distance_Miles))
```

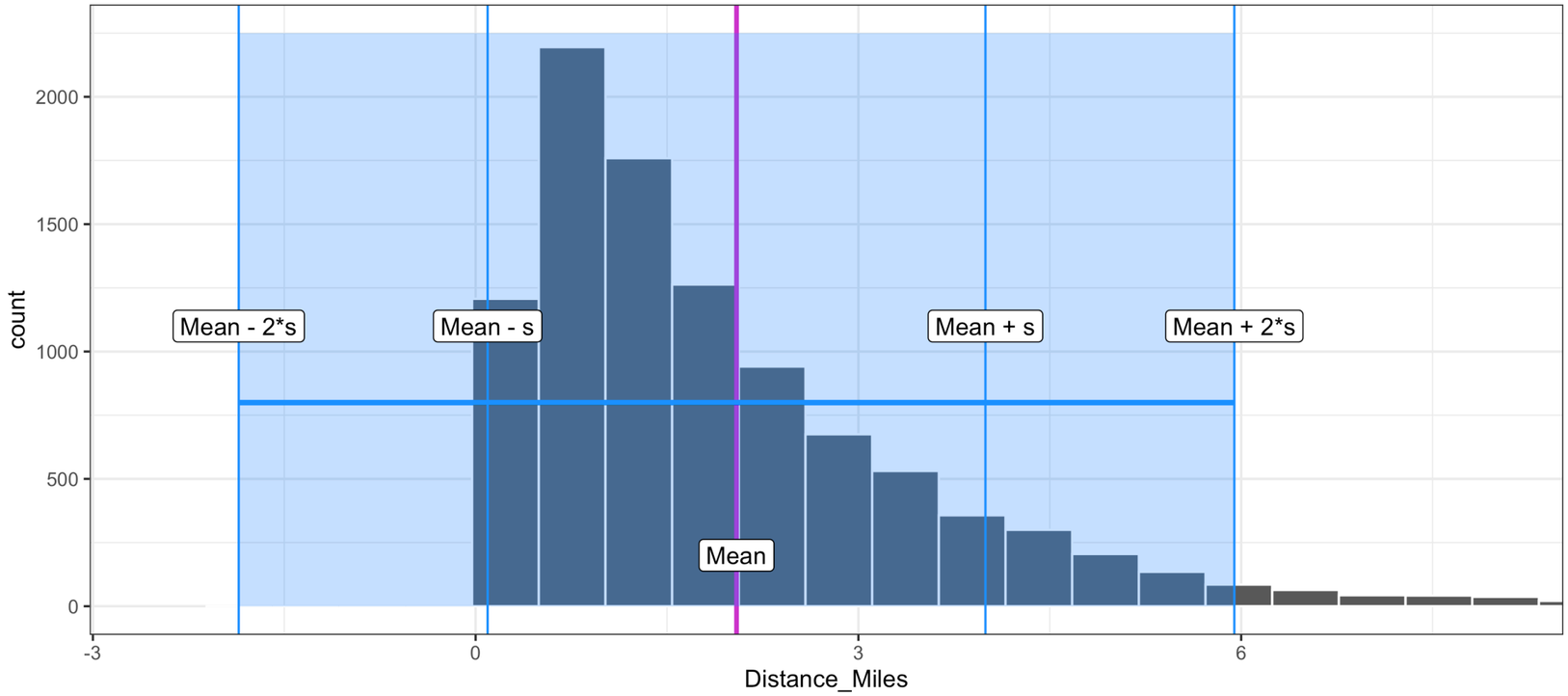
```
[1] 1.950804
```

Visualizing Standard Deviation

- The standard deviation measures the **typical size of deviations from the mean.**
- For most data sets, the large majority of observations are within 2 standard deviations of the mean.

```
1 sd(biketown$Distance_Miles)
```

```
[1] 1.950804
```



Measures of Variability

- In addition to the sample standard deviation and the sample variance, there is the sample **interquartile range** (IQR):

$$\text{IQR} = Q_3 - Q_1$$

- 25% of all observations are less than the *first quartile* Q1
- 25% of all observations are greater than the *third quartile* Q3

Compute with R:

```
1 IQR(biketown$Distance_Miles)
```

```
[1] 1.89
```

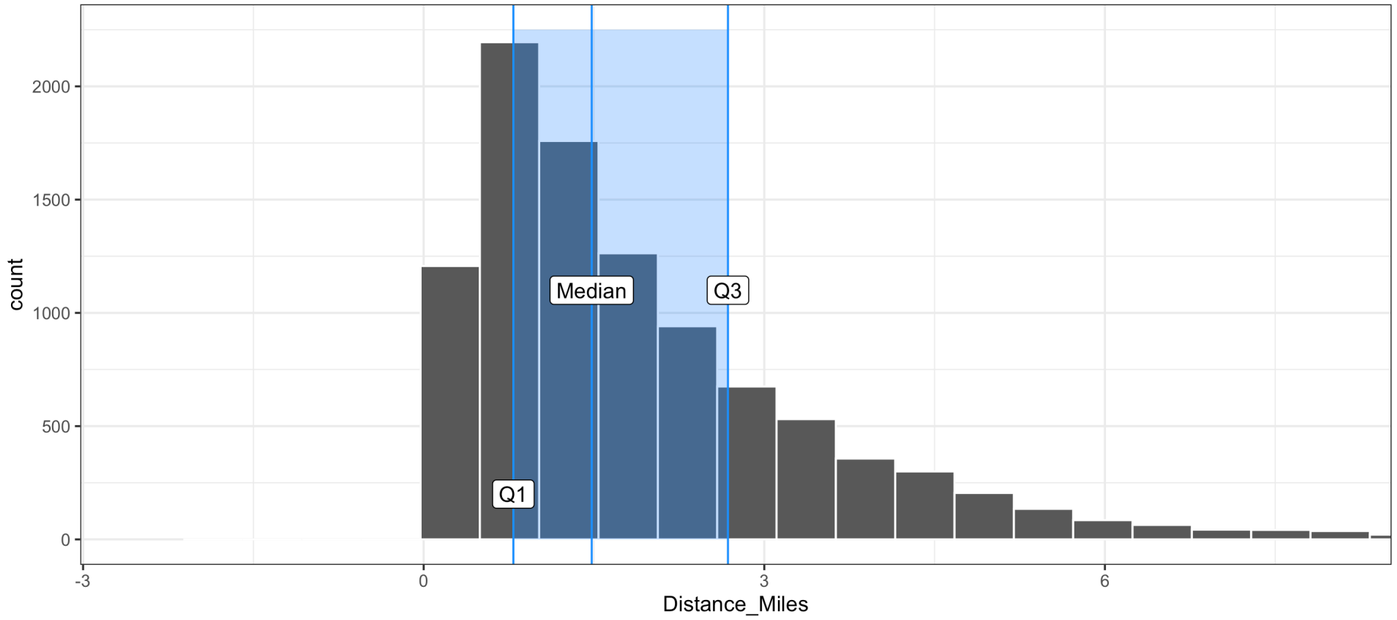
Visualizing the IQR

- In addition to the sample standard deviation and the sample variance, there is the sample **interquartile range** (IQR):

$$\text{IQR} = Q_3 - Q_1$$

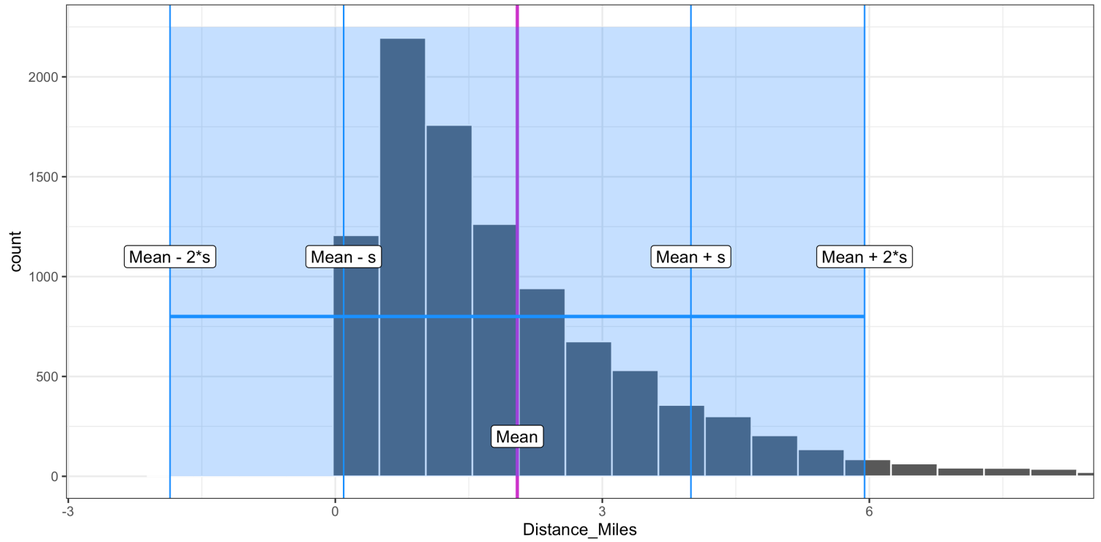
```
1 quantile(biketown$Distance_Miles, c(0.25, 0.5, 0.75))
```

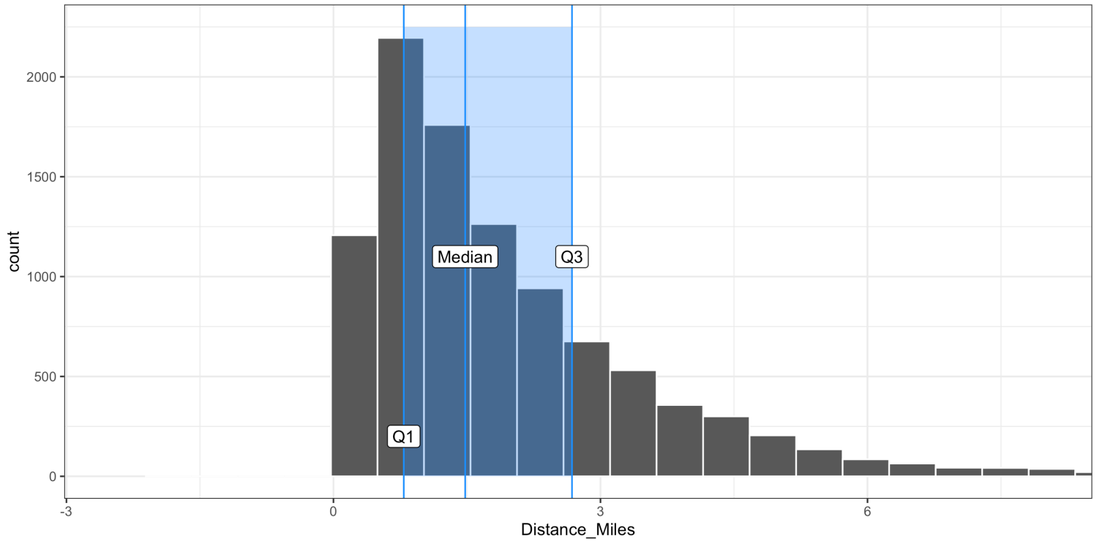
```
 25%  50%  75% 
0.79 1.48 2.68 
```


2000
1500
1000
500
0
count
Median
Q3
Q1
-3
0
3
6
Distance_Miles

Comparing Measures of Variability

- **Q: Which is more robust to outliers, the IQR or s?**





- **Q: Which is more commonly used, the IQR or s?**

Summarizing Categorical Variables

New data set: Cambridge dogs

```r
dogs <- read.csv("https://data.cambridgema.gov/api/views/sckh-3xyx/rows.csv")
str(dogs)
```

```
'data.frame':   2794 obs. of  6 variables:
 $ Dog_Name        : chr  "Grace" "Butch" "RUDI 3" "Phoebe" ...
 $ Dog_Breed       : chr  "Mixed Breed" "Mixed Breed" "Shih Tzu" "Labradoodle" ...
 $ Location_masked : logi  NA NA NA NA NA NA ...
 $ Latitude_masked : num  42.4 42.4 42.4 42.4 42.4 ...
 $ Longitude_masked: num  -71.1 -71.1 -71.1 -71.1 -71.1 ...
 $ Neighborhood    : chr  "Baldwin" "North Cambridge" "Riverside" "West Cambridge" ...
```

Cambridge dogs data set

May want to focus on the dogs with the 5 most common names

```r
dogs <- read.csv("https://data.cambridgema.gov/api/views/sckh-3xyx/rows.csv")

# Useful wrangling that we will come back to
dogs_top5 <- dogs %>%
  mutate(Breed = case_when(Dog_Breed == "Mixed Breed" ~ "Mixed",
                           TRUE ~ "Single")) %>%
  filter(Dog_Name %in% c("Luna", "Charlie", "Lucy", "Cooper", "Rosie"))


head(dogs_top5)
```

```
  Dog_Name          Dog_Breed Location_masked Latitude_masked Longitude_masked
1     Luna        Mixed Breed              NA         42.3939         -71.1311
2   Cooper      Cairn Terrier              NA         42.3785         -71.1237
3  Charlie        Mixed Breed              NA         42.3689         -71.1076
4   Cooper Labrador Retriever              NA         42.3788         -71.1396
5    Rosie       Goldendoodle              NA         42.3827         -71.1424
6    Rosie      DACHSHUND MIX              NA         42.3626         -71.1140
       Neighborhood  Breed
1   North Cambridge  Mixed
2 Neighborhood Nine Single
3     Mid-Cambridge  Mixed
4    West Cambridge Single
5    West Cambridge Single
6         Riverside Single
```

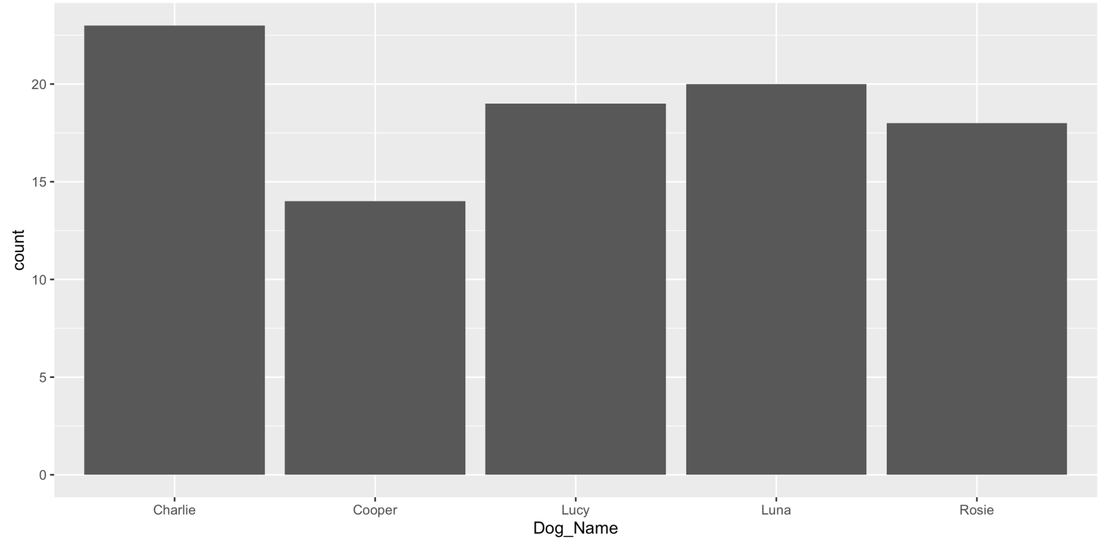
Frequency Table

Distributions of categorical variables can be presented in tables and summarized in bar charts.

```
1 count(dogs_top5, Dog_Name)  # Handy function
```

```
  Dog_Name  n
1  Charlie 23
2   Cooper 14
3     Lucy 19
4     Luna 20
5    Rosie 18
```

```
1 ggplot(data = dogs_top5,
2     mapping = aes(x = Dog_Name)) +
3   geom_bar()
```



Q: Why can't we use mean or standard deviation here?

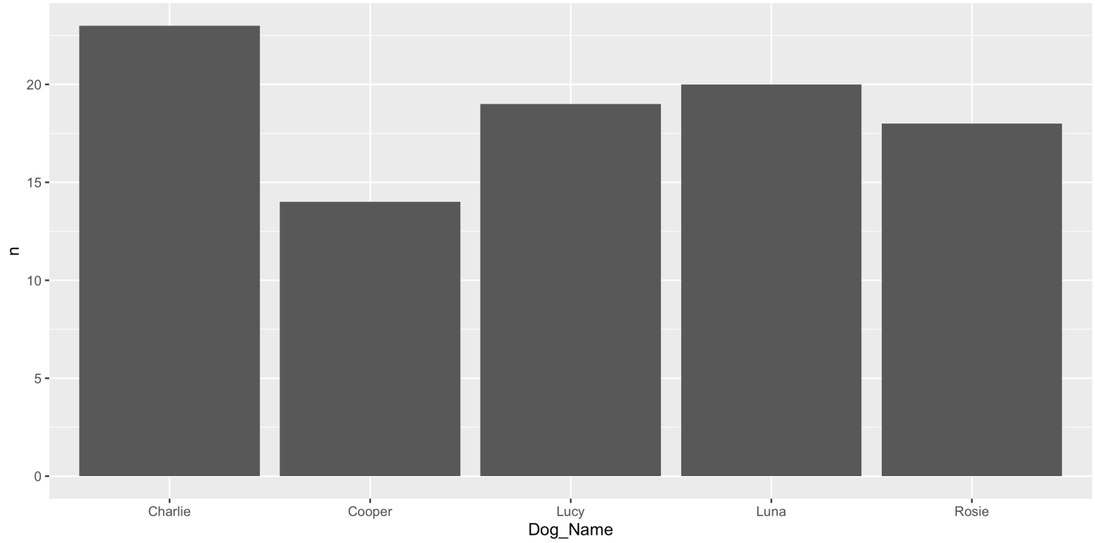
Another `ggplot2` `geom`: `geom_col()`

If you have already aggregated the data, you will use `geom_col()` instead of `geom_bar()`.

```
dog_counts <- count(dogs_top5, Dog_Name)
dog_counts
```

```
  Dog_Name  n
1  Charlie 23
2   Cooper 14
3     Lucy 19
4     Luna 20
5    Rosie 18
```

```
ggplot(data = dog_counts,
       mapping = aes(x = Dog_Name,
                     y = n)) +
  geom_col()
```



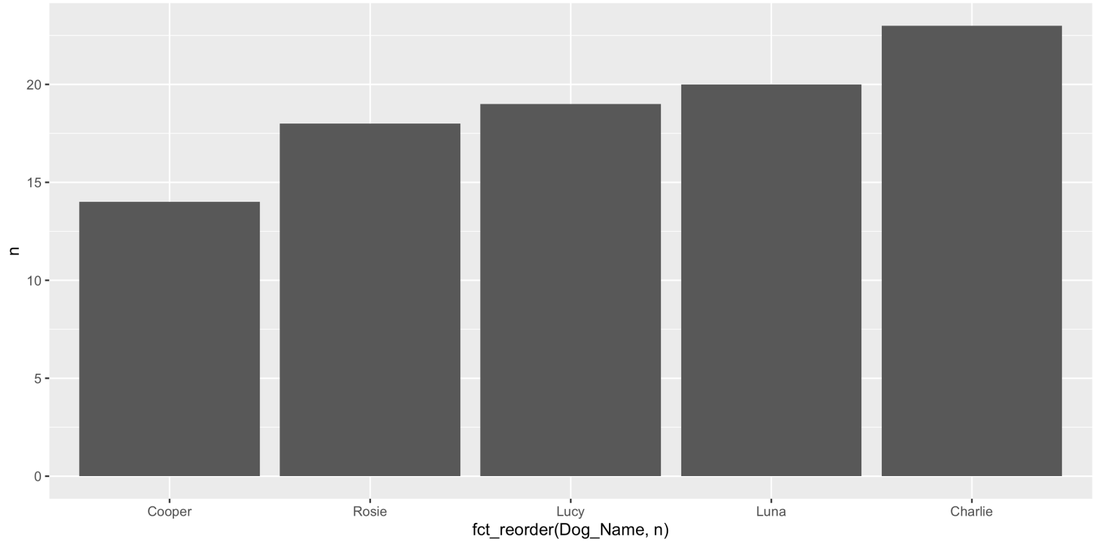
Another `ggplot2` `geom`: `geom_col()`

And you can use `fct_reorder` to order bars by value

```
1 dog_counts <- count(dogs_top5, Dog_Name)
2 dog_counts
```

```
  Dog_Name  n
1  Charlie 23
2   Cooper 14
3     Lucy 19
4     Luna 20
5    Rosie 18
```

```
1 library(forcats)
2 ggplot(data = dog_counts,
3        mapping = aes(x = fct_reorder(Dog_Name, n),
4                      y = n)) +
5   geom_col()
```



Contingency Table

To compare 2 categorical variables, we can use a contingency table

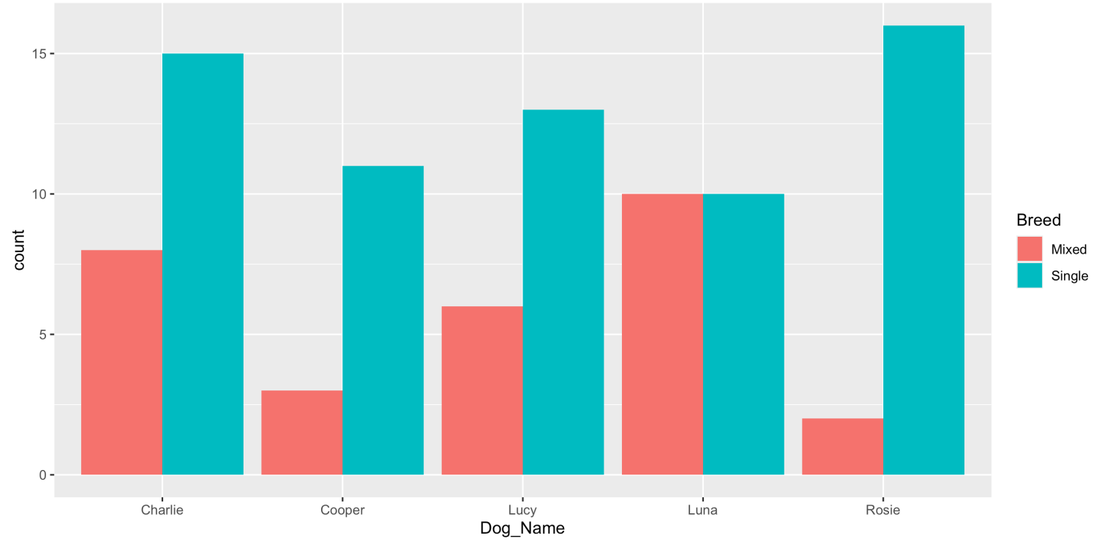
```r
# Using `count()` from dplyr
count(dogs_top5, Dog_Name, Breed)
```

```
   Dog_Name  Breed  n
1   Charlie  Mixed  8
2   Charlie Single 15
3    Cooper  Mixed  3
4    Cooper Single 11
5      Lucy  Mixed  6
6      Lucy Single 13
7      Luna  Mixed 10
8      Luna Single 10
9     Rosie  Mixed  2
10    Rosie Single 16
```

```r
# More common presentation
table(dogs_top5$Dog_Name, dogs_top5$Breed)
```

```
         Mixed Single
Charlie      8     15
Cooper       3     11
Lucy         6     13
Luna        10     10
Rosie        2     16
```

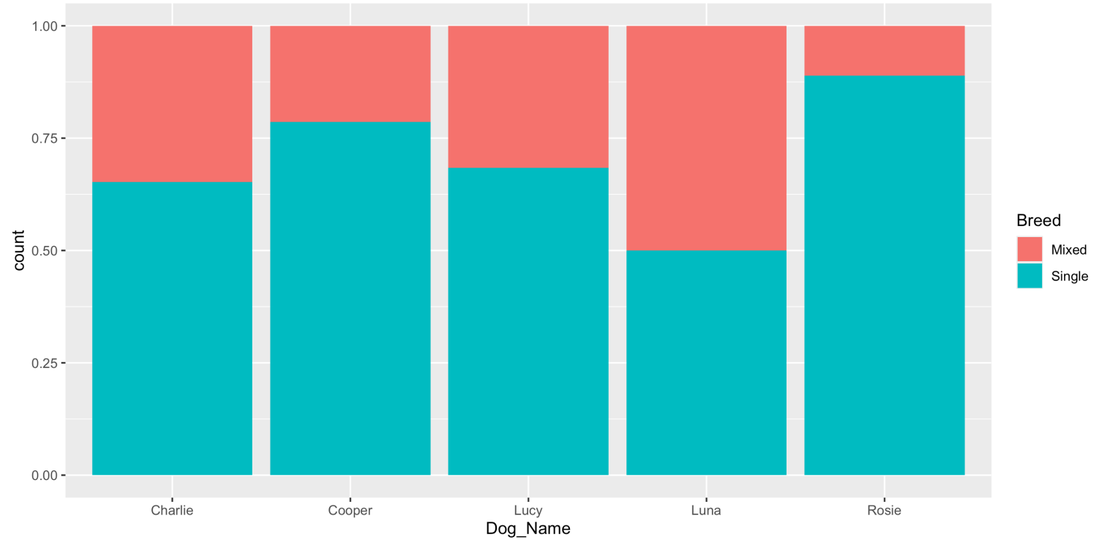
```r
ggplot(data = dogs_top5,
       mapping = aes(x = Dog_Name, fill = Breed)) +
  geom_bar(position = "dodge")
```



Conditional Proportions

- Beyond raw counts, we often summarize categorical data with **conditional** proportions.
  - Especially when looking for relationships!

```
1 ggplot(data = dogs_top5,
2        mapping = aes(x = Dog_Name, fill = Breed)) +
3   geom_bar(position = "fill")
```



Conditional Proportions

```
1 count(dogs_top5, Dog_Name, Breed)
```

```
   Dog_Name  Breed  n
1   Charlie  Mixed  8
2   Charlie Single 15
3    Cooper  Mixed  3
4    Cooper Single 11
5      Lucy  Mixed  6
6      Lucy Single 13
7      Luna  Mixed 10
8      Luna Single 10
9     Rosie  Mixed  2
10    Rosie Single 16
```

```
1 count(dogs_top5, Dog_Name, Breed) %>%
2   group_by(Dog_Name) %>%     ## Conditions on Dog Name
3   mutate(prop = n / sum(n))  ## Adds/changes columns
```

```
# A tibble: 10 × 4
# Groups:   Dog_Name [5]
   Dog_Name Breed      n  prop
   <chr>    <chr>  <int> <dbl>
 1 Charlie  Mixed      8 0.348
 2 Charlie  Single    15 0.652
 3 Cooper   Mixed      3 0.214
 4 Cooper   Single    11 0.786
 5 Lucy     Mixed      6 0.316
 6 Lucy     Single    13 0.684
 7 Luna     Mixed     10 0.5
 8 Luna     Single    10 0.5
 9 Rosie    Mixed      2 0.111
10 Rosie    Single    16 0.889
```

We'll go over these data wrangling functions on Wednesday!

Conditional Proportions

```r
count(dogs_top5, Dog_Name, Breed) %>%
  group_by(Dog_Name) %>%     ## Conditions on Dog Name
  mutate(prop = n / sum(n))  ## Adds/changes columns
```

```
# A tibble: 10 × 4
# Groups:   Dog_Name [5]
   Dog_Name Breed      n  prop
   <chr>    <chr>  <int> <dbl>
 1 Charlie  Mixed      8 0.348
 2 Charlie  Single    15 0.652
 3 Cooper   Mixed      3 0.214
 4 Cooper   Single    11 0.786
 5 Lucy     Mixed      6 0.316
 6 Lucy     Single    13 0.684
 7 Luna     Mixed     10 0.5
 8 Luna     Single    10 0.5
 9 Rosie    Mixed      2 0.111
10 Rosie    Single    16 0.889
```

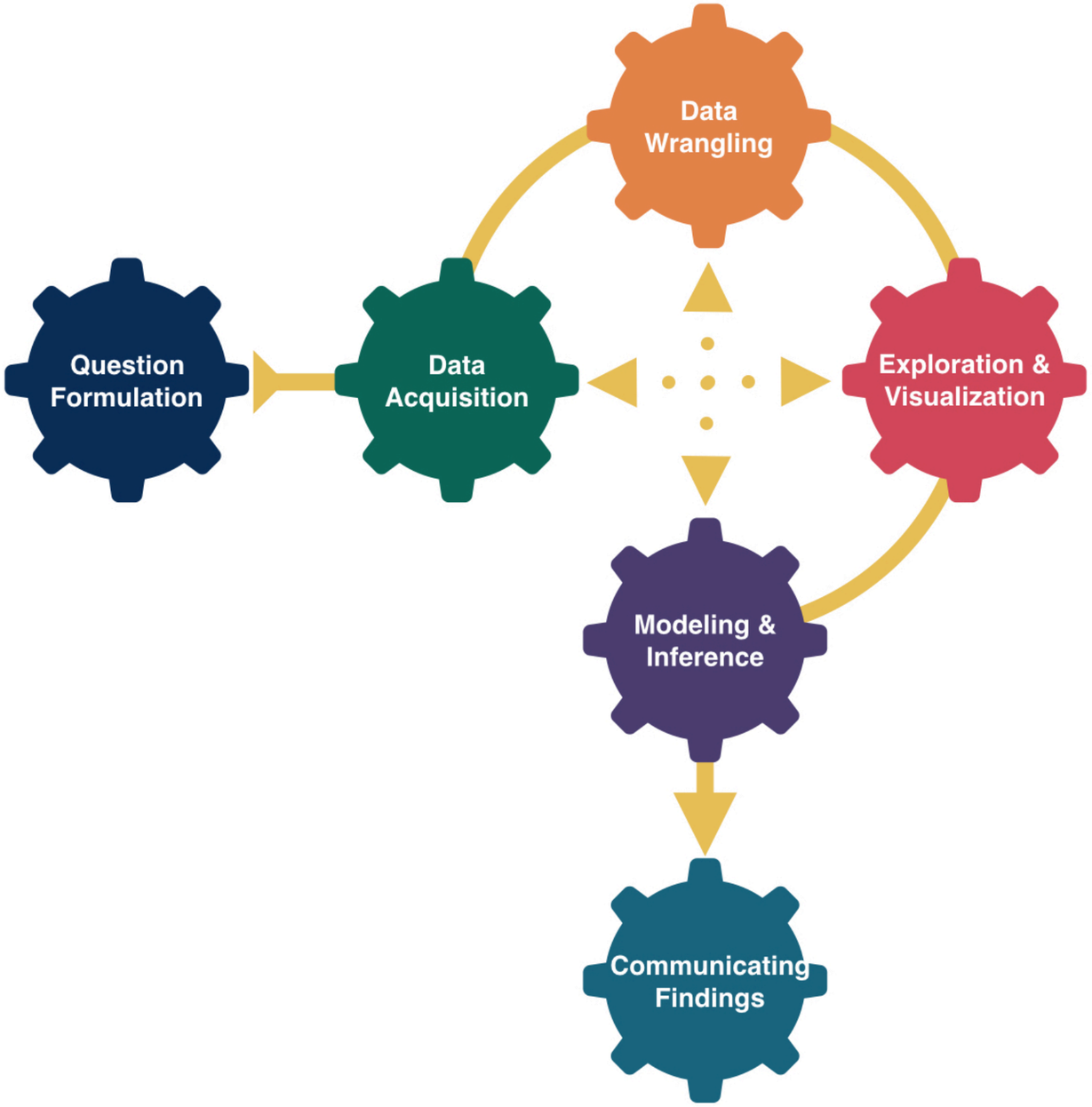
```r
count(dogs_top5, Dog_Name, Breed) %>%
  group_by(Breed) %>%     ## Conditions on Breed
  mutate(prop = n / sum(n))  ## Adds/changes columns
```

```
# A tibble: 10 × 4
# Groups:   Breed [2]
   Dog_Name Breed      n   prop
   <chr>    <chr>  <int>  <dbl>
 1 Charlie  Mixed      8 0.276
 2 Charlie  Single    15 0.231
 3 Cooper   Mixed      3 0.103
 4 Cooper   Single    11 0.169
 5 Lucy     Mixed      6 0.207
 6 Lucy     Single    13 0.2
 7 Luna     Mixed     10 0.345
 8 Luna     Single    10 0.154
 9 Rosie    Mixed      2 0.0690
10 Rosie    Single    16 0.246
```

Q: How does the interpretation change based on which variable you condition on?

Reminders

- The teaching team would love to see you in office hours!
- Next time:
  - We'll define **data wrangling** and
  - Learn to use more functions and methods to summarize and wrangle data



Data Wrangling

Megan Ayers

Math 141 | Spring 2026

Wednesday, Week 2

Goals For Today

- Briefly discuss packages and the difference between base R and `dplyr`
- Learn to wrangle data, mainly with `dplyr`

Packages

- "Base R": core set of built-in functions in R - no extra install required
- R is free and open source - many have contributed **packages**, like `ggplot2`, to make certain things easier to do
- `dplyr`: part of a collection of data science packages called the `tidyverse`

Base R vs `dplyr`

- Throughout the course, we will use **`base R`** and `dplyr` for different tasks.
- We'll try to be consistent with when we use what, but be aware that there are multiple ways to accomplish most tasks.
- Beyond this class as you work with R, I highly recommend developing skills using both base R and `dplyr`. Each has its strengths, both are widely used.

Data for today

```r
dogs <- read.csv("https://data.cambridgema.gov/api/views/sckh-3xyx/rows.csv")

# Useful wrangling that we will come back to
dogs <- dogs[, c("Dog_Name", "Dog_Breed", "Neighborhood")]

dogs_top5 <- dogs %>%
  mutate(Dog_Breed = case_when(Dog_Breed == "Mixed Breed" ~ "Mixed",
                               TRUE ~ "Single")) %>%
  filter(Dog_Name %in% c("Luna", "Charlie", "Lucy", "Cooper", "Rosie"))
```

Data for today

```
1 str(dogs)
```

```
'data.frame':   2794 obs. of  3 variables:
 $ Dog_Name    : chr  "Grace" "Butch" "RUDI 3" "Phoebe" ...
 $ Dog_Breed   : chr  "Mixed Breed" "Mixed Breed" "Shih Tzu" "Labradoodle" ...
 $ Neighborhood: chr  "Baldwin" "North Cambridge" "Riverside" "West Cambridge" ...
```

```
1 str(dogs_top5)
```

```
'data.frame':   94 obs. of  3 variables:
 $ Dog_Name    : chr  "Luna" "Cooper" "Charlie" "Cooper" ...
 $ Dog_Breed   : chr  "Mixed" "Single" "Mixed" "Single" ...
 $ Neighborhood: chr  "North Cambridge" "Neighborhood Nine" "Mid-Cambridge" "West Cambridge" ...
```

Data Wrangling: Transformations done on the data

Why wrangle the data?

To **summarize** the data.

→ To compute the mean and standard deviation of miles ridden by bikeshare users.

To **drop** missing values. (Need to be careful here!)

→ In our Lab 2, we'll see that `ggplot2` will often drop observations before creating a graph.

To **filter** to a particular subset of the data.

→ To subset the `pdxTrees` data to just a few species types.

To **collapse** the categories of a categorical variable.

→ To go from 86 dog breeds to just mixed or single breed.

Data Wrangling: Transformations done on the data

Why wrangle the data?

To **arrange** the data to make it easier to display. → To sort from most common dog name to least common.

To fix how R **stores** a variable (or make a new one). → Converting quantitative variables to/from categorical variables

`dplyr` for Data Wrangling

- Five common wrangling verbs:
  - `count()`
  - `filter()`
  - `arrange()`
  - `mutate()`
  - `summarize()`
- One action:
  - `group_by()`

Motivation for “filtering”

```
1 count(dogs, Dog_Name)
```

	Dog_Name	n
1	ABBY	2
2	ADELAIDE	2
3	AINSLEY	2
4	AJAX	2
5	ALLI	1
6	ANGUS	2
7	ANNIE	3
8	ARCHIE	1
9	Abby	2
10	Abel	2
11	Abigail	2
12	Acadia	2
13	Ace	1

Filtering cases

We extract rows according to logical conditions about variable cases:

```
1 freddy <- filter(dogs, Dog_Name == "Freddy")
2 freddy
```

```
  Dog_Name Dog_Breed      Neighborhood
1   Freddy  Shep Mix Neighborhood Nine
2   Freddy  Shep Mix Neighborhood Nine
```

```
1 not_freddy <- filter(dogs, Dog_Name != "Freddy")
2 head(not_freddy)
```

```
    Dog_Name   Dog_Breed      Neighborhood
1      Grace Mixed Breed           Baldwin
2      Butch Mixed Breed   North Cambridge
3     RUDI 3    Shih Tzu         Riverside
4     Phoebe Labradoodle    West Cambridge
5     Kenobi       Husky Neighborhood Nine
6 WILLOUGHBY    Shih Tzu   North Cambridge
```

Filtering cases

We extract rows according to logical conditions about variable cases:

```
filter(dogs, Dog_Name %in% c("Freddy", "Bo"))
```

```
  Dog_Name      Dog_Breed      Neighborhood
1   Freddy       Shep Mix Neighborhood Nine
2       Bo    Mixed Breed   North Cambridge
3   Freddy       Shep Mix Neighborhood Nine
4       Bo    Mixed Breed   North Cambridge
5       Bo Cocker Spaniel     Mid-Cambridge
```

```
filter(dogs, Dog_Name == "Freddy" | Dog_Name == "Bo")
```

```
  Dog_Name      Dog_Breed      Neighborhood
1   Freddy       Shep Mix Neighborhood Nine
2       Bo    Mixed Breed   North Cambridge
3   Freddy       Shep Mix Neighborhood Nine
4       Bo    Mixed Breed   North Cambridge
5       Bo Cocker Spaniel     Mid-Cambridge
```

```
filter(dogs, Dog_Name == "Rosie" & Dog_Breed == "Goldendoodle")
```

```
  Dog_Name    Dog_Breed   Neighborhood
1    Rosie Goldendoodle West Cambridge
2    Rosie Goldendoodle West Cambridge
```

The pipe: %>%

The pipe operator %>% is a handy way to string multiple commands together in dplyr. Instead of:

```
1 not_freddy <- filter(dogs, Dog_Name != "Freddy")
2 head(not_freddy)
```

```
   Dog_Name   Dog_Breed      Neighborhood
1      Grace Mixed Breed           Baldwin
2      Butch Mixed Breed   North Cambridge
3     RUDI 3    Shih Tzu         Riverside
4     Phoebe Labradoodle    West Cambridge
5     Kenobi       Husky Neighborhood Nine
6 WILLOUGHBY    Shih Tzu   North Cambridge
```

… we can write:

```
1 dogs %>% filter(Dog_Name != "Freddy") %>% head()
```

```
   Dog_Name   Dog_Breed      Neighborhood
1      Grace Mixed Breed           Baldwin
2      Butch Mixed Breed   North Cambridge
3     RUDI 3    Shih Tzu         Riverside
4     Phoebe Labradoodle    West Cambridge
5     Kenobi       Husky Neighborhood Nine
6 WILLOUGHBY    Shih Tzu   North Cambridge
```

- **Q:** Which do you prefer?
- As you develop coding skills, consider:
  - Readability
  - Minimizing the # of defined variables
  - Ease of debugging
  - Personal preference

Extracting variables

We can use $ to extract a single variable

```
dogs_top5$Dog_Name
```

```
 [1] "Luna"    "Cooper"  "Charlie" "Cooper"  "Rosie"   "Rosie"   "Luna"
 [8] "Rosie"   "Luna"    "Rosie"   "Lucy"    "Charlie" "Cooper"  "Cooper"
[15] "Lucy"    "Luna"    "Cooper"  "Lucy"    "Charlie" "Charlie" "Luna"
[22] "Lucy"    "Charlie" "Charlie" "Lucy"    "Lucy"    "Charlie" "Lucy"
[29] "Luna"    "Cooper"  "Cooper"  "Luna"    "Luna"    "Charlie" "Charlie"
[36] "Luna"    "Lucy"    "Charlie" "Charlie" "Lucy"    "Rosie"   "Lucy"
[43] "Rosie"   "Cooper"  "Rosie"   "Lucy"    "Rosie"   "Lucy"    "Cooper"
[50] "Luna"    "Cooper"  "Luna"    "Lucy"    "Rosie"   "Cooper"  "Charlie"
[57] "Charlie" "Rosie"   "Charlie" "Rosie"   "Rosie"   "Luna"    "Lucy"
[64] "Luna"    "Charlie" "Rosie"   "Rosie"   "Charlie" "Luna"    "Cooper"
[71] "Charlie" "Charlie" "Rosie"   "Cooper"  "Lucy"    "Luna"    "Luna"
[78] "Rosie"   "Luna"    "Lucy"    "Lucy"    "Luna"    "Rosie"   "Charlie"
[85] "Luna"    "Charlie" "Lucy"    "Charlie" "Charlie" "Rosie"   "Charlie"
[92] "Luna"    "Cooper"  "Lucy"
```

To extract multiple variables, can use syntax like `df[, c("var1", "var2")]`.

```
head(dogs_top5[, c("Dog_Name", "Dog_Breed")])
```

```
  Dog_Name Dog_Breed
1     Luna     Mixed
2   Cooper    Single
3  Charlie     Mixed
4   Cooper    Single
```

```
5    Rosie    Single
6    Rosie    Single
```

FYI: We can also filter cases with syntax like `df[condition, ]`

Rows and columns can be subset at the same time:

`df[--rows to extract--, --columns to extract--]`.

In this class we will rarely use this syntax, but it is common in the wild.

```
keep_names <- c("Maggie", "Stella")
dogs[dogs$Dog_Name %in% keep_names, c("Dog_Name", "Dog_Breed")]
```

```
     Dog_Name      Dog_Breed
42     Maggie    Mixed Breed
414    Stella    Mixed Breed
578    Stella     Poodle mix
776    Stella   Goldendoodle
797    Maggie    Mixed Breed
836    Maggie    Mixed Breed
1357   Maggie    Mixed Breed
1475   Maggie    Mixed Breed
1653   Stella   Goldendoodle
1915   Stella french bulldog
2000   Maggie    Mixed Breed
2074   Stella     Poodle mix
```

Arranging cases

Use `arrange()` to change the order of cases in a data frame

```
dogs_top5 %>%
  arrange(Dog_Name)
```

```
   Dog_Name Dog_Breed        Neighborhood
1   Charlie     Mixed       Mid-Cambridge
2   Charlie    Single      West Cambridge
3   Charlie    Single   Neighborhood Nine
4   Charlie    Single       Cambridgeport
5   Charlie    Single Cambridge Highlands
6   Charlie    Single      West Cambridge
7   Charlie     Mixed       Mid-Cambridge
8   Charlie    Single       Cambridgeport
9   Charlie     Mixed       Mid-Cambridge
10  Charlie    Single       Cambridgeport
11  Charlie     Mixed       Mid-Cambridge
12  Charlie     Mixed            The Port
13  Charlie     Mixed     Strawberry Hill
```

Creating a new variable

We use `mutate()` to create new variables, or overwrite existing ones

```
1 head(dogs_top5)
```

```
  Dog_Name Dog_Breed      Neighborhood
1     Luna     Mixed   North Cambridge
2   Cooper    Single Neighborhood Nine
3  Charlie     Mixed     Mid-Cambridge
4   Cooper    Single    West Cambridge
5    Rosie    Single    West Cambridge
6    Rosie    Single         Riverside
```

```
1 dogs_top5 %>%
2   mutate(animal = "dog",
3          Dog_Name = toupper(Dog_Name)) %>%
4   head()
```

```
  Dog_Name Dog_Breed      Neighborhood animal
1     LUNA     Mixed   North Cambridge    dog
2   COOPER    Single Neighborhood Nine    dog
3  CHARLIE     Mixed     Mid-Cambridge    dog
4   COOPER    Single    West Cambridge    dog
5    ROSIE    Single    West Cambridge    dog
6    ROSIE    Single         Riverside    dog
```

New Data Setting: Bureau of Labor Statistics (BLS) Consumer Expenditure Survey

BLS Mission: “Measures labor market activity, working conditions, price changes, and productivity in the U.S. economy to support public and private decision making.”

Data: Last quarter of the 2016 BLS Consumer Expenditure Survey.

```
1 ce_raw <- read.csv("data/fmli.csv", na = c("NA", "."))
2 str(ce_raw)
```

```
'data.frame':   6301 obs. of  51 variables:
 $ NEWID   : int  3324174 3324204 3324214 3324244 3324274 3324284 3324294 3324304 3324324 3324334 ...
 $ PRINEARN: int  1 1 1 1 2 1 1 1 2 1 ...
 $ FINLWT21: num  25985 6581 20208 18078 20112 ...
 $ FINCBTAX: int  116920 200 117000 0 2000 942 0 91000 95000 40037 ...
 $ BLS_URBN: int  1 1 1 1 1 1 1 1 2 1 ...
 $ POPSIZE : int  2 3 4 2 2 2 1 2 5 2 ...
 $ EDUC_REF: int  16 15 16 15 14 11 10 13 12 12 ...
 $ EDUCA2  : int  15 15 13 NA NA NA NA 15 15 14 ...
 $ AGE_REF : int  63 50 47 37 51 63 77 37 51 64 ...
 $ AGE2    : int  50 47 46 NA NA NA NA 36 53 67 ...
 $ SEX_REF : int  1 1 2 1 2 1 2 1 1 2 ...
 $ SEX2    : int  2 2 1 NA NA NA NA 2 2 1 ...
 $ REF_RACE: int  1 4 2 1 1 1 1 1 1 1 ...
 $ RACE2   : int  1 4 1 NA NA NA NA 1 1 1 ...
 $ HISP_REF: int  2 2 2 2 2 1 1 2 2 2 ...
 $ HISP2   : int  2 2 1 NA NA NA NA 2 2 2 ...
 $ FAM_TYPE: int  3 4 1 8 9 9 8 3 1 1 ...
 $ MARITAL1: int  1 1 1 5 3 3 2 1 1 1 ...
 $ REGION  : int  4 4 3 4 4 3 4 1 3 2 ...
```

Wrangling CE Data

Want to better understand a family's income and expenditures

```
ce <- ce_raw[, c("NEWID", "PRINEARN", "FINCBTAX", "BLS_URBN",
                 "HIGH_EDU", "TOTEXPCQ", "IRAX")]
dim(ce)
```

```
[1] 6301    7
```

Variables:

- `NEWID`: ID for the household
- `PRINEARN`: ID for which member of the household is the principal earner
- `FINCBTAX`: Final income before taxes for the year
- `BLS_URBN`: 1 = urban, 2 = rural
- `HIGH_EDU`: Highest education in the household. 00 = Never attended, 10 = Grades 1-8, 11 = Grades 9-12, no degree, 12 = High school graduate, 13 = Some college, no degree, 14 = Associates degree, 15 = Bachelor's degree, 16 = Masters, Professional/doctorate degree
- `TOTEXPCQ` = Total household expenditures for the current quarter
- `IRAX` = Total in retirement funds

Wrangling CE Data

Q: What is this code doing?

```
ce <- ce %>%
  mutate(YEARLY_EXP = TOTEXPCQ*4)
ce
```

	NEWID	PRINEARN	FINCBTAX	BLS_URBN	HIGH_EDU	TOTEXPCQ	IRAX	YEARLY_EXP
1	3324174	1	116920	1	16	0.0000	1000000	0.0000
2	3324204	1	200	1	15	0.0000	10000	0.0000
3	3324214	1	117000	1	16	0.0000	0	0.0000
4	3324244	1	0	1	15	0.0000	NA	0.0000
5	3324274	2	2000	1	14	0.0000	NA	0.0000
6	3324284	1	942	1	11	0.0000	0	0.0000
7	3324294	1	0	1	10	0.0000	0	0.0000
8	3324304	1	91000	1	15	0.0000	15000	0.0000
9	3324324	2	95000	2	15	0.0000	NA	0.0000
10	3324334	1	40037	1	14	0.0000	477000	0.0000
11	3324394	2	109000	1	12	0.0000	NA	0.0000
12	3324404	1	4104	1	12	0.0000	NA	0.0000
13	3324424	1	0	1	12	0.0000	NA	0.0000

Using logical operators

Q: What is this code doing?

```
ce_sub <- ce %>%
  filter(YEARLY_EXP > 0, BLS_URBN == 1, HIGH_EDU != "00")
ce_sub
```

```
      NEWID PRINEARN FINCBTAX BLS_URBN HIGH_EDU   TOTEXPCQ    IRAX  YEARLY_EXP
1   3335204        1    37000        1       14  2491.9167       0   9967.6668
2   3335214        1   103000        1       16  6128.3334      NA  24513.3336
3   3335224        1    14686        1       13  1071.8000      NA   4287.2000
4   3335244        2    33396        1       12  1630.0000       0   6520.0000
5   3335264        1        0        1       13  3213.3667      NA  12853.4668
6   3335274        1        0        1       15  4673.5833       0  18694.3332
7   3335294        1   745136        1       16  8693.3727  280000  34773.4908
8   3335304        1    36000        1       16  3733.1667      NA  14932.6668
9   3335314        2    45000        1       15  3627.2500    3000  14509.0000
10  3335334        1    20862        1       13   802.3333       0   3209.3332
11  3335344        1        0        1       16  7670.7833      NA  30683.1332
12  3335354        2    47662        1       13  4179.8833  400000  16719.5332
13  3335374        1    10560        1       15  2728.3333      NA  10913.3332
```

Using logical operators

Q: What is this code doing?

```
ce_sub <- ce %>%
  filter(YEARLY_EXP > 0, (BLS_URBN == 1 | HIGH_EDU != "00"))
ce_sub
```

	NEWID	PRINEARN	FINCBTAX	BLS_URBN	HIGH_EDU	TOTEXPCQ	IRAX	YEARLY_EXP
1	3335204	1	37000	1	14	2491.9167	0	9967.6668
2	3335214	1	103000	1	16	6128.3334	NA	24513.3336
3	3335224	1	14686	1	13	1071.8000	NA	4287.2000
4	3335244	2	33396	1	12	1630.0000	0	6520.0000
5	3335264	1	0	1	13	3213.3667	NA	12853.4668
6	3335274	1	0	1	15	4673.5833	0	18694.3332
7	3335294	1	745136	1	16	8693.3727	280000	34773.4908
8	3335304	1	36000	1	16	3733.1667	NA	14932.6668
9	3335314	2	45000	1	15	3627.2500	3000	14509.0000
10	3335334	1	20862	1	13	802.3333	0	3209.3332
11	3335344	1	0	1	16	7670.7833	NA	30683.1332
12	3335354	2	47662	1	13	4179.8833	400000	16719.5332
13	3335374	1	10560	1	15	2728.3333	NA	10913.3332

Using logical operators

Q: Which version leaves more cases in the data frame?

```
ce_sub <- ce %>%
  filter(YEARLY_EXP > 0, BLS_URBN == 1, HIGH_EDU != "00")
```

```
ce_sub <- ce %>%
  filter(YEARLY_EXP > 0, (BLS_URBN == 1 | HIGH_EDU != "00"))
```

Using logical operators

Q: Which version leaves more cases in the data frame?

```
1 ce_sub <- ce %>%
2   filter(YEARLY_EXP > 0, BLS_URBN == 1, HIGH_EDU != "00")
3 dim(ce_sub)
```

```
[1] 3954    8
```

```
1 ce_sub <- ce %>%
2   filter(YEARLY_EXP > 0, (BLS_URBN == 1 | HIGH_EDU != "00"))
3 dim(ce_sub)
```

```
[1] 4178    8
```

Recoding Variables with `case_when()`

```
1 count(ce, BLS_URBN)
```

```
  BLS_URBN    n
1        1 5952
2        2  349
```

Q: What is this code doing?

```
1 ce <- ce %>%
2   mutate(BLS_URBN = case_when(BLS_URBN == 1 ~ "Urban",
3                               BLS_URBN == 2 ~ "Rural",
4                               TRUE ~ NA))
5
6 count(ce, BLS_URBN)
```

```
  BLS_URBN    n
1    Rural  349
2    Urban 5952
```

Creating Variables with `case_when()`

```
ce %>%
  mutate(HIGH_EDU = as.numeric(HIGH_EDU)) %>%
  count(HIGH_EDU)
```

```
  HIGH_EDU    n
1        0    8
2       10  110
3       11  302
4       12 1272
5       13 1297
6       14  714
7       15 1528
8       16 1070
```

```
ce <- ce %>%
  mutate(HIGH_EDU2 = case_when(HIGH_EDU <= 11 ~ "Less than high school degree",
                               HIGH_EDU <= 13 ~ "High school degree",
                               HIGH_EDU >= 14 ~ "College degree",
                               TRUE ~ NA))

count(ce, HIGH_EDU2)
```

```
                  HIGH_EDU2    n
1             College degree 3312
2          High school degree 2569
3 Less than high school degree  420
```

Variable Names

Sometimes datasets come with terrible variable names.

```
1 names(ce)
```

```
[1] "NEWID"      "PRINEARN"   "FINCBTAX"   "BLS_URBN"   "HIGH_EDU"
[6] "TOTEXPCQ"   "IRAX"       "YEARLY_EXP" "HIGH_EDU2"
```

We can fix that with `rename()`.

```
1 ce <- ce %>%
2   rename(INCOME = FINCBTAX)
```

Handling Missing Data

Want to compute mean income and mean retirement funds.

```
1 mean(ce$INCOME)
```

```
[1] 62480.12
```

```
1 mean(ce$IRAX)
```

```
[1] NA
```

```
1 ce_aggressive <- na.omit(ce_raw)
2 ce_aggressive
```

```
 [1] NEWID    PRINEARN FINLWT21 FINCBTAX BLS_URBN POPSIZE
EDUC_REF EDUCA2
 [9] AGE_REF  AGE2     SEX_REF  SEX2     REF_RACE RACE2
HISP_REF HISP2
[17] FAM_TYPE MARITAL1 REGION   SMSASTAT HIGH_EDU EHOUSNGC
TOTEXPCQ FOODCQ
[25] TRANSCQ  HEALTHCQ ENTERTCQ EDUCACQ  TOBACCCQ STUDFINX
IRAX     CUTENURE
[33] FAM_SIZE VEHQ     ROOMSQ   INC_HRS1 INC_HRS2 EARNCOMP
NO_EARNR OCCUCOD1
[41] OCCUCOD2 STATE    DIVISION TOTXEST  CREDFINX CREDITB
CREDITX  BUILDING
[49] ST_HOUS  INT_PHON INT_HOME
<0 rows> (or 0-length row.names)
```

- Q: Guesses about what's going on here?
- A: `na.omit()` drops a row if it has a missing value in *any* column.

Handling Missing Data

```
ce_moderate <- ce %>%
  filter(!is.na(IRAX), !is.na(INCOME))

mean(ce_moderate$INCOME)
```

```
[1] 57771.14
```

```
mean(ce_moderate$IRAX)
```

```
[1] 91075.09
```

Or, to be very conservative about preserving data:

```
mean(ce$INCOME, na.rm = TRUE)
```

```
[1] 62480.12
```

```
mean(ce$IRAX, na.rm = TRUE)
```

```
[1] 91075.09
```

Q: Why does the mean income value change between these approaches, even though there are no NA income values?

A: The moderate approach still drops `INCOME` values if a row is missing `IRAX`.

Summarizing within groups using `group_by()` and `summarize()`

```
ce_moderate %>%
  group_by(BLS_URBN) %>%
  summarize(mean_IRAX = mean(IRAX))
```

```
# A tibble: 2 × 2
  BLS_URBN mean_IRAX
  <chr>        <dbl>
1 Rural       37008.
2 Urban       94512.
```

```
ce_moderate %>%
  group_by(BLS_URBN, HIGH_EDU2) %>%
  summarize(mean_IRAX = mean(IRAX))
```

```
# A tibble: 6 × 3
# Groups:   BLS_URBN [2]
  BLS_URBN HIGH_EDU2                    mean_IRAX
  <chr>    <chr>                            <dbl>
1 Rural    College degree                 105148.
2 Rural    High school degree              15543.
3 Rural    Less than high school degree        0
4 Urban    College degree                 168767.
5 Urban    High school degree              30533.
6 Urban    Less than high school degree     8270.
```

Naming Wrangled Data

When I make a new data frame, what name should I give it? Importantly, should I **write over my original dataframe** or should I **save a new dataframe**?

- Advice:
  - Is your new data frame structurally different? If so, give it a **new name**.
  - Are you removing values you will need for a future analysis within the same document? If so, give it a **new name**.
  - Are you just adding to or cleaning the data? If so, then **write over** the original.

Sage Advice from ModernDive

> “Crucial: Unless you are very confident in what you are doing, it is worthwhile not starting to code right away. Rather, first sketch out on paper all the necessary data wrangling steps not using exact code, but rather high-level pseudocode that is informal yet detailed enough to articulate what you are doing. This way you won’t confuse what you are trying to do (the algorithm) with how you are going to do it [(writing code)].”

In-Class Exercise: Data wrangling pseudocode

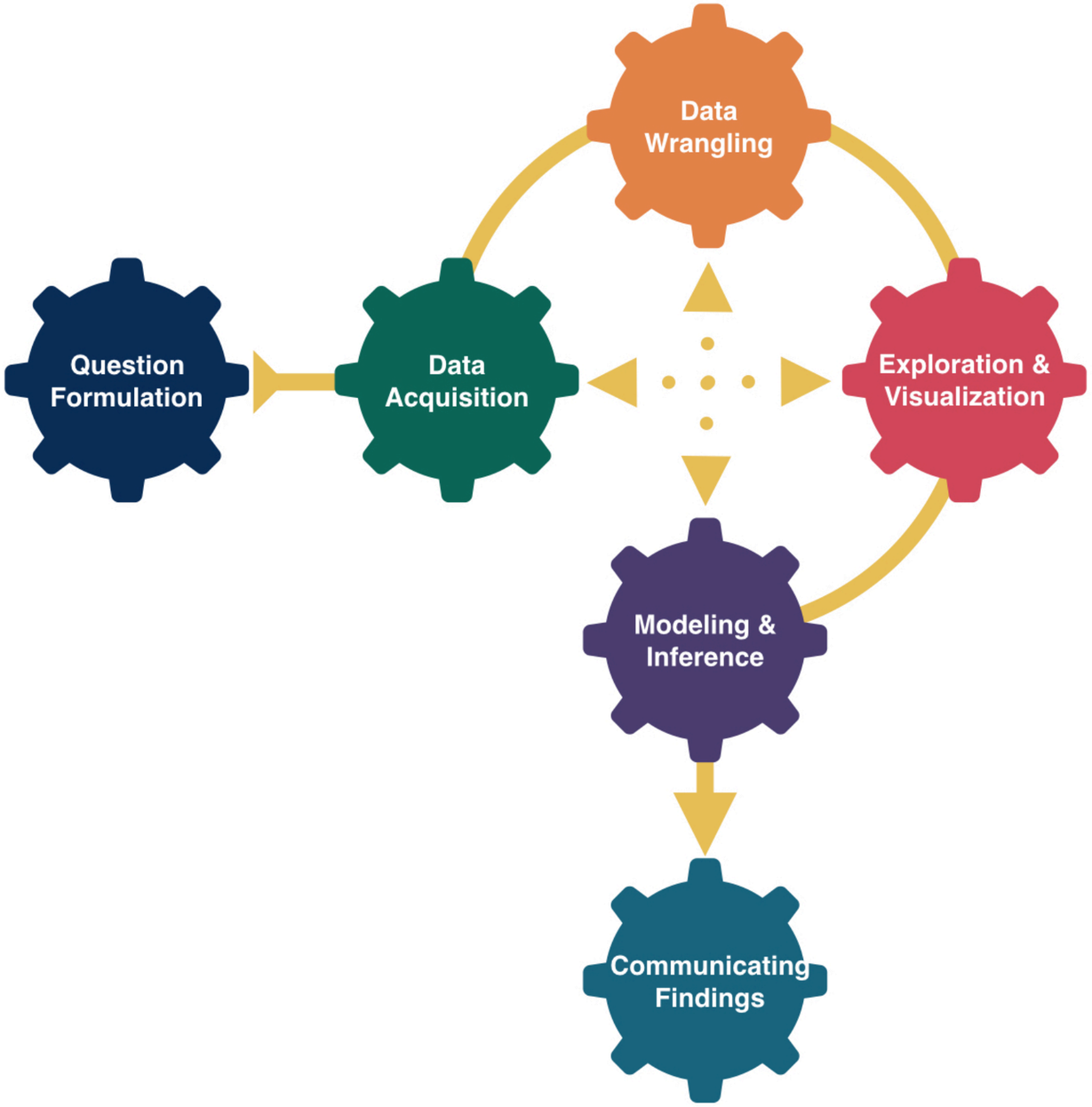
Anonymized data from a 2017 study on the effect of the higher education system on upward mobility (Chetty et al. 2017).

```
load("data/colleges.Rdata")
head(colleges[, c("region", "state", "tier_name", "name", "mr_kq5_pq1")])
```

```
# A tibble: 6 × 5
  region state tier_name                  name                          mr_kq5_pq1
  <fct>  <chr> <fct>                      <chr>                              <dbl>
1 3      TX    Selective private          Abilene Christian Univers…    0.0144
2 3      GA    Nonselective 4-year public Abraham Baldwin Agricultu…    0.0149
3 4      CO    Selective public           Adams State University        0.0188
4 1      NY    Selective private          Adelphi University            0.0326
5 2      MI    Selective private          Adrian College                0.00956
6 3      FL    Nonselective 4-year private Adventist University Of H…   0.0379
```

- `region`: 1 for Northeast, 2 for Midwest, 3 for South, and 4 for West
- `state`: State name
- `tier_name`: Tier defined by selectivity and type
- `mr_kq5_pq1`: Mobility rate, top 20% of the income distribution.

In groups of 3-4, on paper or the board, write pseudocode to answer: what states in the west have the highest average mobility rate across their private and elite colleges?



Probability I

Megan Ayers

Math 141 | Spring 2026

Friday, Week 2

Announcements/reminders

- Compiled PDF of all slides now on the course website
- HW 1 due today at 11:59pm

Goals For Today

- Introduce **probability theory**
- See the **Law of Large Numbers**
- Define and practice using key **rules of probability**

Probability Theory

Random Processes and Events

Probability Theory: the study and quantification of uncertainty/randomness in outcomes of repeated experiments.

A **random process** is one which we know what results *could* happen, but don't know which particular result *will* happen.

- Can be used to model processes that are complicated, but not truly random, to figure out how they work
- Example: Rolling a 6-sided die

An **event** is a potential result of some particular random process.

- Example: If we roll a 6-sided die, one event is *"the die rolls a 6"*. Another is *"the die rolls an odd number"*.
- The statement *"it will rain at 9am in Portland on 2/6"* is *an* event, but not one for the die-rolling process.

Perception of Probability

We will soon define the word “probability.” Before that, consider the following questions in small groups:

1. When flipping a fair coin, we might say that “the probability of flipping Heads is 0.5.” How do you interpret this sentence?
   a. If I flip this coin over and over, roughly 50% will be Heads.
   b. Heads and Tails are equally plausible.
   c. Both a and b make sense.

2. An election is coming up and a pollster claims “candidate A has a 0.9 probability of winning”. How do you interpret this sentence?
   a. If we observe the election over and over, candidate A will win roughly 90% of the time.
   b. Candidate A is much more likely to win than to lose.
   c. Both a and b make sense.

Perception of Probability (Notes)

- If you answered (a), you may naturally subscribe to the **"Frequentist"** view of probability (*long-run proportion*)
- If you answered (b), you may naturally subscribe to the **"Bayesian"** view of probability (*relative chance*)
- If you answered (c), you fall somewhere in between!

Neither definition is "correct". Statisticians are divided on how to interpret probability.

- In this class, we'll discuss frequentist probability (see next slide)

Probability

The “frequentist” definition of probability is…

> **Definition: Probability**
>
> The **probability** of a particular event is the proportion of times the event would occur, if we observed the random process an *arbitrarily large* number of times.

To say that *a coin has 50% probability of landing heads*, means that…

- We expect the proportion of heads in a large number of coin flips to be close to 0.5

Since probabilities are defined as a proportion, **they will always be values between 0 and 1**.

For brevity, we’ll represent statements like *the probability of the event “the coin lands heads” is 50%* using the notation:

$$P(\text{Heads}) = 0.5 \quad \text{or} \quad P(H) = 0.5$$

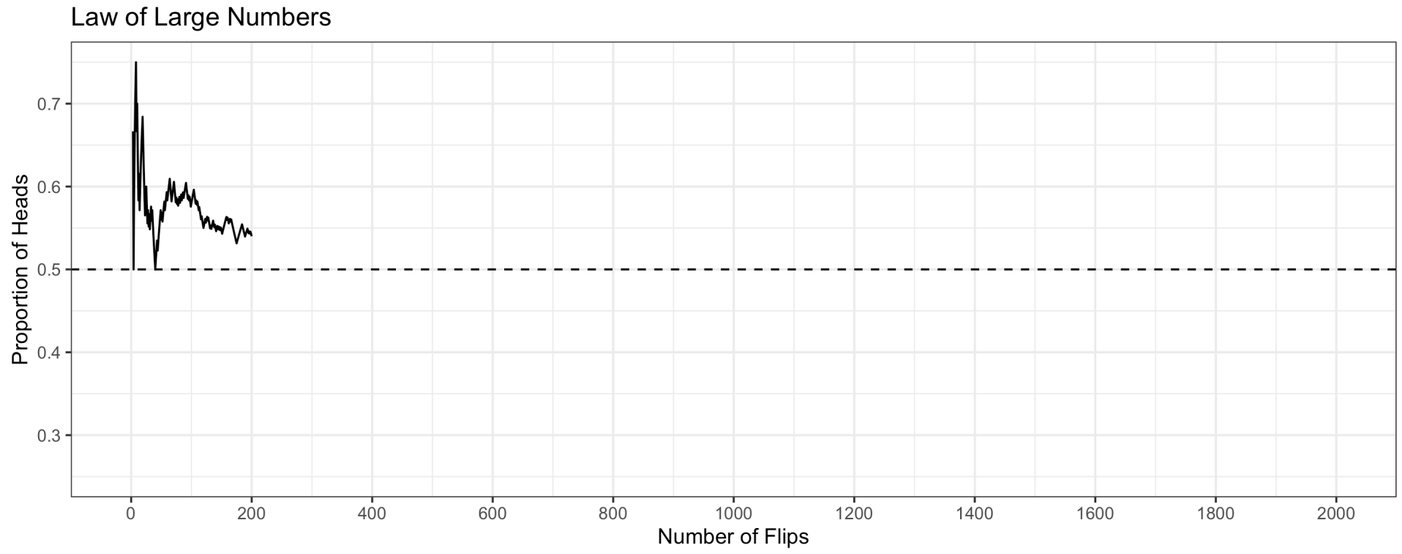
Law of Large Numbers

- **Q:** If I flip a fair coin 10 times, do we expect the proportion of heads observed to be 0.5?
- **Q:** What if I flip it 1000 times?
- **Q:** What do you expect to happen to the proportion of heads as we increase the number of flips?

Law of Large Numbers

The probability of an event refers to the **long-run tendency of the proportion.**

- Observed proportions can deviate from expected probabilities in a finite number of trials
- But as you conduct more and more trials, the deviation decreases

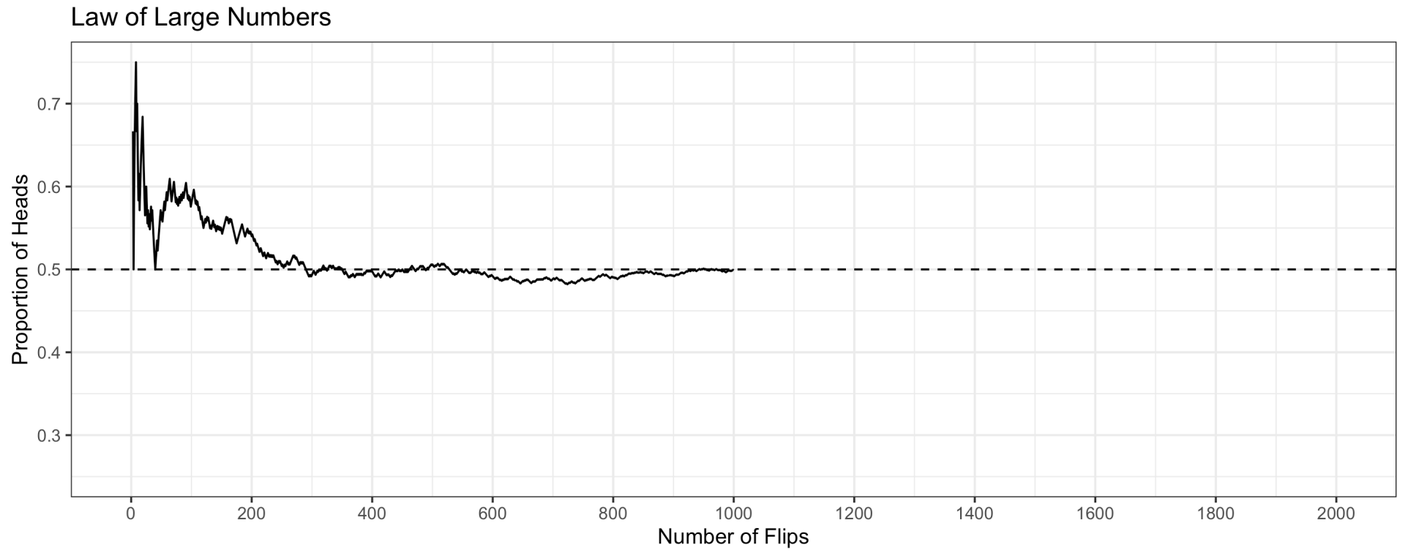


The proportion of heads deviates wildly from 0.5 during the first 200 flips

Law of Large Numbers

The probability of an event refers to the **long-run tendency of the proportion.**

- Observed proportions can deviate from expected probabilities in a finite number of trials
- But as you conduct more and more trials, the deviation decreases

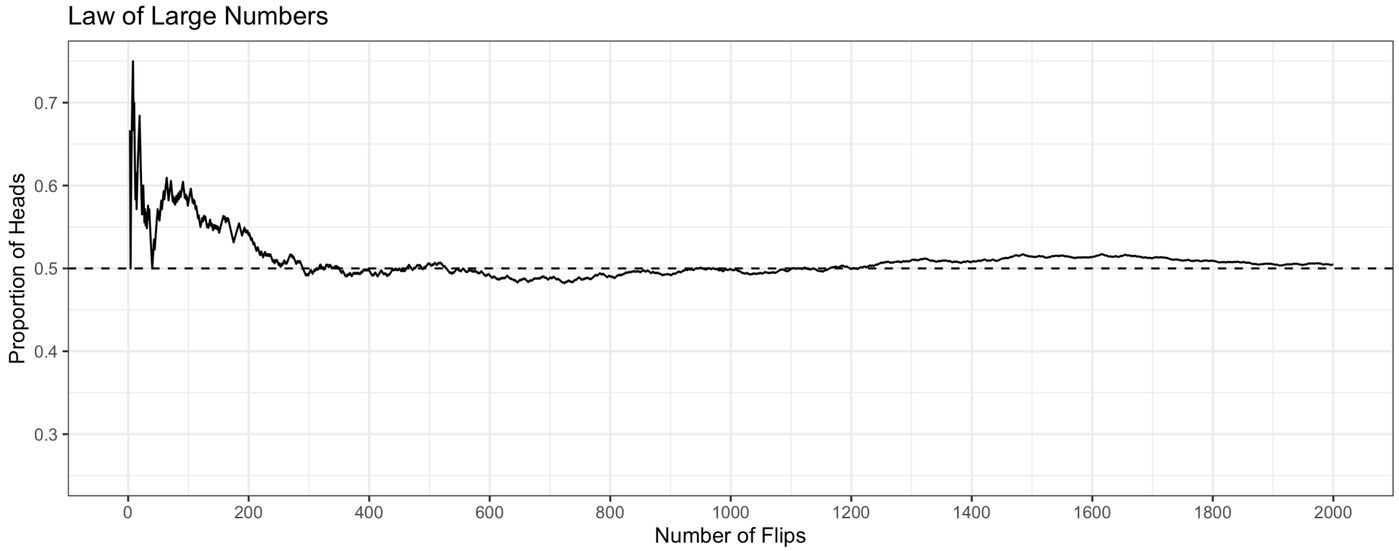


The proportion of heads gets closer to 0.5 after 1000 flips

Law of Large Numbers

The probability of an event refers to the **long-run tendency of the proportion.**

- Observed proportions can deviate from expected probabilities in a finite number of trials
- But as you conduct more and more trials, the deviation decreases

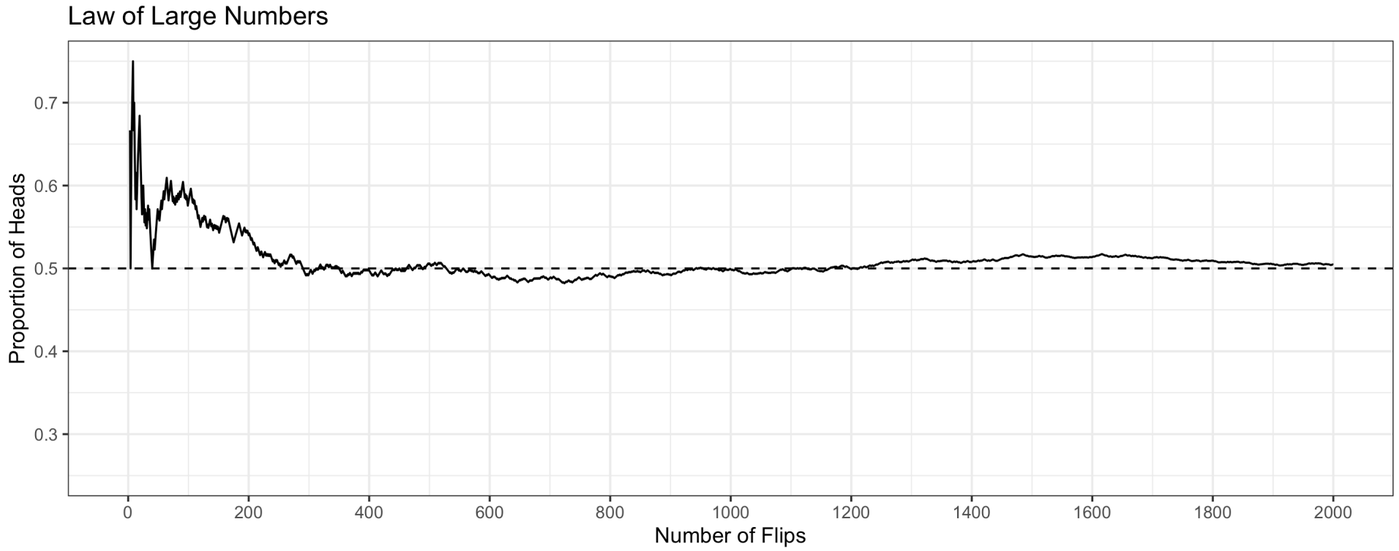


The proportion of heads is very close to 0.5 after 2000 flips!

Law of Large Numbers

The probability of an event refers to the **long-run tendency of the proportion.**

- Observed proportions can deviate from expected probabilities in a finite number of trials
- But as you conduct more and more trials, the deviation decreases



This is the **Law of Large Numbers** in effect!

Check Your Understanding

Suppose you're interested in winning the Powerball Jackpot, where the chance of winning is 1 in 292 million. You buy a single lottery ticket and wait to see if you won.

1. When playing the lottery (a *random process*), what are the two possible outcomes (**events**)?

2. Write down the probability of your two events.

3. If you played the lottery 1 million times, should you **expect** to win? What about 292 million times?

4. If you played the lottery 1 billion times, are you **guaranteed** to win?

Check Your Understanding (Answers)

1. When playing the lottery (a *random process*), what are the two possible outcomes (**events**)?
   - Winning and Losing

2. Write down the probability of your two events.
   - P(Winning) = 1/292,000,000
   - P(Losing) = 291,999,999/292,000,000 or 1 - P(Winning)

3. If you played the lottery 1 million times, should you **expect** to win? What about 292 million times?
   - You shouldn't *expect* to win after playing 1 million times. But you might win after playing 292 million times!

4. If you played the lottery 1 billion times, are you **guaranteed** to win?
   - You are *never guaranteed* to win! Each time you play, your chance of winning is so small!

Probability Models

We can formalize the connection between random processes/events with probabilities using a **Probability Model**

A **probability model** has two components:

1. A list of the possible results of a random process (called **events**)
2. A rule (called the **probability function**) that assigns to each event a probability between 0 and 1, in a consistent manner.

Example: Probability Model for a Coin Toss:

1. Heads or Tails
2. $P(Heads) = 0.5$ and $P(Tails) = 0.5$

When discussing probability, we always (explicitly or implicitly) define a probability model.

- Probability models help us make sense of complex, random processes.

Rules of Probability

Definition: Disjoint Events

Two events A and B are said to be **mutually exclusive** (or **disjoint**) if it is not possible for both to occur at the same time.

Example: Single Roll of a Die

- Let A denote the event "a 1 is rolled on 6-sided die"
- Let B denote the event "a 2 is rolled on 6-sided die"
- A and B are disjoint events for the random process

The Addition Rule

Theorem: Addition Rule

The probability that at least one event occurs in a pair of disjoint events is the sum of their individual probabilities:

$$P(A \text{ or } B) = P(A) + P(B)$$

Q: When a die is rolled, what is the probability that an **odd number** is rolled?

$$P(\text{Odd}) = P(\text{roll a } 1, 3, \text{ or } 5)$$

The Addition Rule

Theorem: Addition Rule

The probability that at least one event occurs in a pair of disjoint events is the sum of their individual probabilities:

$$P(A \text{ or } B) = P(A) + P(B)$$

Q: When a die is rolled, what is the probability that an **odd number** is rolled?

$$\begin{aligned} P(\text{Odd}) &= P(\text{roll a } 1, 3, \text{ or } 5) \\ &= P(\text{roll a } 1) + P(\text{roll a } 3) + P(\text{roll a } 5) \qquad \text{(by the Addition Rule)} \end{aligned}$$

The Addition Rule

Theorem: Addition Rule

The probability that at least one event occurs in a pair of disjoint events is the sum of their individual probabilities:

$$P(A \text{ or } B) = P(A) + P(B)$$

Q: When a die is rolled, what is the probability that an **odd number** is rolled?

$$\begin{aligned} P(\text{Odd}) &= P(\text{roll a } 1, 3, \text{ or } 5) \\ &= P(\text{roll a } 1) + P(\text{roll a } 3) + P(\text{roll a } 5) \qquad \text{(by the Addition Rule)} \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \end{aligned}$$

The Addition Rule

Theorem: Addition Rule

The probability that at least one event occurs in a pair of disjoint events is the sum of their individual probabilities:

$$P(A \text{ or } B) = P(A) + P(B)$$

Q: When a die is rolled, what is the probability that an **odd number** is rolled?

$$\begin{aligned} P(\text{Odd}) &= P(\text{roll a } 1, 3, \text{ or } 5) \\ &= P(\text{roll a } 1) + P(\text{roll a } 3) + P(\text{roll a } 5) \qquad \text{(by the Addition Rule)} \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{3}{6} \end{aligned}$$

The Addition Rule: Only for Disjoint events!

Theorem: Addition Rule

The probability that at least one event occurs in a pair of disjoint events is the sum of their individual probabilities:

$$P(A \text{ or } B) = P(A) + P(B)$$

Q: In a die-roll, what's the probability we get an even number?

- $P(\text{Even}) = P(\text{ Roll a } 2, 4, \text{ or } 6) = \frac{3}{6} = \frac{1}{2}$

Q: In a die-roll, what's the probability we get at least a 3?

- $P(\geq 3) = P(\text{ Roll a } 3, 4, 5, \text{ or } 6) = \frac{4}{6} = \frac{2}{3}$

Q: What's the probability you get an even number OR at least a 3?

- $P(\text{ Even OR} \geq 3) = P(\text{ Roll a } 2, 3, 4, 5, 6) = \frac{5}{6}$

- But $P(\text{Even}) + P(\geq 3) = \frac{1}{2} + \frac{2}{3} = \frac{7}{6}$… that's not even between 0 and 1!

Complementary Events

The **complement** to an event A (denoted A^c) is the event that occurs exactly when the original does not.

- If A is the event that a 1 is rolled on a die, then A^c is the event that any other number (2,3,4,5, or 6) is rolled.

> **Theorem: Complement Rule**
>
> The probability that the complement of an event occurs is 1 minus the probability of the event:
>
> $$P(A^c) = 1 - P(A)$$

What is the probability that any number other than a 1 is rolled on a fair 6-sided die?

$$P(\text{roll something other than a 1}) = 1 - P(\text{roll a 1}) = 1 - \frac{1}{6} = \frac{5}{6}$$

Check Your Understanding

Theorem: Addition Rule

The probability that at least one event occurs in a pair of disjoint events is the sum of their individual probabilities:

$$P(A \text{ or } B) = P(A) + P(B)$$

Theorem: Complement Rule

The probability that the complement of an event occurs is 1 minus the probability of the event:

$$P(A^c) = 1 - P(A)$$

In the city of Portland, there's rain on 60% of days and snow on 1% of days.

1. On a given day, what's the probability that it **doesn't rain?**
2. On a given day, what's the probability that it rains **OR** doesn't rain?
3. What if I told you it rains OR snows 61% of days in Portland. What's the **flaw** in my reasoning?

Check Your Understanding (Answers)

In the city of Portland, there's rain on 60% of days and snow on 1% of days.

1. On a given day, what's the probability that it **doesn't rain?**

$$P(\text{Doesn't Rain}) = P(\text{Rain}^c) = 1 - P(\text{Rain}) = 1 - 0.6 = 0.4$$

2. On a given day, what's the probability that it rains **OR** doesn't rain?

$$P(\text{Rain or Doesn't Rain}) = P(\text{Rain}) + P(\text{Doesn't Rain}) = 0.6 + 0.4 = 1$$

3. What if I told you it rains OR snows 61% of days in Portland. What's the **flaw** in my reasoning?

It can rain *and* snow in one day! Thus, we *can't use the addition rule!*

Recap

Today we defined:

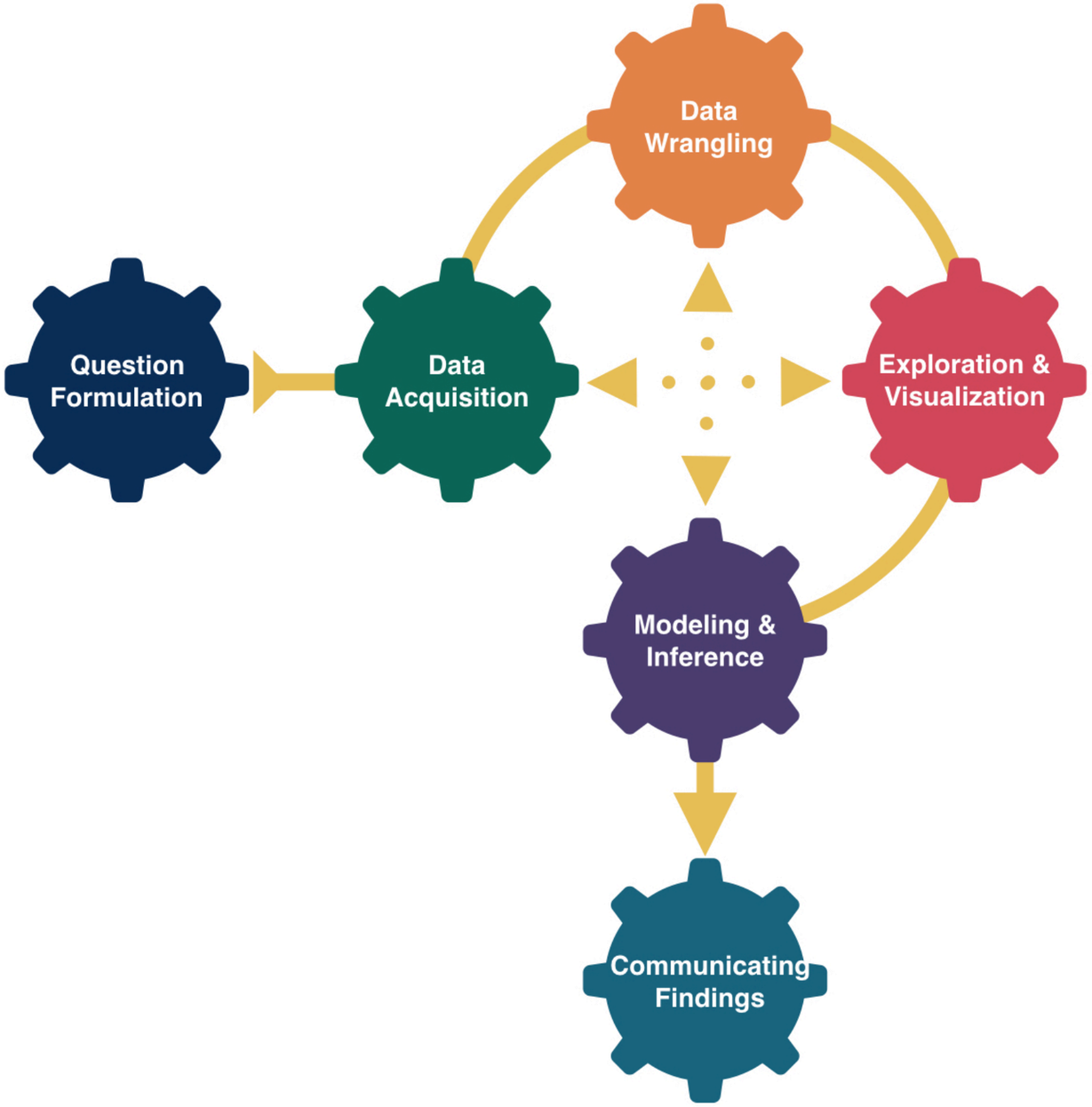
- Random processes, events, probability, probability models
- Law of Large Numbers
- Addition Rule and Complement Rule

Next time

- More probability rules and an activity!

Quick note

- Even if you arrive to class late, you can fill out the attendance form by coming up after class!



Probability II

Megan Ayers

Math 141 | Spring 2026

Monday, Week 3

Announcements/Reminders

- Groupwork expectations

Last Lecture

- Introduced **probability theory**
- Saw the **Law of Large Numbers**
- Defined and practiced using key **rules of probability**

Goals for Today

- Introduce **Conditional Probability**
- Define **Independence** and the **Multiplication Rule**
- Practice using the **Law of Total Probability** and **Bayes' Rule**

Conditional Probability

Activity 1: Coffee or Tea?

A survey was given to 100 Math 141 students in 2017. Some results are summarized below:

	Coffee	Tea	total
First-year	7	10	17
Sophomore	25	20	45
Junior	13	12	25
Senior	8	5	13
total	**53**	**47**	**100**

1. What is the probability that a **random student** prefers *coffee*?
2. What is the probability that a **random student** was a *sophomore*?
3. What is the probability that a **random student** was a *sophomore* and preferred *coffee*?
4. What is the probability that a **random *sophomore*** preferred *coffee*?

Conditional Probability

Conditional Probability: probability of something occurring, **given** that another event has **already** occurred.

We write *the conditional probability of Event A given Event B has occurred* as,

$$P(A \mid B)$$

In the previous example,

- Question: What is the probability that a random *sophomore* preferred *coffee*?
- Event A: The student prefers coffee
- Event B: The student is a sophomore
- Answer: P(Coffee | Sophomore) = P(A|B)

Conditional Probability

How do we calculate conditional probabilities?

$$P(\text{Coffee}|\text{Sophomore}) = \frac{P(\text{Sophomore and prefers Coffee})}{P(\text{Sophomore})}$$

In general,

Theorem: Conditional Probability Rule

The Conditional Probability of an event A given another event B is

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Answers to Activity 1

- Event A: The student prefers coffee
- Event B: The student is a sophomore

1. **Probability student prefers coffee:** $P(\text{Coffee}) = P(A) = \frac{53}{100}$

2. **Probability student is a sophomore:** $P(\text{Sophomore}) = P(B) = \frac{45}{100}$

3. **Probability student is a sophomore and prefers coffee:**
   $P(\text{Sophomore and Coffee}) = P(\text{A and B}) = \frac{25}{100}$

4. **Probability a random sophomore prefers coffee:**

$$P(\text{A} \mid \text{B}) = \frac{P(\text{A and B})}{P(\text{B})} = \frac{25}{100} \Big/ \frac{45}{100} = \frac{25}{45}$$

Using Conditional Probability

Using Conditional Probabilities...

In the next few slides, we're going to use the concept of conditional probability to explore 4 more concepts:

1. Independence
2. Multiplication Rule
3. Law of Total Probability
4. Bayes' Rule

Independence

We say two events are **independent** if knowing one occurs doesn't change the probability that another occurs.

- For example, rolling a 3 on a die (*event B*) doesn't change the probability that it rains today (*event A*), so these two events are independent.

- Learning that an even number has been rolled on a die (*event B*) does change the probability that a 4 was rolled (*event A*), so these two events are not independent.

Theorem: Criteria for Independence

Two events A and B are independent if and only if

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B)$$

Q: Are the events "preferring coffee" (A) and "being a sophomore at Reed" (B) independent?

No! That's because $P(A) = \frac{53}{100}$ and $P(A|B) = \frac{25}{45}$, so $P(A) \neq P(A|B)$

Multiplication Rule

What if we want to know the probability that two events **both** occur?

Rearrange our formula for conditional probability!

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$\implies P(A|B)P(B) = P(A \text{ and } B)$$

Theorem: Multiplication Rule

For any events A and B,

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

Multiplication Rule

Theorem: Multiplication Rule

For any events A and B,

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

In the previous example…

- Event A: The student prefers coffee
- Event B: The student is a sophomore

and

- $P(B) = \frac{45}{100}$
- $P(A|B) = \frac{25}{45}$
- $P(A \text{ and } B) = \frac{25}{100}$

Does it work?

$$P(A|B)P(B) = \left(\frac{25}{45}\right)\left(\frac{45}{100}\right) = \frac{25}{100}$$

Checks out!

The Law of Total Probability

Consider two events:

- Event A: The student prefers coffee
- Event B: The student is a sophomore

In this scenario, notice for any student who prefers coffee (A), it's possible…

- they're a sophomore (B)
- they're **not** a sophomore (B^c)

Thus,

$$\begin{aligned} P(A) &= P(A \text{ and } B) + P(A \text{ and } B^c) && \text{(Addition Rule)} \\ &= P(A|B)P(B) + P(A|B^c)P(B^c) && \text{(Multiplication Rule)} \end{aligned}$$

The Law of Total Probability

Theorem: The Law of Total Probability

Let A and B be events. Then

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

- The Law of Total Probability is useful for finding "overall" probabilities when we only have the "pieces".
- We'll see a real-world example next!

Patient Health

Suppose you're a medical researcher interested in population health.

- 80% of individuals *see a physician* each year
- 20% do not

- Among people who see a physician, 95% are generally healthy.
- Among people who *don't see* a physician, 70% are generally healthy.

Q: What proportion of adults are generally healthy?

Let

- Event A: Individual is healthy
- Event B: Individual sees a physician each year
- Event: B^c: Individual doesn't see a physician each year

This question is asking: $P(A) =$**???**

Patient Health

Theorem: The Law of Total Probability

Let A and B be events. Then

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

We want to answer $P(\text{Healthy})$.

Using the Law of Total Probability:

$$\begin{aligned} P(\text{Healthy}) &= P(\text{Healthy}|\text{Doctor})P(\text{Doctor}) + P(\text{Healthy}|\text{Doctor}^c)P(\text{Doctor}^c) \\ &= P(A|B)P(B) + P(A|B^c)P(B^c) \\ &= (.95)(.80) + (.70)(.20) \\ &= 0.90 \end{aligned}$$

Thus, 90% of adults are generally healthy!

Bayes' Rule

What if we know $P(B|A)$, but we want to know $P(A|B)$?

Theorem: Bayes' Rule

Let A and B be events. Then,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Proof: From the definition of multiplication rule,

$$P(A \text{ and } B) = P(A|B) \times P(B)$$
$$P(A \text{ and } B) = P(B|A) \times P(A)$$

Activity

Activity 2: Testing for a Rare Disease

Suppose that we have a rapid COVID-19 test where:

- Given a person **does not** have COVID, the test is (correctly) negative 99% of the time.
- Given a person **does** have COVID, the test is (correctly) positive 80% of the time.

Assume that the overall prevalence of COVID at the time of the test was 1%.

Q: Suppose a person takes this test and receives a positive diagnosis – **what is the probability that the person has COVID?**

- **Hint 1**: Write out the object in question in terms of a conditional probability
- **Hint 2**: Define the relevant events and their complements in this setting
- **Hint 3**: Write out all the info given in terms of probabilities and events A and B.
- **Hint 4**: Start with Bayes' Rule! You'll also need the Law of Total Probability!

Activity 2: Solution

Q: Suppose a person takes this test and receives a positive diagnosis – what is the probability that the person has COVID?

- This is asking for $P(\text{Have Covid} \mid \text{Positive Test})$
- C = Has COVID, C^c = No COVID, $+$ = Positive Test, $+^c$ = Negative Test
- Express what we know using probability language…

Activity 2: Solution

- Test correctly diagnoses a person who **does not** have COVID 99% of the time
  $P(+^c|C^c) = 0.99$ and $P(+|C^c) = 0.01$
- Test correctly diagnoses a patient who **does** have COVID 80% of the time.
  $P(+|C) = 0.80$ and $P(+^c|C) = 0.20$
- The overall prevalence of COVID at the time of the test was 1%.
  $P(C) = 0.01$ and $P(C^c) = 0.99$

Bayes' Rule:

$$P(C|+) = \frac{P(+|C)P(C)}{P(+)} = \frac{0.80 * 0.01}{P(+) = ???}$$

Law of Total Probability:

$$P(+) = P(+|C)P(C) + P(+|C^c)P(C^c) = 0.80(0.01) + 0.01(0.99) = 0.0179$$

Combining the above:

$$\implies P(C|+) = \frac{0.80 * 0.01}{0.0179} \approx \boxed{0.447}$$

Bonus Activity: Bayes' Rule

Theorem: Bayes' Rule

Let A and B be events. Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Q: Consider Bayes' Rule. Under what circumstances will $P(A|B) = P(B|A)$?

Q: Suppose $P(B|A) = 1$:

- What does this suggest about A and B?
- What is $P(A|B)$ in this case?

Bonus Activity: Answers

Theorem: Bayes' Rule

Let A and B be events. Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Q: Under what circumstances will $P(A|B) = P(B|A)$?

Answer: Whenever $P(A) = P(B)$

Q: Suppose $P(B|A) = 1$:

- What does this suggest about A and B?

A: Whenever A occurs, so does B.

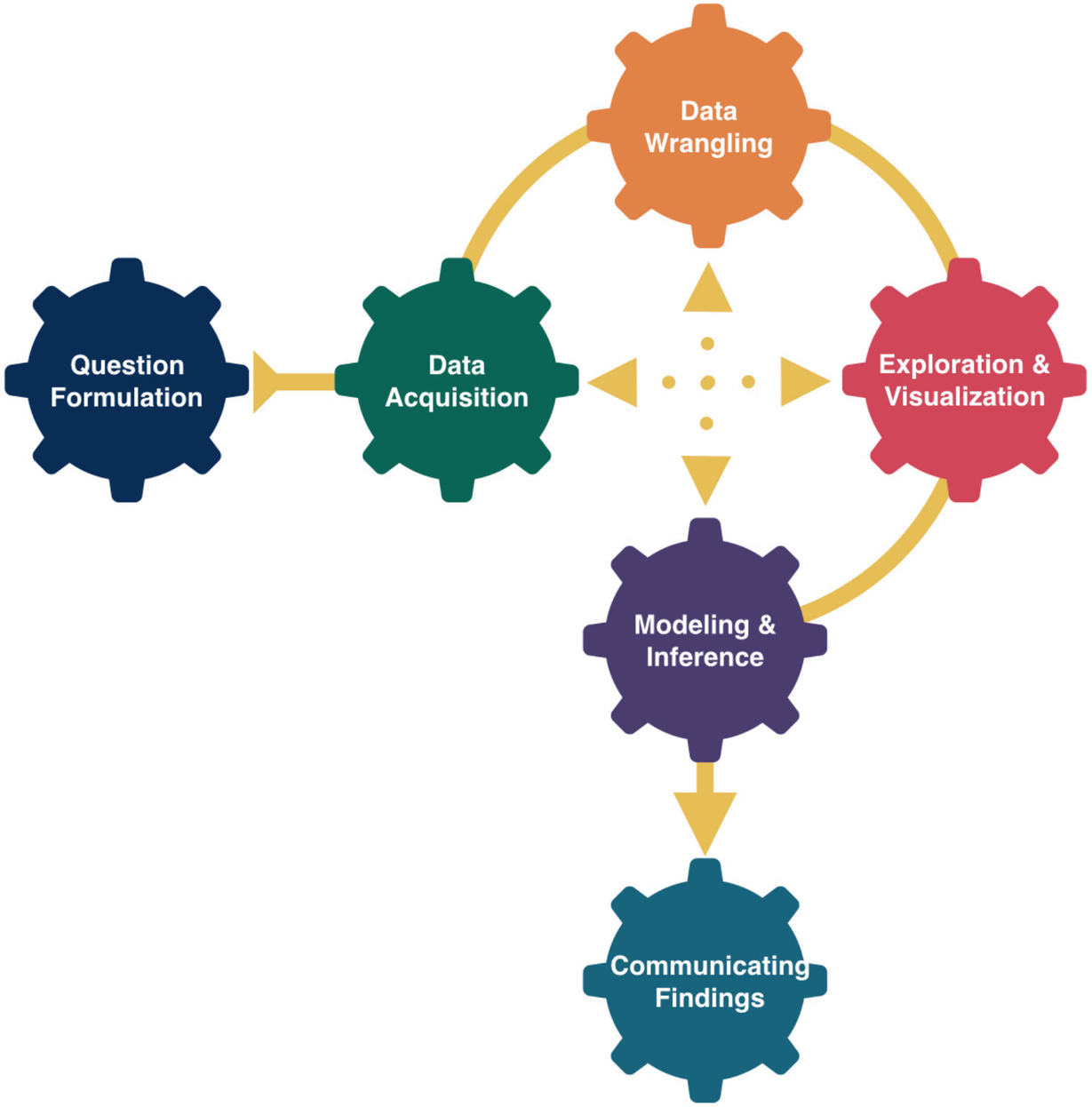
- What is $P(A|B)$ in this case?

A: $P(A|B) = P(A)/P(B)$

Story time: Independence and "Lumpy" Randomness

A professor from the Georgia Institute of Technology asked students on their homework assignment to either flip a coin 200 times and record the results, or pretend to flip a coin and fake 200 results. After HW submission, students were amazed that he could correctly identify whether everyone's results were real or fake…

- Was he omniscient?
- Nope! His strategy was simply to look for runs of at least six heads or six tails in a row. **Why did this work?**
  - Independence and a large number of trials means a run of six is very likely from real coin flips.
  - Humans writing pretend coin flips are influenced by previously written outcomes.
- This idea has connections to many topics, ranging from video games to fraud detection!



Data Collection and Sampling

Announcements/Reminders

HEDS Sense of Community Survey

Open through Friday, February 13

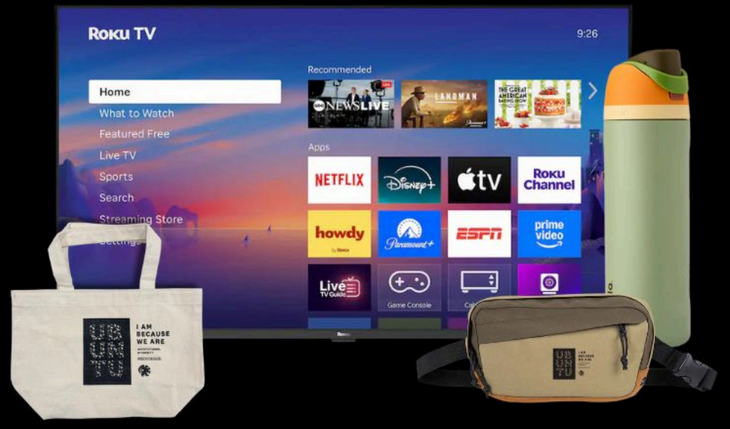
This 10-minute survey will help us learn about your experiences of belonging and connection at Reed.

Check your email for a unique link to take the survey

Participants have a chance to win daily, weekly and grand prizes.

Grand prize

One survey participant will win a **private three-course dinner** at the **Parker House** for themselves & 5 Reed community member



Goals for Today

- Discuss principles of data collection/acquisition
- Investigate 3 methods of drawing random samples

When to Get Coding Help

😢 *“I have no idea how to do this problem.”*

→ Ask someone to point you to a similar example from the lecture or readings.

→ Talk it through with a course assistant, a fellow Math 141 student, or Megan so together we can verbalize the process of going from Q to A.

😡 *“I am getting a weird error but really think my code is correct/on the right track/matches the examples from class.”*

→ It is time for a second pair of eyes. Don’t stare at the error for over 10 minutes.

🤩 And lots of other times too! 😬

When to Get Help

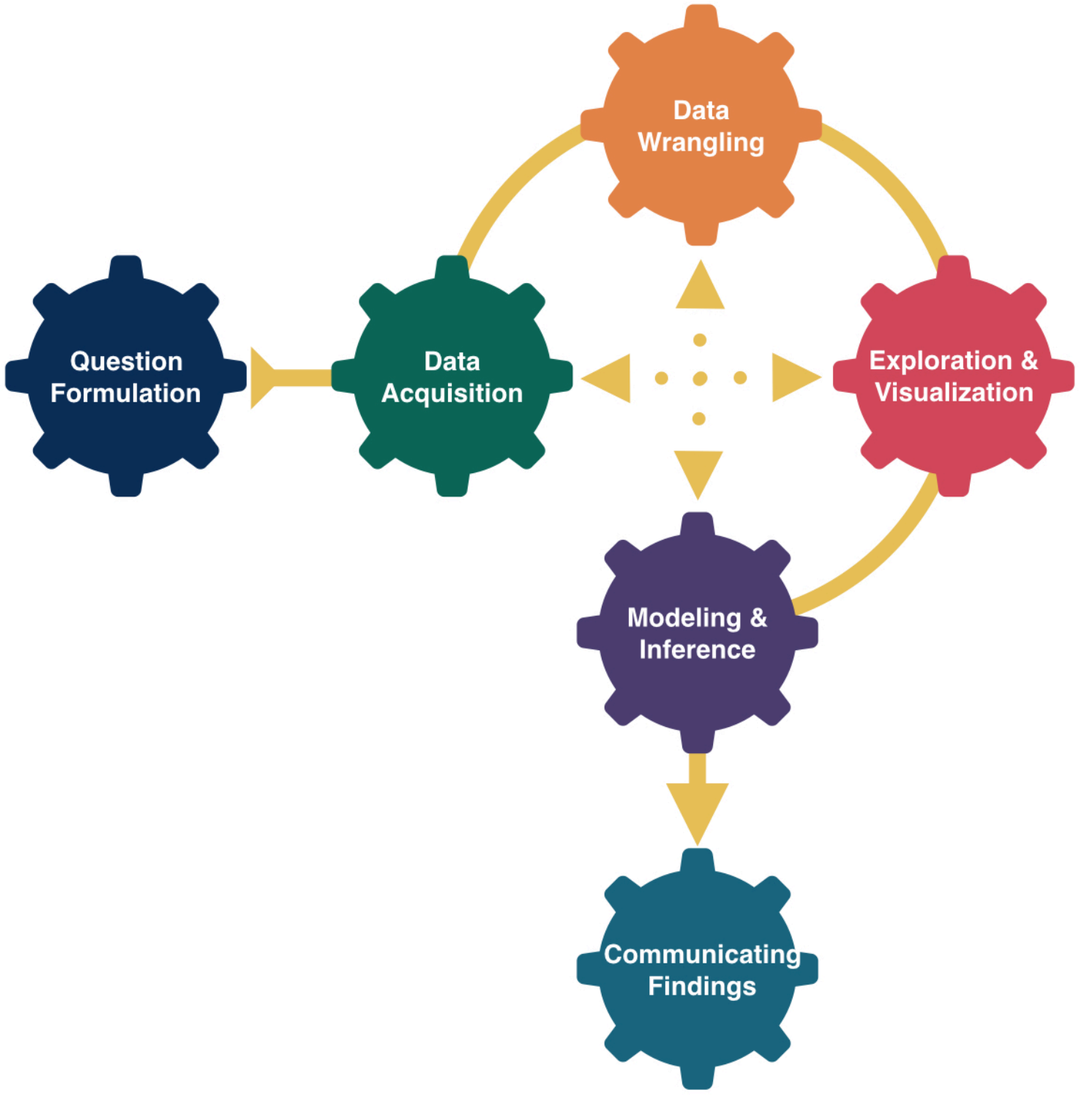
Remember:

→ Struggling is part of learning.

→ But let us help you ensure it is a **productive** struggle.

→ Struggling does NOT mean you are bad at stats, it actually means you are doing the work to **learn** the material!

Now for Data Collection


Data Wrangling
Question Formulation
Data Acquisition
Exploration & Visualization
Modeling & Inference
Communicating Findings

Who are the data supposed to represent?

Every statistical investigation should clearly identify and compare:

- The **population** to be studied
- The **sample** from which measurements (data) will be taken

Key questions:

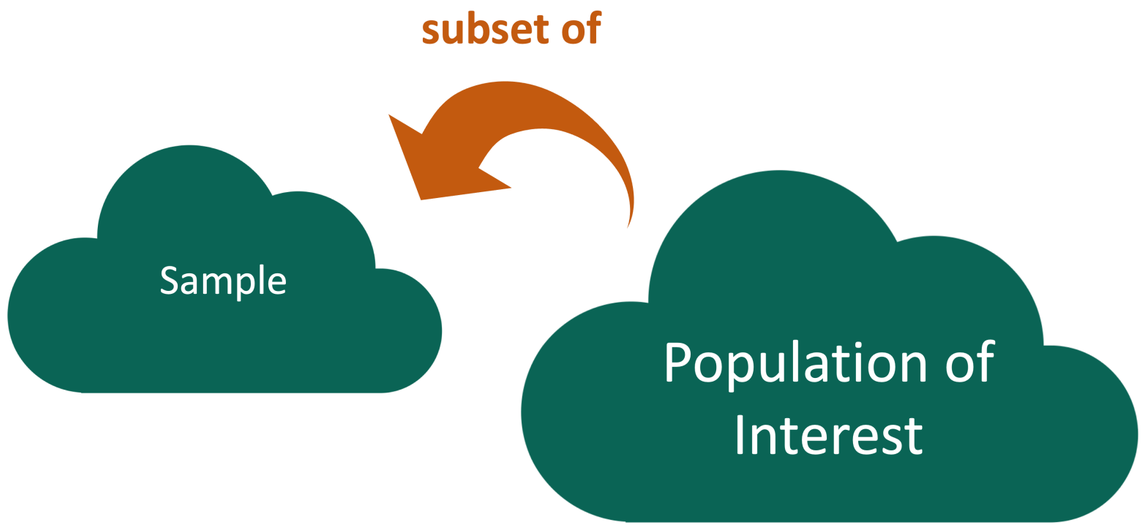
- What evidence is there that the sample is **representative** of the population?
- Who is present? Who is absent?
- Who is overrepresented? Who is underrepresented?

Who are the data supposed to represent?

In a **census**, we have data on the entire population!

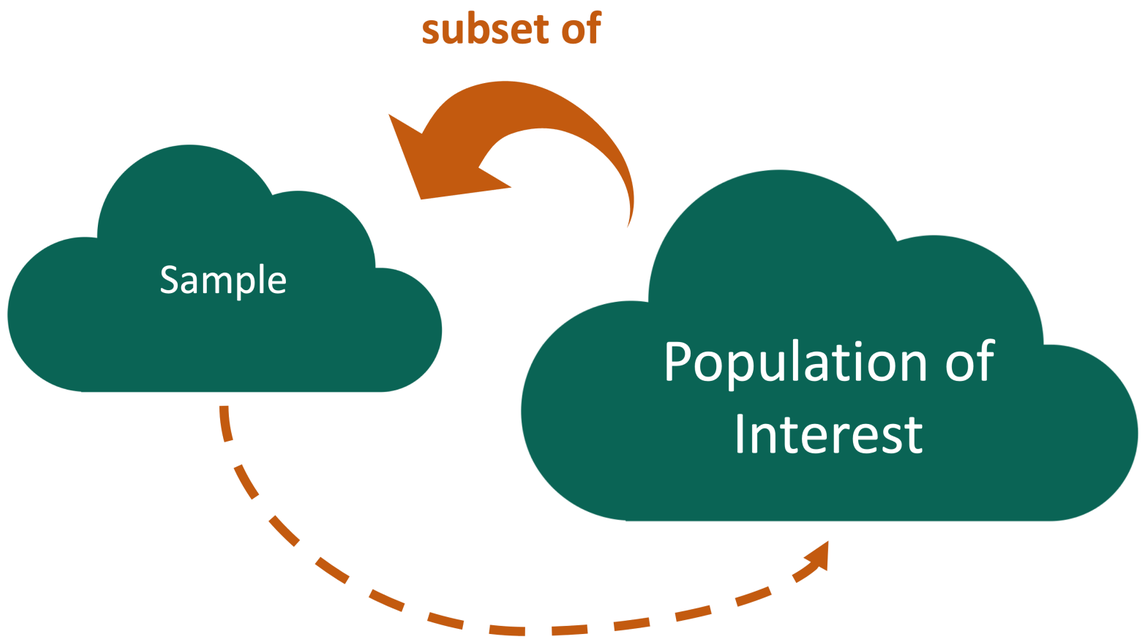
But usually, we don't have the money, time, or ability to do this.

Who are the data supposed to represent?



Instead, we use a **sample** of the population, and use the sample to draw conclusions about the **population**.

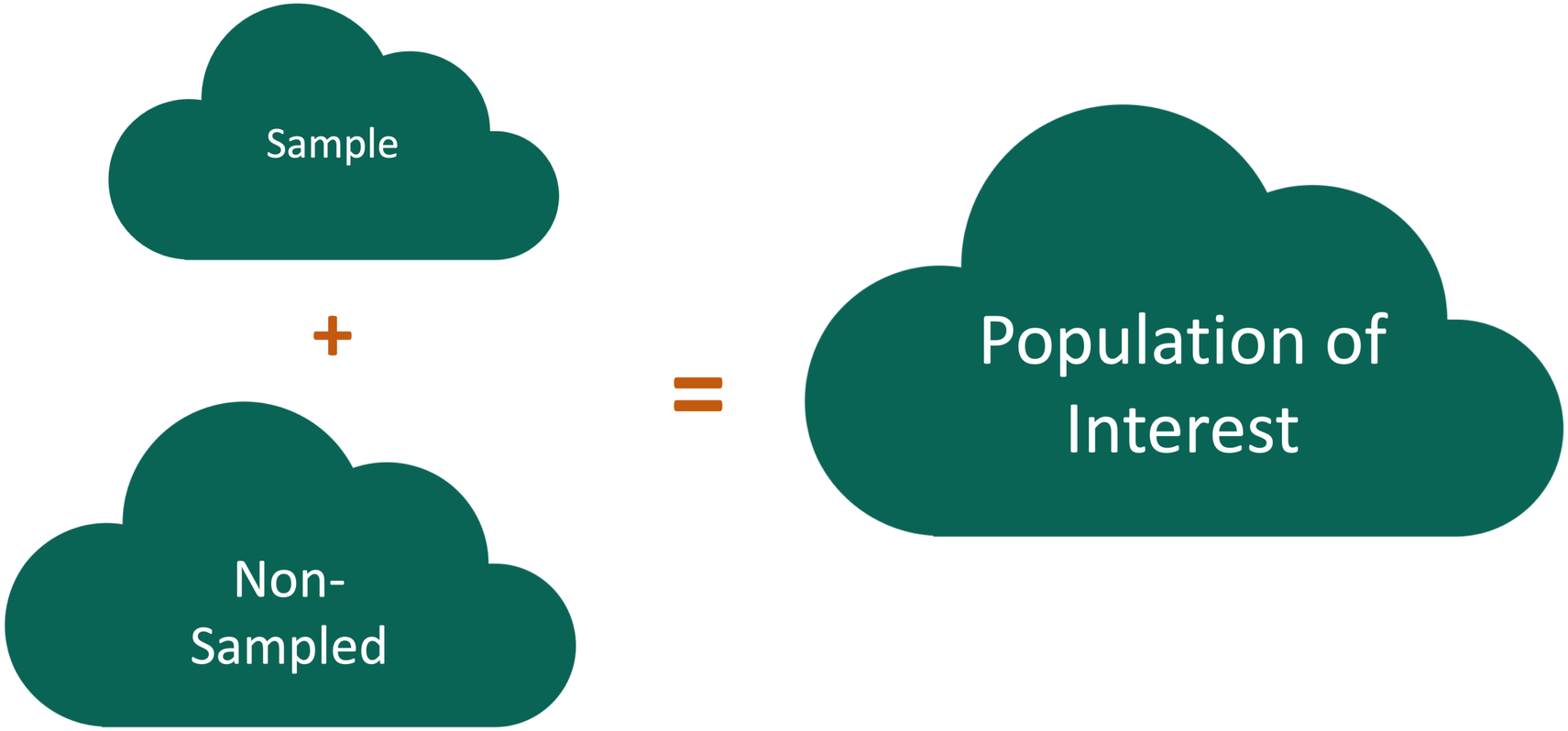
Who are the data supposed to represent?



Key questions:

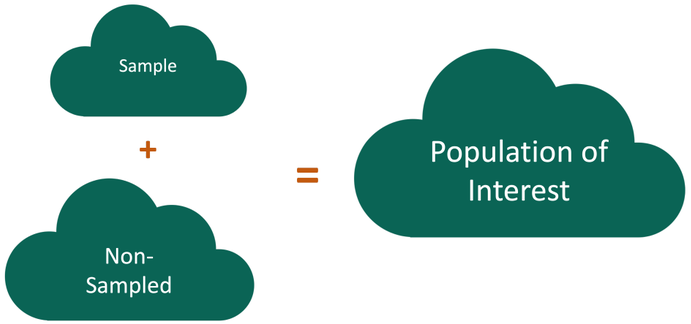
- What evidence is there that the **sample** is **representative** of the **population**?
- Who is present? Who is absent?
- Who is overrepresented? Who is underrepresented?

Sampling Bias



Sampling bias: When certain individuals are more likely to be sampled than others

Sampling Bias



Sampling bias: When certain individuals are more likely to be sampled than others

Q: Consider a telephone poll for an election - where might we get sampling bias?

- Non-response: individual can't or won't contribute
- Undercoverage: some groups are less likely to be called
- Inaccurate response
- Self-selection: membership in the sample is voluntary
- Convenience: selecting a convenient but non-representative block to sample

Sampling Bias Example

The **Literary Digest** was a political magazine that correctly predicted the presidential outcomes from 1916 to 1932. In 1936, they conducted the most extensive (to that date) public opinion poll. They mailed questionnaires to over 10 million people (about 1/3 of US households) whose names and addresses they obtained from telephone books and vehicle registration lists.

Population of Interest:

Sample:

Sampling bias:

Sampling Bias Example

We want to know how Portlanders feel about a new coffee shop in Woodstock.

- The coffee shop has a Yelp rating of 3.5/5 stars with 10 reviews.

Q1: Can we conclude that a typical Portlander would rate this coffee shop at 3.5 stars?

Q2: What sources of bias are present in this sample?

Q3: A year later, the coffee shop still has 3.5 stars, but 1000 reviews. Does the verdict change?

Q4: A second coffee shop opens up nearby with a Yelp rating of 4 stars and 1000 reviews. Can we conclude Portlanders prefer the second restaurant to the first?

Random Sampling

Use random sampling (a random mechanism for selecting cases from the population) to remove sampling bias.

Types of random sampling

We'll explore 3 types of random sampling

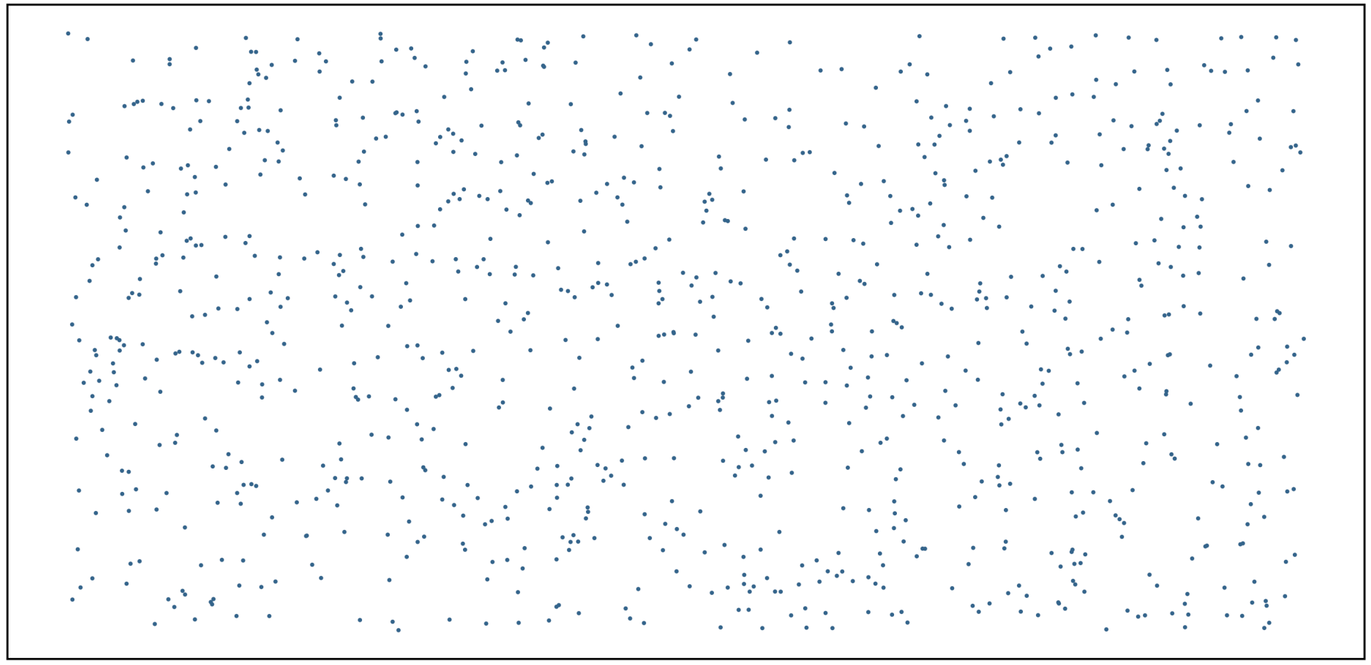
- Simple random sampling
- Cluster sampling
- Stratified random sampling

Simple Random Sampling

Motivating question: what is the average amount of student loan debt in Oregon?

Simple Random Sampling: Imagine that a unique ID for each student *in the population* is written on a slip of paper…

- Shuffle the slips of paper in a bowl
- Draw n IDs/slips one-by-one to create a sample

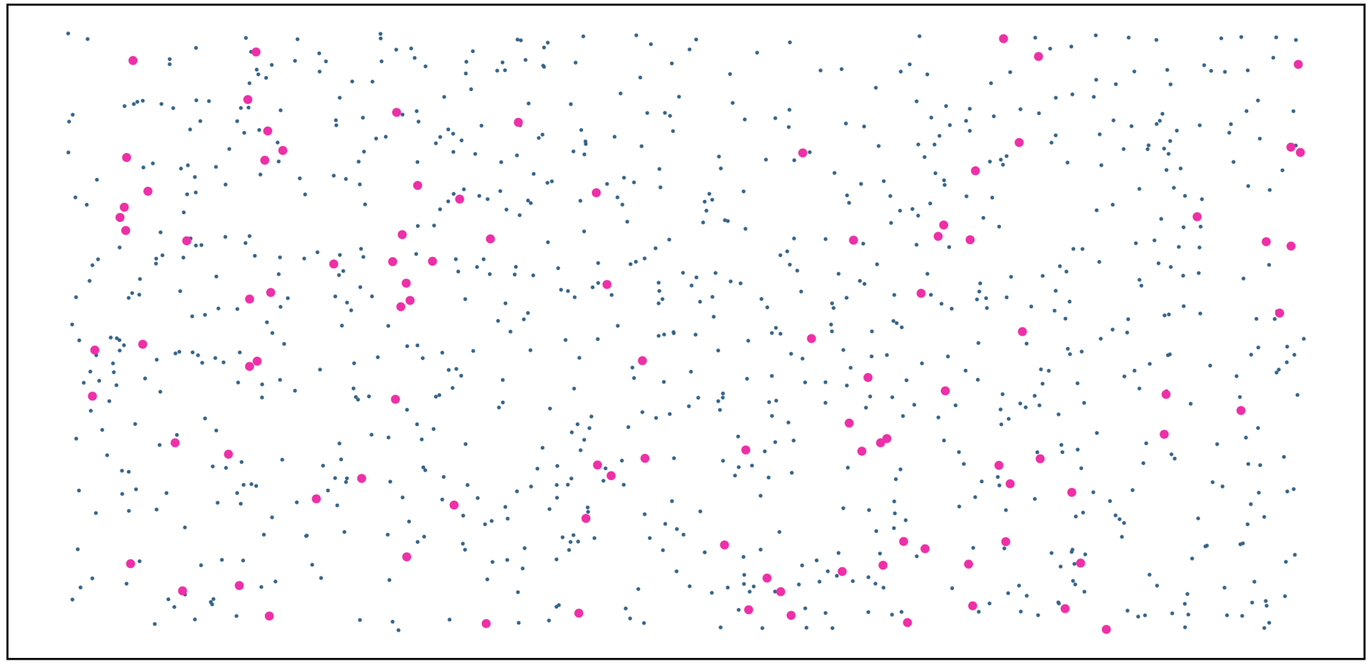


Simple Random Sampling

Motivating question: what is the average amount of student loan debt in Oregon?

Simple Random Sampling: Imagine that a unique ID for each student *in the population* is written on a slip of paper…

- Shuffle the slips of paper in a bowl
- Draw n IDs/slips one-by-one to create a sample



Simple Random Sampling

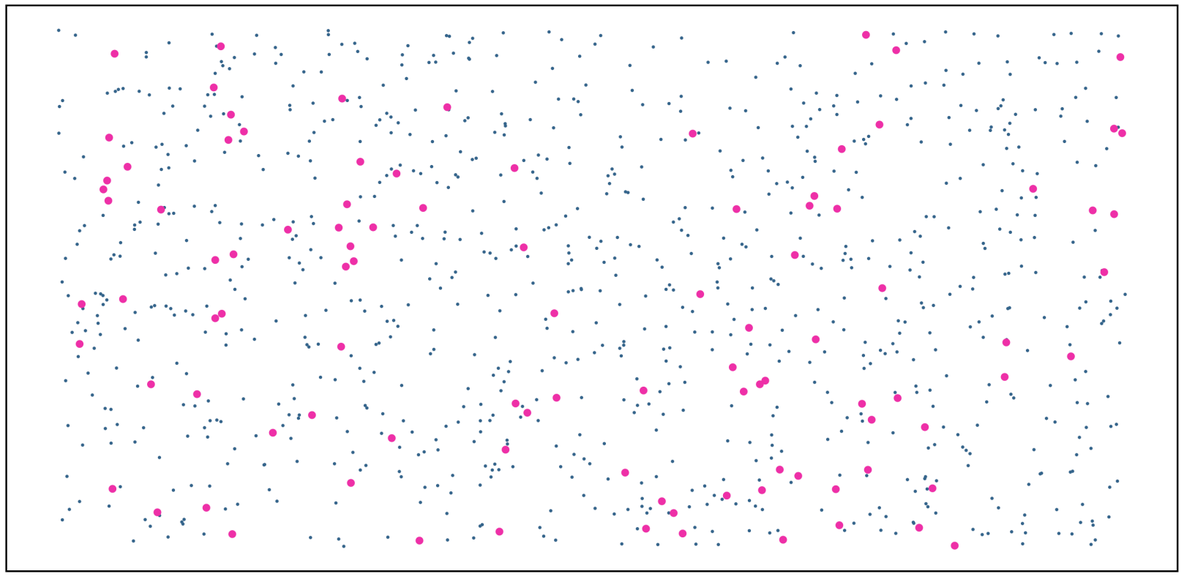
Consequences:

- Every member of the population has an **equal chance of being selected** for the sample
- There is **no inherent correlation** between any two members of the sample

Q: Can a simple random sample be non-representative?

A: Yes, even if all goes as planned!

- For large sample sizes, it's unlikely
- The sample will be **representative on average**



Q: Why aren't all samples generated using simple random sampling?

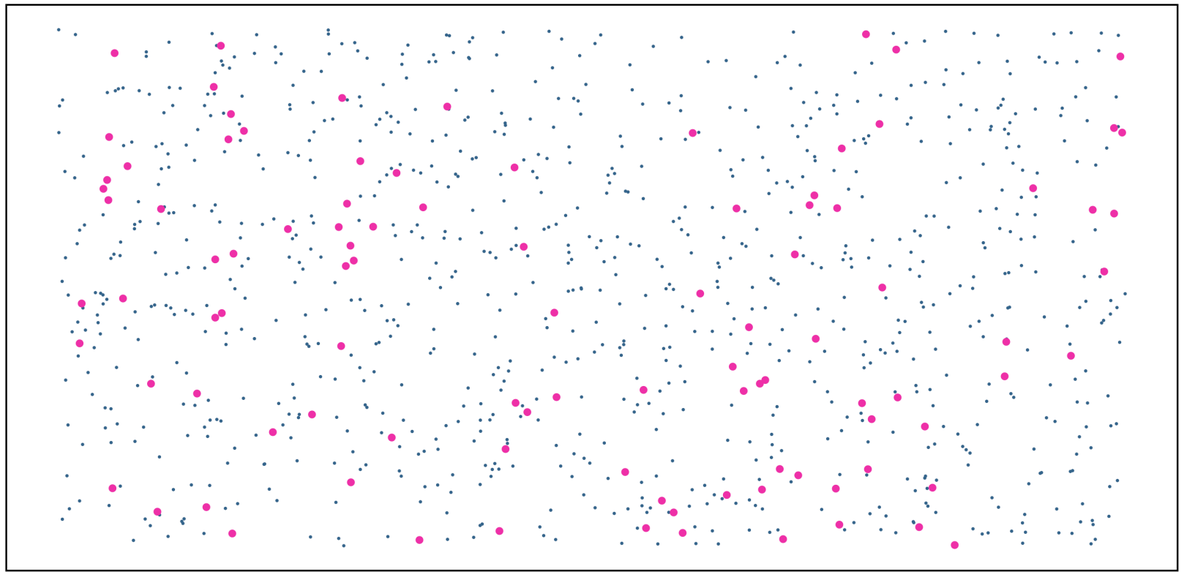
Simple Random Sampling

Advantages:

- Relatively simple to interpret and analyze
- Non-biased (in theory)

Disadvantages:

- May not be as “precise” as other sampling techniques
- Can be difficult to perform in practice

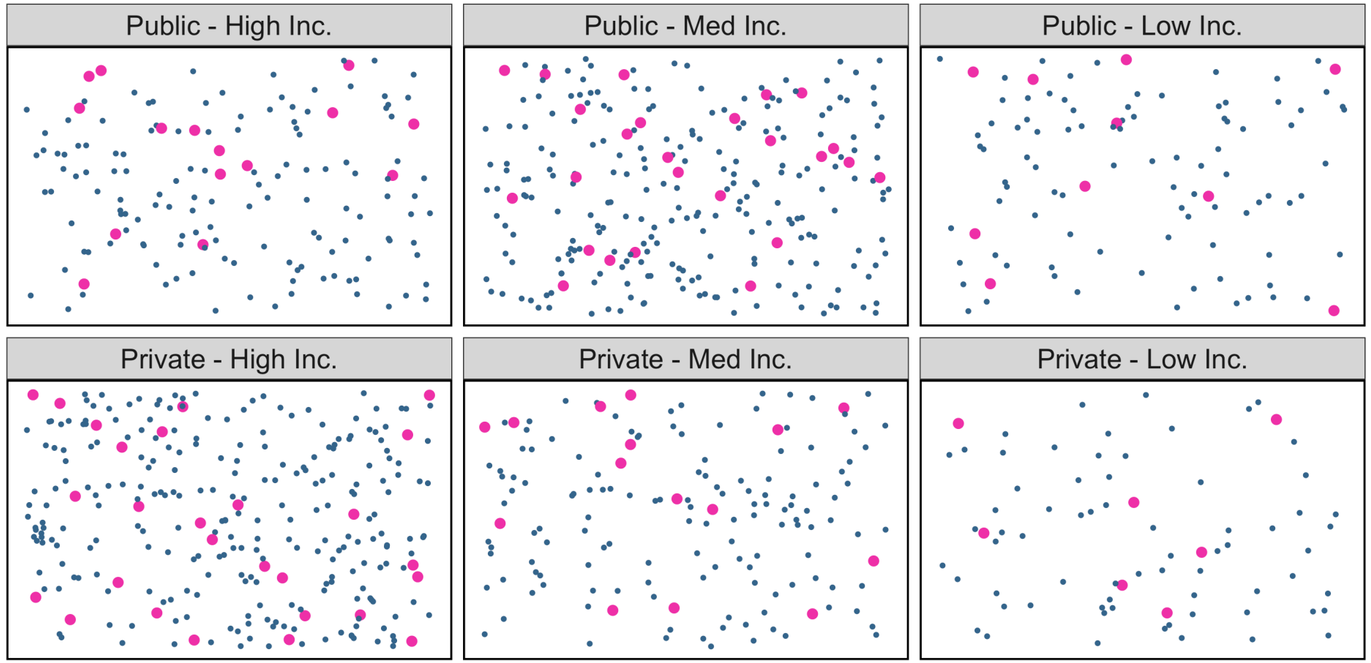


Stratified Random Sampling

Motivating question: what is the average amount of student loan debt in Oregon?

Stratified Random Sampling: "Strata" are made up of similar individuals, then simple random samples are taken from each stratum.

- e.g. define strata based on public vs private college and family income ranges



Stratified Random Sampling

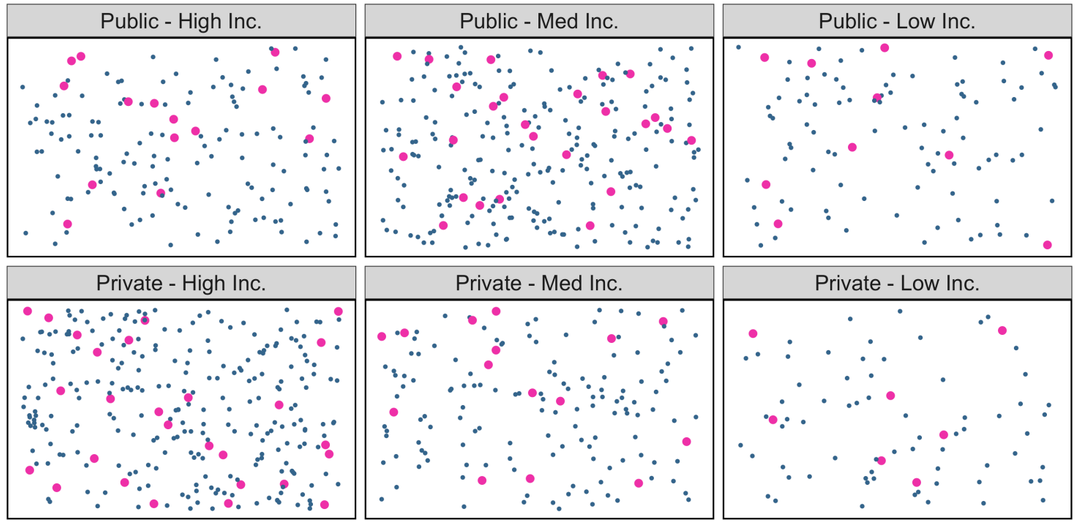
Stratified Random Sampling: "Strata" are made up of similar individuals, then simple random samples are taken from each stratum.

Advantages:

- Can be more "precise" than simple random sampling, requiring lower sample size
- Hedges against non-representative samples

Disadvantages:

- Statistical analysis is more complex
- Strata creation isn't always straightforward (need additional data, and to have compelling reasons for strata definitions)

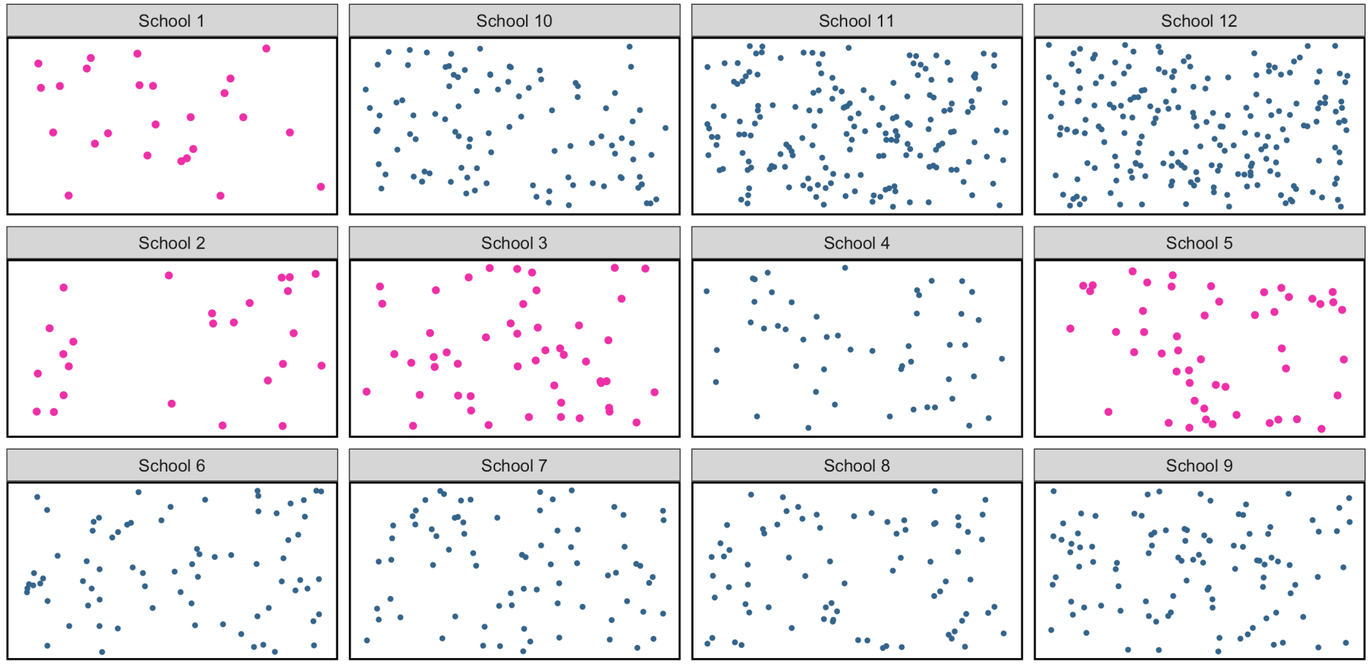


Cluster Random Sampling

Motivating question: what is the average amount of student loan debt in Oregon?

Cluster Random Sampling: "Clusters" are non-homogeneous. We take a simple random sample of the clusters, and use all observations in those clusters as the sample.

- e.g. we take a simple random sample of schools, include all students in those schools in the sample



Cluster Random Sampling

Cluster Random Sampling: “Clusters” are non-homogeneous. We take a simple random sample of the clusters, and use all observations in those clusters as the sample.

National Health and Nutrition Examination Survey

Mission: "Assess the health and nutritional status of adults and children in the United States."

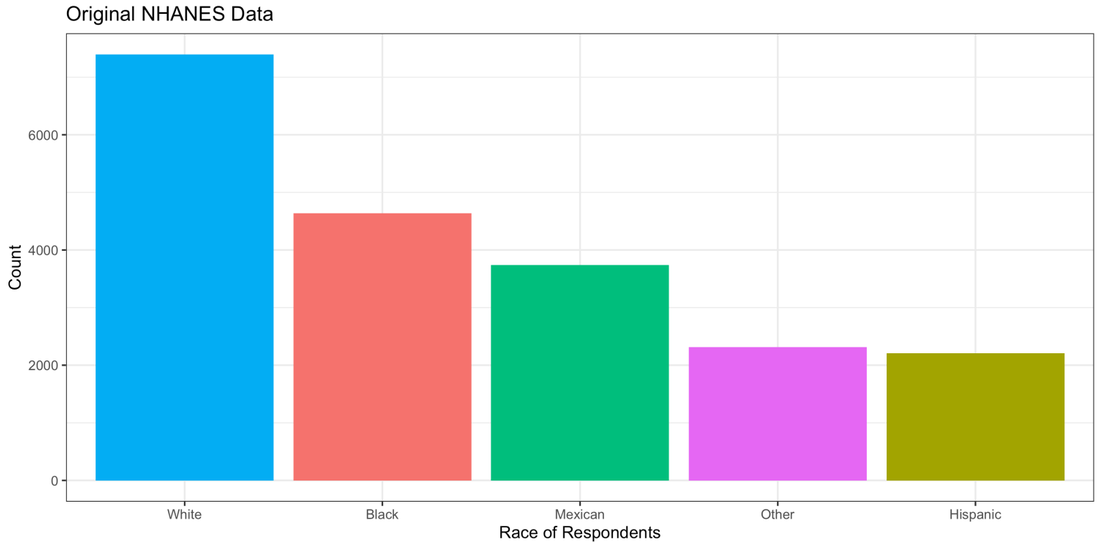
How are these data collected?

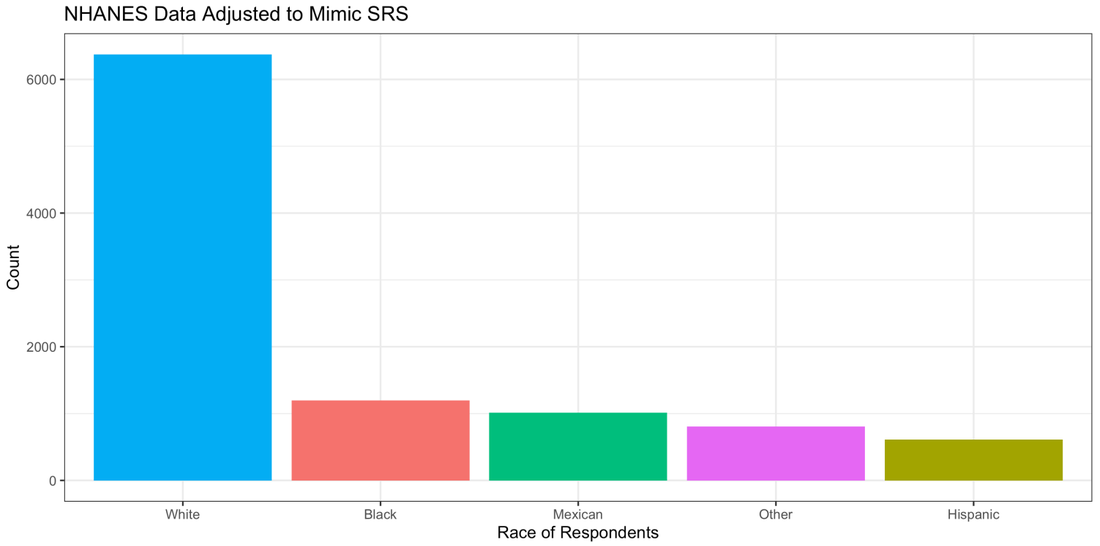
NHANES Sampling Design

- **Stage 1**: US is stratified by geography and distribution of minority populations. Counties are randomly selected within each stratum.
- **Stage 2**: From the sampled counties, city blocks are randomly selected. (City blocks are clusters.)
- **Stage 3**: From sampled city blocks, households are randomly selected. (Households are clusters.)
- **Stage 4**: From sampled households, people are randomly selected. For the sampled households, a mobile health vehicle goes to the house and medical professionals take the necessary measurements.

Why don’t they use simple random sampling?

Careful Using Non-Simple Random Sample Data





Detour: Data Ethics

Data Ethics

> "Good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations." – Committee on Professional Ethics of the American Statistical Association (ASA)

The ASA has created "Ethical Guidelines for Statistical Practice"

→ These guidelines are for EVERYONE doing statistical work.

→ There are ethical decisions at all steps of the Data Analysis Process.

→ We will periodically refer to specific guidelines throughout this class.

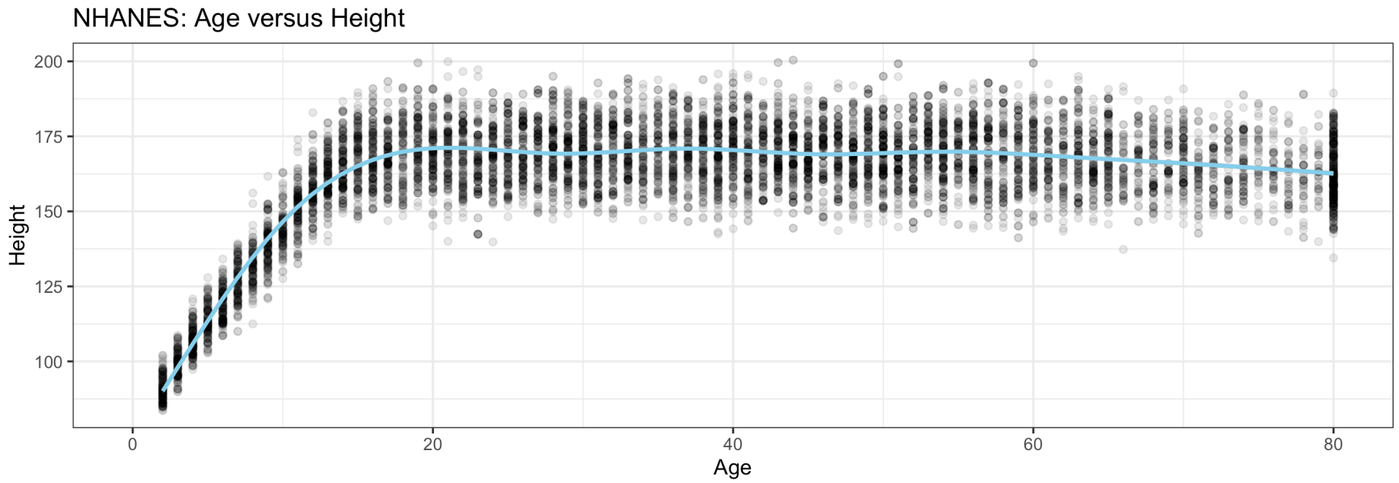
> "Above all, professionalism in statistical practice presumes the goal of advancing knowledge while avoiding harm; using statistics in pursuit of unethical ends is inherently unethical."

Responsibilities to Research Subjects

> “The ethical statistician protects and respects the rights and interests of human and animal subjects at all stages of their involvement in a project. This includes respondents to the census or to surveys, those whose data are contained in administrative records, and subjects of physically or psychologically invasive research.”

Responsibilities to Research Subjects

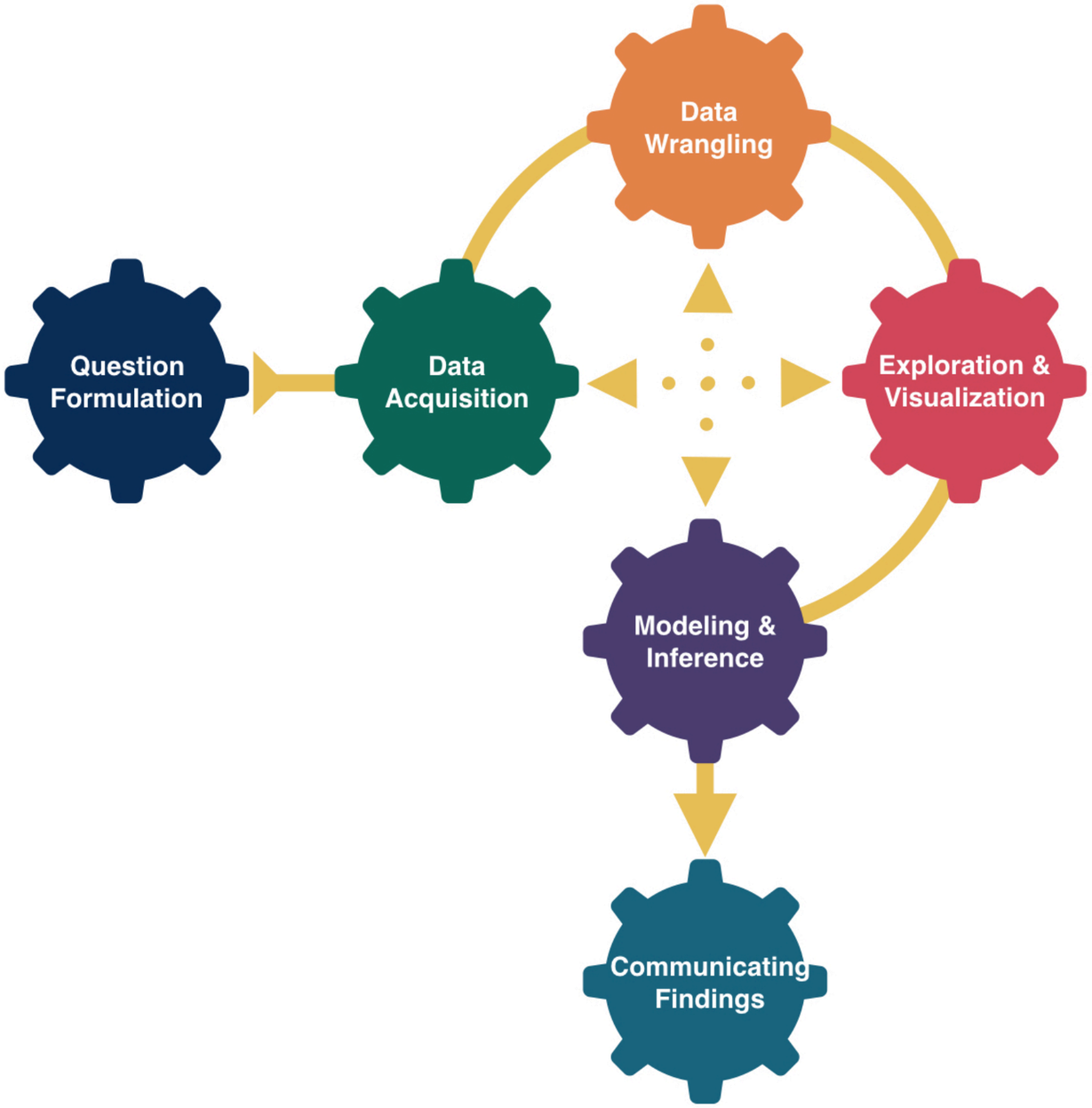
Why do you think the Age variable maxes out at 80?



> "Protects the privacy and confidentiality of research subjects and data concerning them, whether obtained from the subjects directly, other persons, or existing records."

Next time:

- Study design



Study Design

Megan Ayers

Math 141 | Spring 2026

Friday, Week 3

Reminders/Announcements

- Feedback for Lab 01 posted on Gradescope
- Feedback for HW 01 coming soon

Math/Stats course interest form

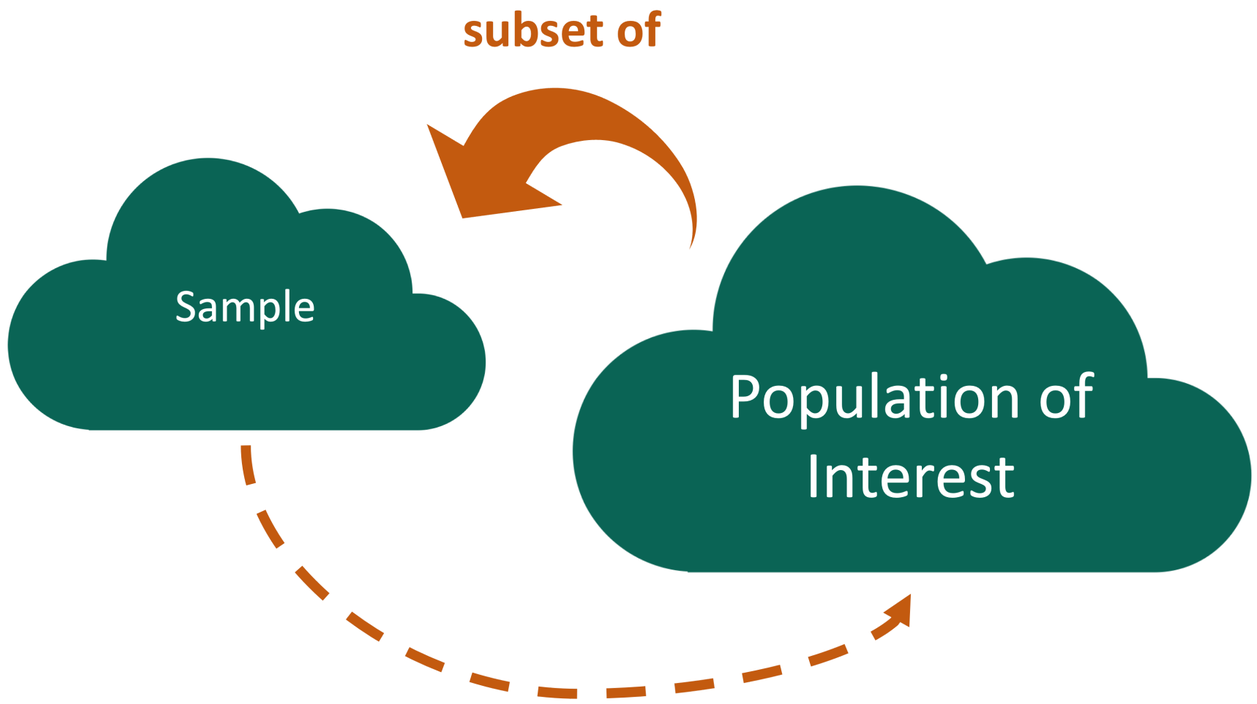
> The Mathematics and Statistics Department is currently in the process of drafting a schedule for next academic year. In order to have a better sense of how many sections of each class to offer, we would like to know your plans for next year.

Goals for Today

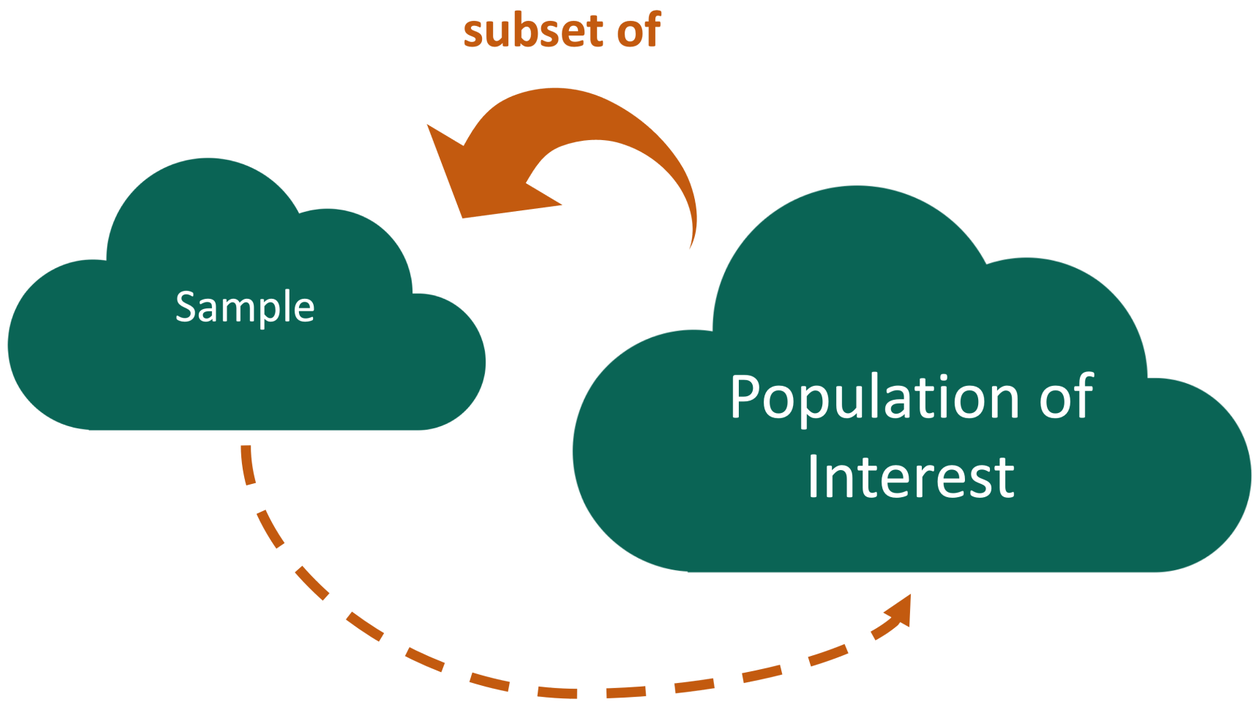
- Recap sampling bias
- Discuss drawing conclusions from our sample and types of studies

Data Collection

Who are the data supposed to represent?



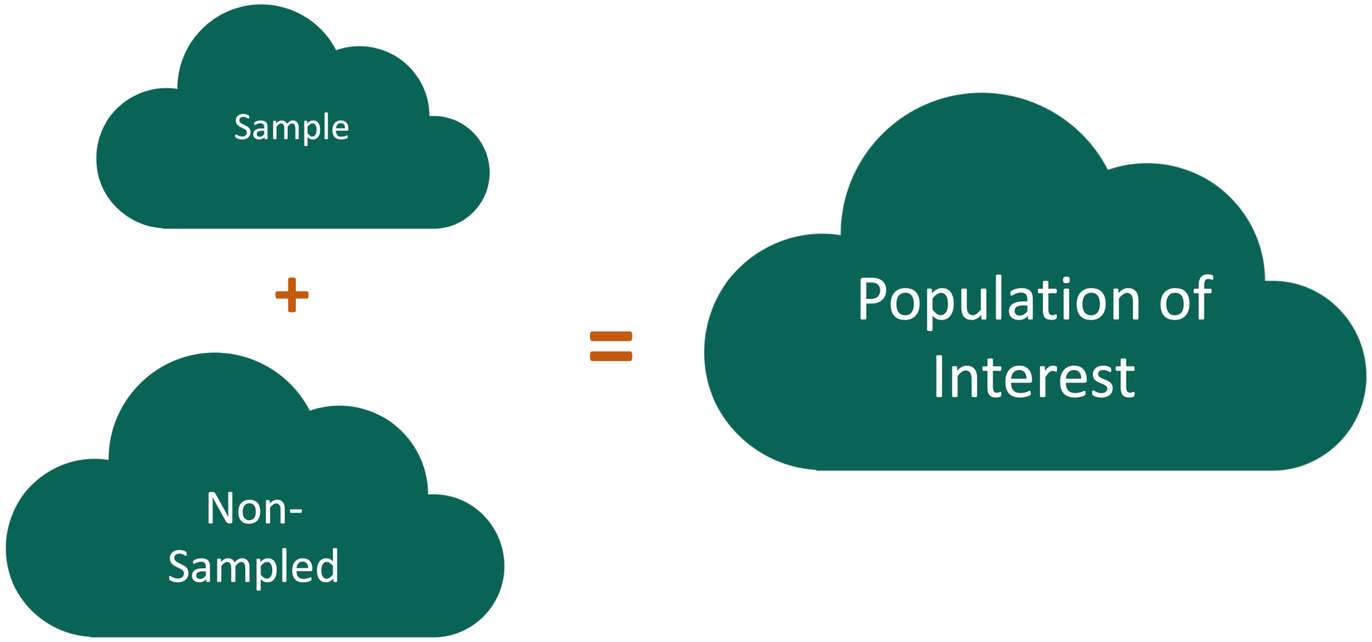
Who are the data supposed to represent?



Key questions:

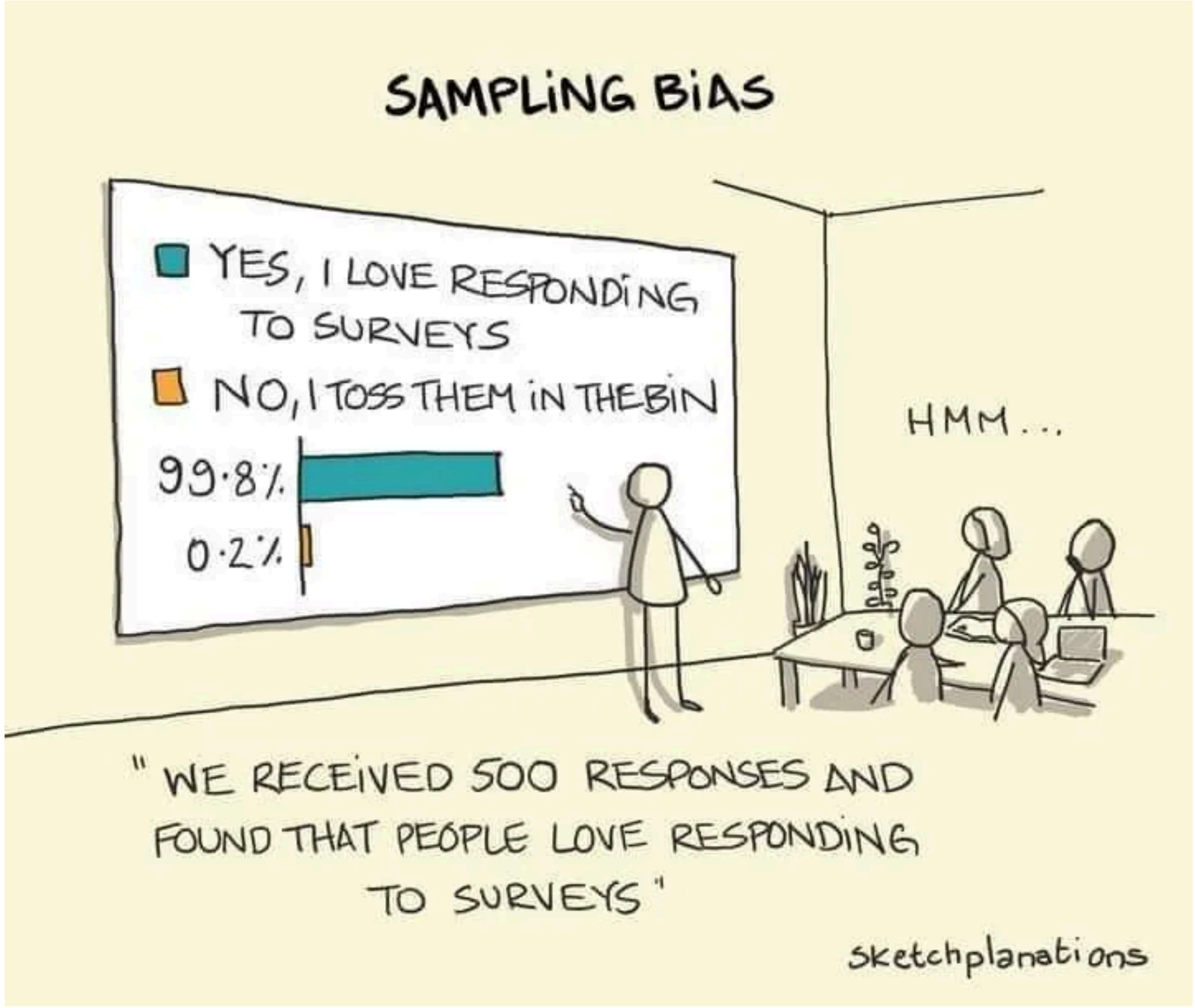
- What evidence is there that the **respondents** are **representative** of the **population**?
- Who is present? Who is absent?
- Who is overrepresented? Who is underrepresented?

Nonresponse bias



Nonresponse bias: The respondents are **systematically** different from the non-respondents for the variables of interest.

Nonresponse bias



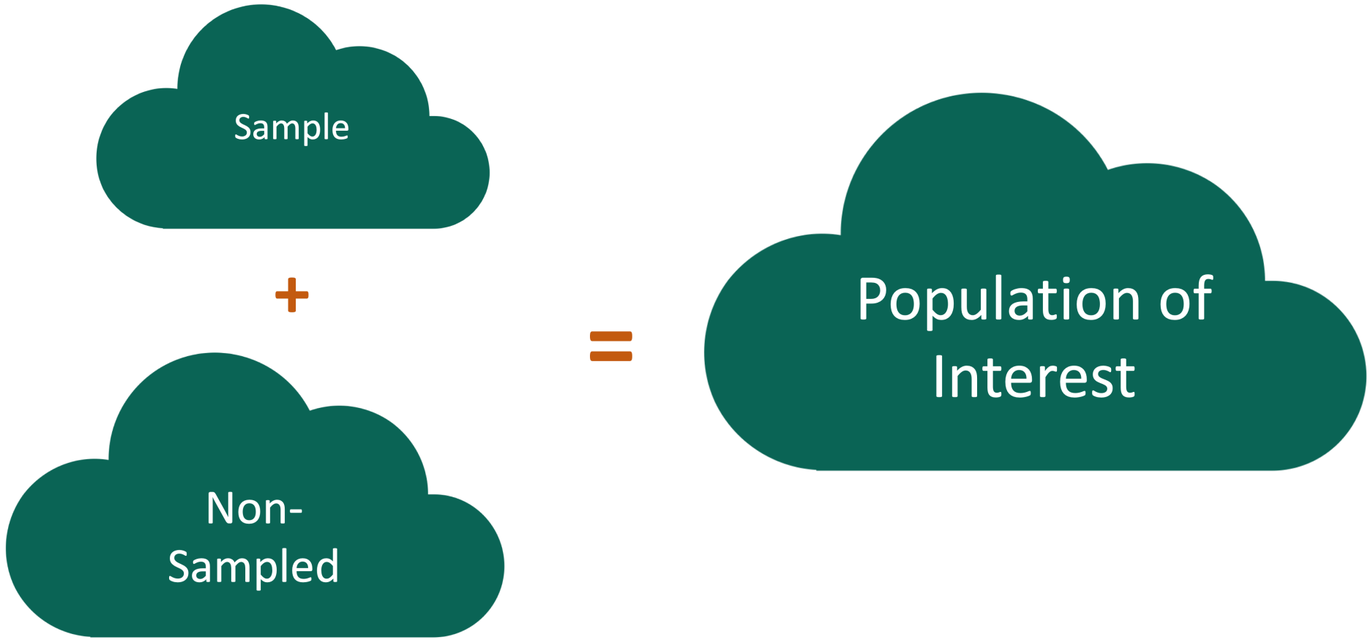
Nonresponse bias: The respondents are **systematically** different from the non-respondents for the variables of interest.

Come Back to Literary Digest Example

Of the 10 million people surveyed, more than 2.4 million responded with 57% indicating that they would vote for Republican Alf Landon in the upcoming presidential election instead of the current President Franklin Delano Roosevelt.

Non-response bias? Sample creation issues?

Tackling Nonresponse bias



- Use **multiple modes** (mail, phone, in-person) and **multiple attempts** for reaching sampled cases.
- Explore key demographic variables to see how respondents and non-respondents vary.
- In survey statistics, we can create **survey weights** to adjust for potential nonresponse bias.

Is Bigger Always Better?

For our **Literary Digest Example**, Gallup predicted Roosevelt would win based on a survey of **50,000** people (instead of 2.4 million).

Quality over quantity!

Thoughts on Sampling

- **Random** sampling is important to ensure the sample is **representative** of the population.
  - Word we will use: **generalizability**
- Representativeness isn't about **size**.
  - Small random samples will tend to be more representative than large non-random samples.
- However, I bet most samples you will encounter **won't** have arisen from a random mechanism.
- How do we draw conclusions about the population from **non-random samples**?
  - Determine if your sampled cases (and respondents) are systematically different from the non-sampled cases (and non-respondents) for the variables you care about.
  - Adjust your population of interest.

Now let's shift our discussion to the conclusions we can draw from the sample we have.

Typical Analysis Goals

Descriptive: Want to estimate quantities related to the population.

→ *How many trees are in the Amazon?*

Predictive: Want to predict the value of a variable.

→ *Can I use remotely sensed data to predict forest types in the Amazon?*

Causal: Want to determine if changes in a variable cause changes in another variable.

→ *Do financial contracts prevent people from deforesting their land in the Amazon?*

Typical Analysis Goals

For these goals will differentiate between the roles of the variables:

- **Response variable**: Variable I want to better understand
- **Explanatory/predictor variables**: Variables I think might explain/predict the response variable

Q: What is the role of each variable for each goal?

→ *How many trees are in the Amazon?*

→ *Can I use remotely sensed data to predict forest types in the Amazon?*

→ *Do financial contracts prevent people from deforesting their land in the Amazon?*

Key Mechanism for Causal Goal

Random assignment: Cases are randomly assigned to levels of the **explanatory variable**

- **Random assignment** allow us to conclude if the explanatory variable **causes** changes in the response variable.

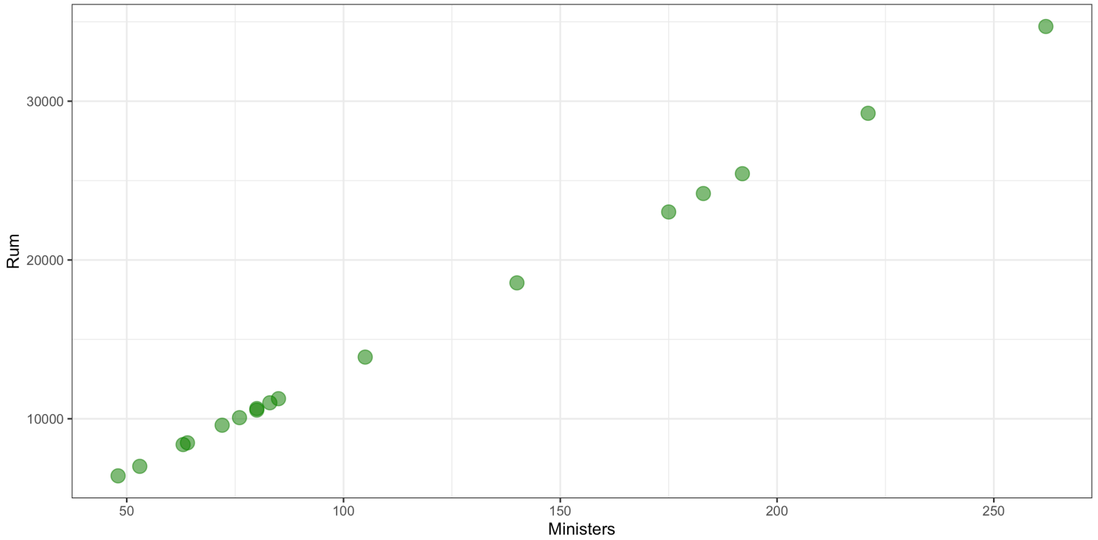
Example: COVID Vaccine Trials

To study the effectiveness of the Moderna vaccine (mRNA-1273), researchers carried out a study on over 30,000 adult volunteers with no known previous COVID-19 infection. Volunteers were randomly assigned to either receive two doses of the vaccine or two shots of saline. The incidence of symptomatic COVID-19 was 94% lower in those who received the vaccine than those who did not.

Question: Why does random assignment allow us to conclude that this vaccine was effective at preventing (early strains of) COVID-19?

Careful with Non-Random Assignment Data

We have data on the number of Methodist ministers in New England and the number of barrels of rum imported into Boston each year. The data range from 1860 to 1940.



- **Q**: Should we conclude that ministers drink a lot of rum? Or maybe that rum drinking encourages church attendance?

- **Confounding variable**: A third variable that is associated with both the explanatory variable and the response variable.
- Unclear if the explanatory variable or the confounder (or some other variable) is causing changes in the response.

Causal Inference

- **Spurious relationship**: Two variables are associated but not causally related
  - In the age of big data, lots of good examples out there.

→ *"Correlation does not imply causation."*

→ *"Correlation does not imply not causation."*

- **Causal inference**: Methods for measuring causal relationships, both with and without random assignment.

An (In)famous Historical Example: Smoking and Lung Cancer

Correlation does not imply causation … but sometimes there's causation!

- In 1950, a large study showed extremely strong association between smoking and lung cancer.
- In a 1958 article in *Nature*, **R.A. Fisher** argued that smoking does not cause lung cancer
- He argued "correlation does not imply causation"
- Context: Fisher was a smoker, and happened to be being paid by big tobacco
- How do we know Fisher was wrong? Tools from causal inference!

Notes on Correlation and Causation

- Even if we aren't intending to make rigorous causal claims, we often use the terms **explanatory** and **response** variables in analyses and modeling.
- **Correlation** is bi-directional: If X is correlated with Y, then Y is correlated with X
- **Causation** is mono-directional: If X causes Y, Y may not cause X.
- Academics in quantitative fields can be **very** sensitive to implications of causal claims - be careful with your vocab choices ("caused" vs "is associated with" vs "trends with").
- At the same time, p-hacking (for both causal and correlational findings) is a common unethical practice! We'll return to this.

Types of Studies

Observational Studies

- A study in which the researchers don't actively control the value of any variable, but simply observe the values as they naturally exist.
- **Example**: Hand washing study
  - To estimate what percent of people in the US wash their hands after using a public restroom, researchers pretended to comb their hair while observing 6000 people in public restrooms throughout the United States. They found that 85% of the people who were observed washed their hands after going to the bathroom.

(Randomized) Experiment

- A study in which the researcher actively controls one or more of the explanatory variables through random assignment.
- **Example**: COVID Trial
- Common features:
  - **Control** group that gets no treatment or a standard treatment.
  - **Placebo**: A fake treatment to control for the **placebo effect** where if people believe they are receiving a treatment, they may experience the desired effect regardless of whether the treatment is any good.
  - **Blinding**: When the subjects and/or researchers don't know the explanatory group assignments.

(Randomized) Experiment

- A study in which the researcher actively controls one or more of the explanatory variables through random assignment.

- **Another Example**: Experiment in yesterday's lab!
  - I randomly assigned you each a piece of paper with either 82 or 531 on it
  - I asked you to guess the number of dog breeds in the world
  - We'll investigate the results in Lab 04 to see if there was an "anchoring effect"

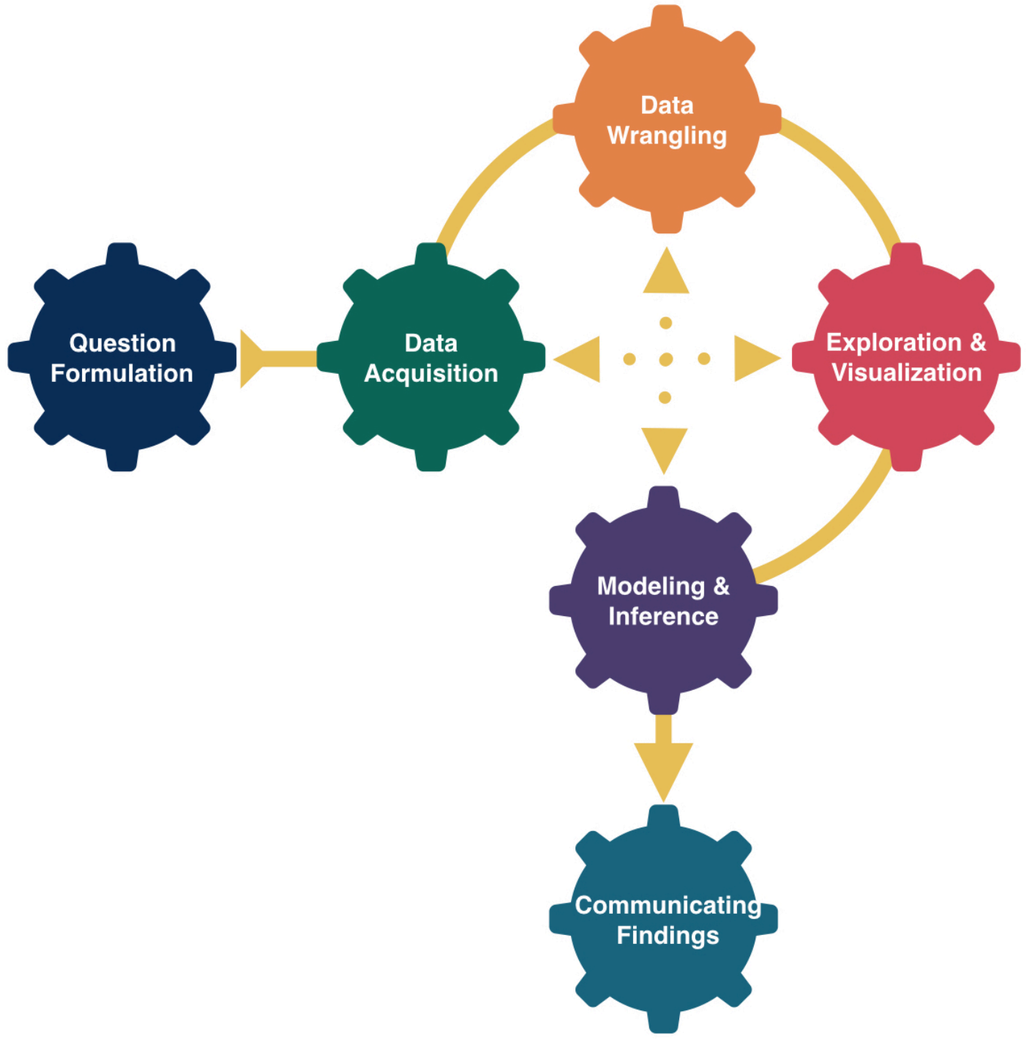
Thoughts on Data Collection Goals

- Random assignment allows you to explore **causal** relationships between your explanatory variables and the predictor variables by removing the possibility of a confounding variable
- How do we draw causal conclusions from studies without random assignment?
  - With extreme care! Try to **control** for all possible confounding variables.
  - Discuss the associations/correlations you found. Use domain knowledge to address potentially causal links.
  - Take *Math 394: Causal Inference* to learn more about causal inference.
- But also consider the goals of your analysis. Often the research question isn't causal.
- **Bottom Line:** We often have to use imperfect data to make decisions. But whenever possible, use random assignment or find *pseudo-random* contexts for the strongest causal claims.

John Snow, Cholera, and Shoe Leather: Think-pair-share

John Snow was a 19th century physician who is considered a founder of modern epidemiology. In 1854, he was investigating the drivers of a cholera epidemic in London. He suspected that contaminated water was the key cause. He realized there seemed to be no rhyme or reason to the water source that homes were connected to, and that one source was near sewage collection points, and the other further upstream. By surveying residents of homes across London, Snow created a data set that allowed him to compare cholera outcomes between those with water servicing from each source.

- What columns would you expect to be in John Snow's data set?
- Was this an experiment or an observational study?
- Why is it important that water servicing wasn't determined by location (e.g. city block or neighborhood)?
- Statistician David Freedman described Snow's work as a historic statistical success story, in part because of the amount of "shoe leather" involved. What do you think he meant?



Linear Models I: Introduction

Megan Ayers

Math 141 | Spring 2026

Monday, Week 4

Goals for Today

- Discuss the ideas of statistical modeling
- Learn two new summary statistics
- Introduce simple linear regression

Typical Analysis Goals

Descriptive: Want to estimate quantities (summary statistics) related to the population.

→ *How many trees are in the Amazon?*

Predictive: Want to predict the value of a variable.

→ *Can I use remotely sensed data to predict forest types in the Amazon?*

Causal: Want to determine if changes in a variable cause changes in another variable.

→ *Do financial contracts prevent people from deforesting their land in the Amazon?*

We will focus mainly on **descriptive modeling** in this course, and occasionally on **predictive modeling** (take Math 243: Statistical Learning to learn more). If you want to learn more about **causality**, take Math 394: Causal Inference.

Form of the Model

$$y = f(x) + \epsilon$$

where ϵ represents an error term.

Goal:

- Determine a reasonable form for $f()$. (Ex: Line, curve, ...)
- Estimate $f()$ with $\widehat{f}()$ using the data.
- Generate predicted values: $\hat{y} = \widehat{f}(x)$.

Simple Linear Regression Model

Consider this model when:

- Response variable (y): quantitative
- Explanatory variable (x): quantitative
  - Have only ONE explanatory variable.
- AND, $f()$ can be approximated by a line.

Example: The Ultimate Halloween Candy Power Ranking

> “The social contract of Halloween is simple: Provide adequate treats to costumed masses, or be prepared for late-night tricks from those dissatisfied with your offer. To help you avoid that type of vengeance, and to help you make good decisions at the supermarket this weekend, we wanted to figure out what Halloween candy people most prefer. So we devised an experiment: Pit dozens of fun-sized candy varietals against one another, and let the wisdom of the crowd decide which one was best.” – Walt Hickey

> “While we don’t know who exactly voted, we do know this: 8,371 different IP addresses voted on about 269,000 randomly generated matchups. So, not a scientific survey or anything, but a good sample of what candy people like.”

Example: The Ultimate Halloween Candy Power Ranking

Which would you prefer as a trick-or-treater?

Battle: : Candy

Hershey's Special Dark

Payday

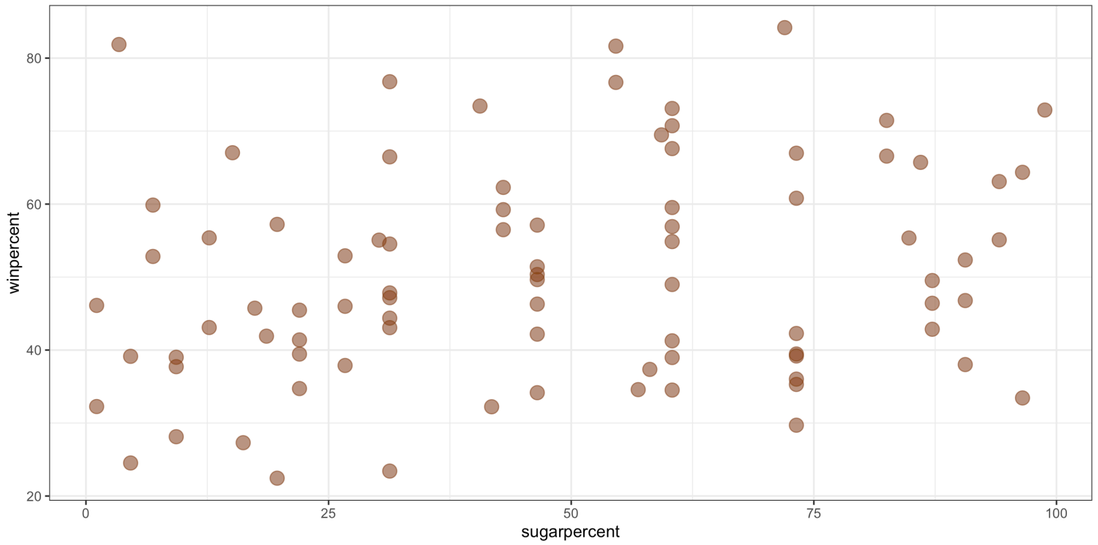
Example: The Ultimate Halloween Candy Power Ranking

```
1 candy <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/candy-data.csv"
2   mutate(sugarpercent = sugarpercent*100)
3
4 glimpse(candy)
```

```
Rows: 85
Columns: 13
$ competitorname   <chr> "100 Grand", "3 Musketeers", "One dime", "One quarter…
$ chocolate        <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
$ fruity           <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1,…
$ caramel          <dbl> 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,…
$ peanutyalmondy   <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ nougat           <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
$ crispedricewafer <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ hard             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1,…
$ bar              <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
$ pluribus         <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1,…
$ sugarpercent     <dbl> 73.2, 60.4, 1.1, 1.1, 90.6, 46.5, 60.4, 31.3, 90.6, 6…
$ pricepercent     <dbl> 0.860, 0.511, 0.116, 0.511, 0.511, 0.767, 0.767, 0.51…
```

Example: The Ultimate Halloween Candy Power Ranking

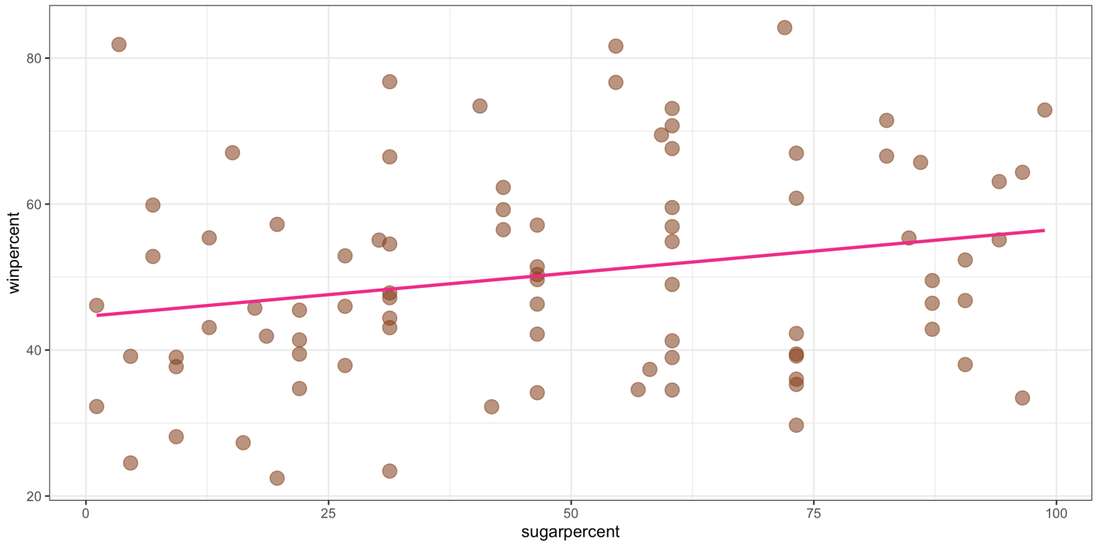
- Linear trend?
- Direction of trend?



Example: The Ultimate Halloween Candy Power Ranking

- A simple linear regression model would be suitable for these data:

$$y = \beta_0 + \beta_1 x + \epsilon$$



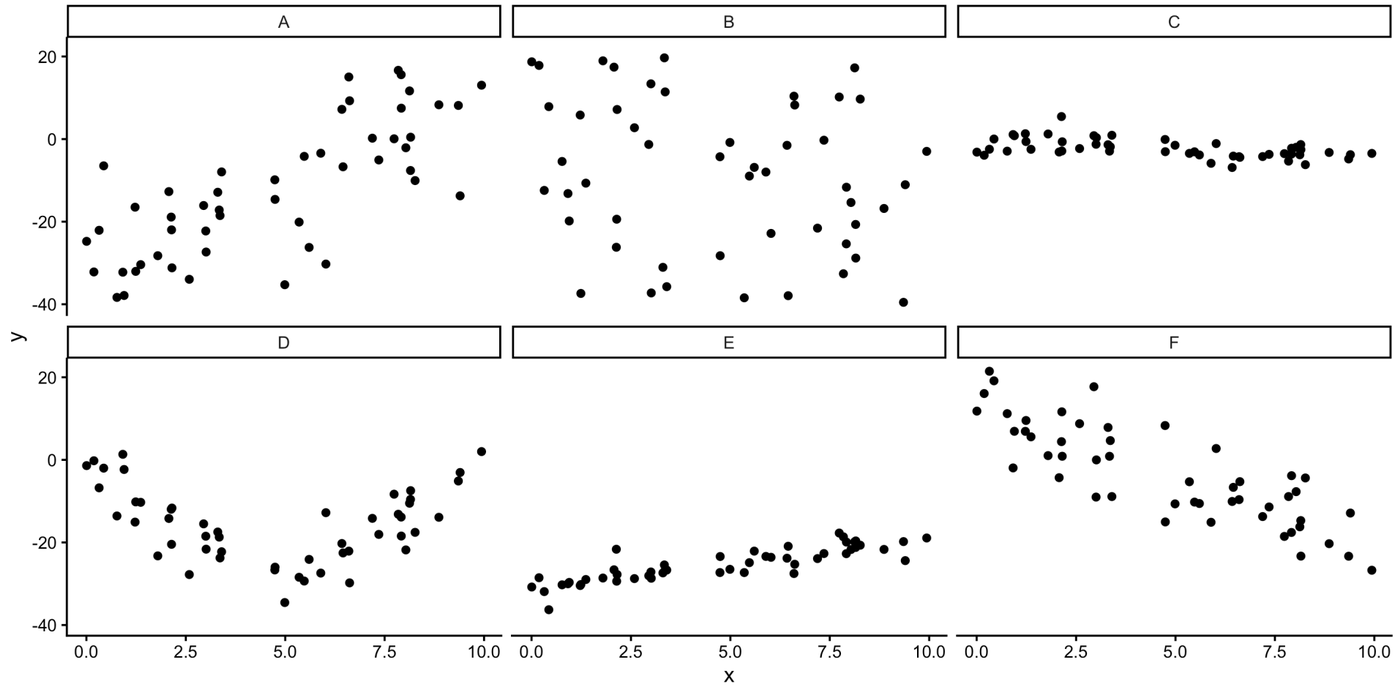
- β_0 and β_1 are fixed numbers
- β_0 represents the **intercept** of a line
- β_1 represents the **slope** of a line

- Knowing β_0 and β_1 would help us summarize and describe the relationship between x and y.

- We want to find β_1 (slope) and β_0 (intercept) so that the line fits our data well

  → Need summary statistics that quantify the strength and relationship of the linear trend

  → These will help us find a value for slope and determine how well a line fits our data



(Sample) Correlation Coefficient

- Measures the **strength** and **direction** of **linear** relationship between two quantitative variables
- Symbol: r
- Always between -1 and 1
- Sign indicates the direction of the relationship
- Magnitude indicates the strength of the linear relationship

```
cor(candy$sugarpercent, candy$winpercent)
```

```
[1] 0.2291507
```

r is calculated using the sample means ($\bar{x}$, $\bar{y}$) and standard deviations (s_x, s_y) of the variables x and y:

$$r = \frac{1}{s_x s_y} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Sidenote: (Sample) Covariance

Sample Correlation Coefficient:

$$r = \frac{1}{s_x s_y} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Sample Covariance:

$$cov(x, y) = r \times s_x s_y = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

The sample correlation coefficient is a **standardized sample covariance**, which is what causes it to only take values from -1 to 1. The sample covariance can take any real value.

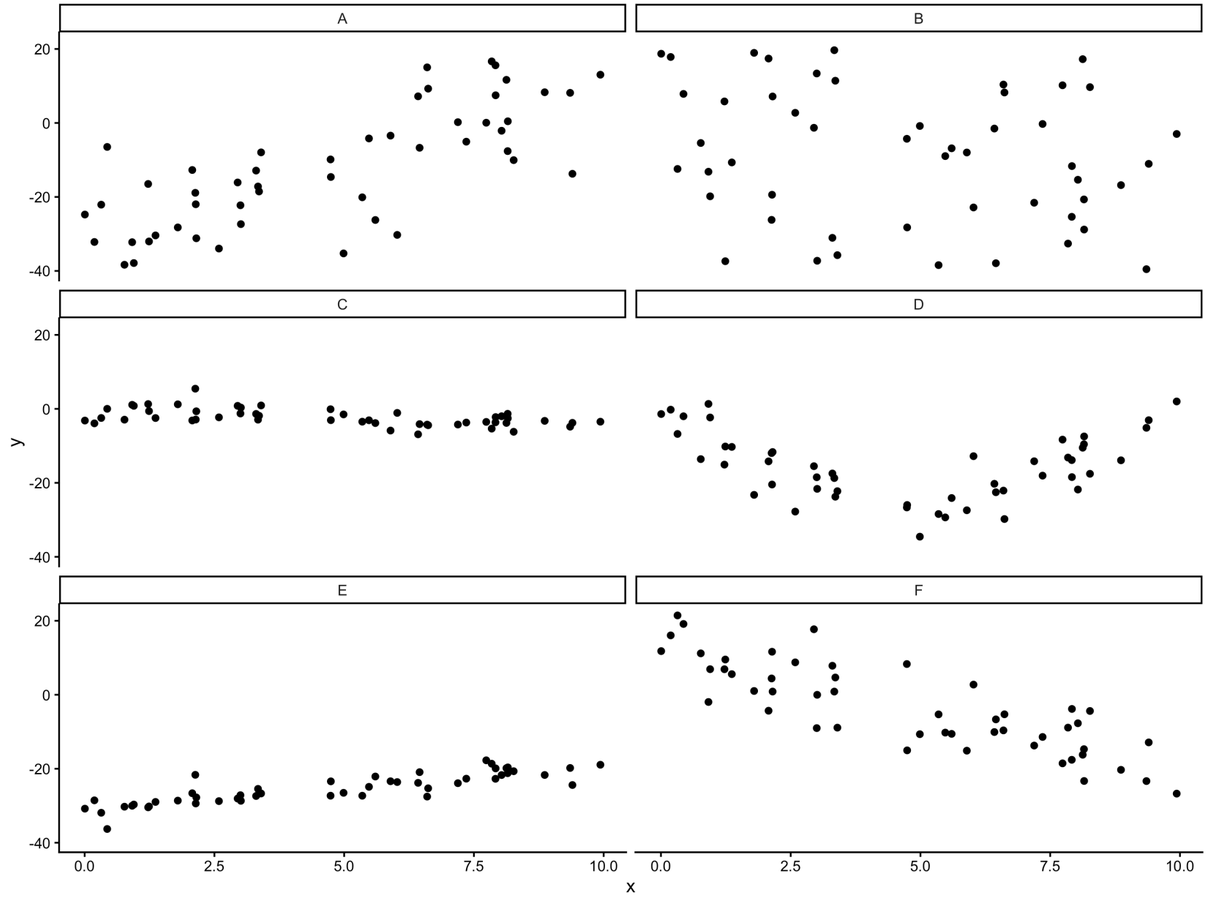
Correlation, Covariance, and Slope

Q: Which will have the largest *positive* **correlation**?

Q: Which will have the largest *negative* **correlation**?

Q: Which will have **correlation** closest to 0?

```
A: 0.7568    B: −0.2172    C: −0.5373    D: −0.1133
E: 0.863    F: −0.8343
```



- Correlation: how strong is the linear relationship *relative to noise*
- Correlation and slope are related, but not the same (ex. **A** vs **E**)

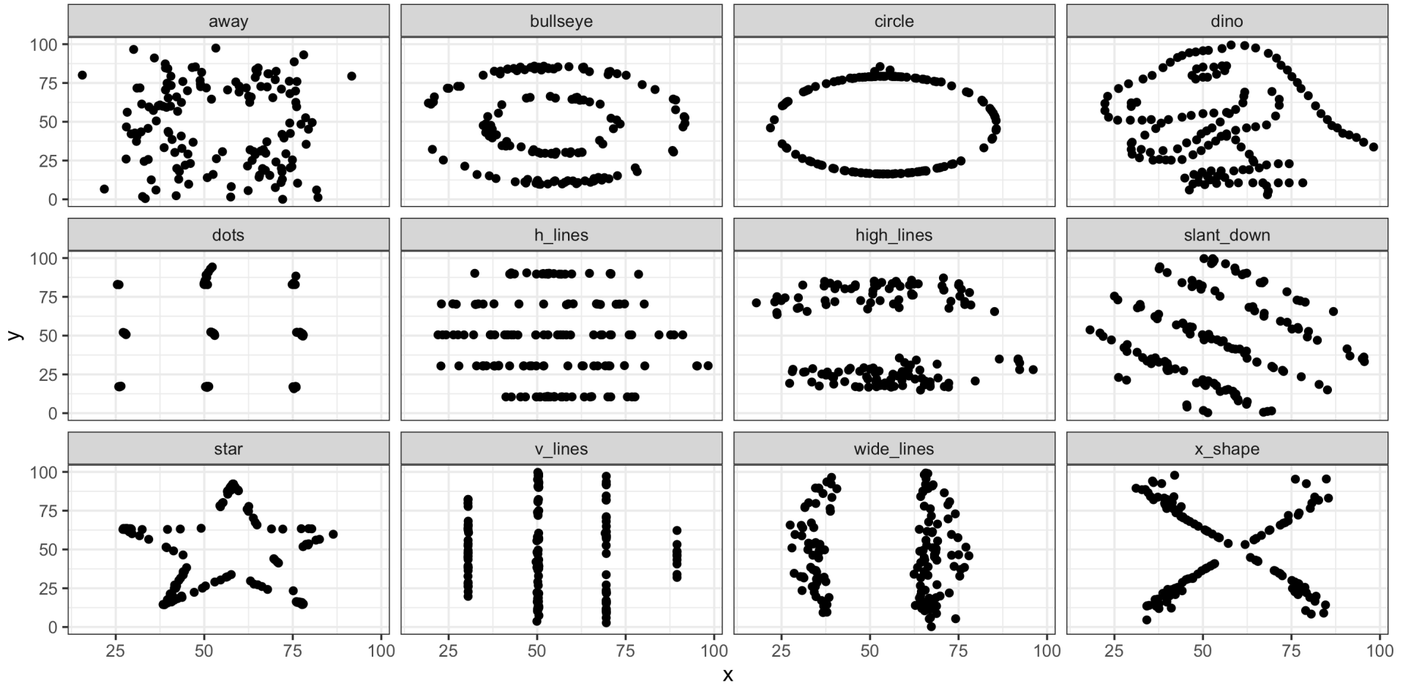
New Example

```
 1 # Correlation coefficients
 2 dat2 %>%
 3   group_by(dataset) %>%
 4   summarize(cor = cor(x, y))
```

```
# A tibble: 12 × 2
   dataset        cor
   <chr>        <dbl>
 1 away       -0.0641
 2 bullseye   -0.0686
 3 circle     -0.0683
 4 dino       -0.0645
 5 dots       -0.0603
 6 h_lines    -0.0617
 7 high_lines -0.0685
 8 slant_down -0.0690
 9 star       -0.0630
10 v_lines    -0.0694
11 wide_lines -0.0666
```

- Conclude that x and y have the same relationship across these different datasets because the correlation is the same?

Always graph the data when exploring relationships!

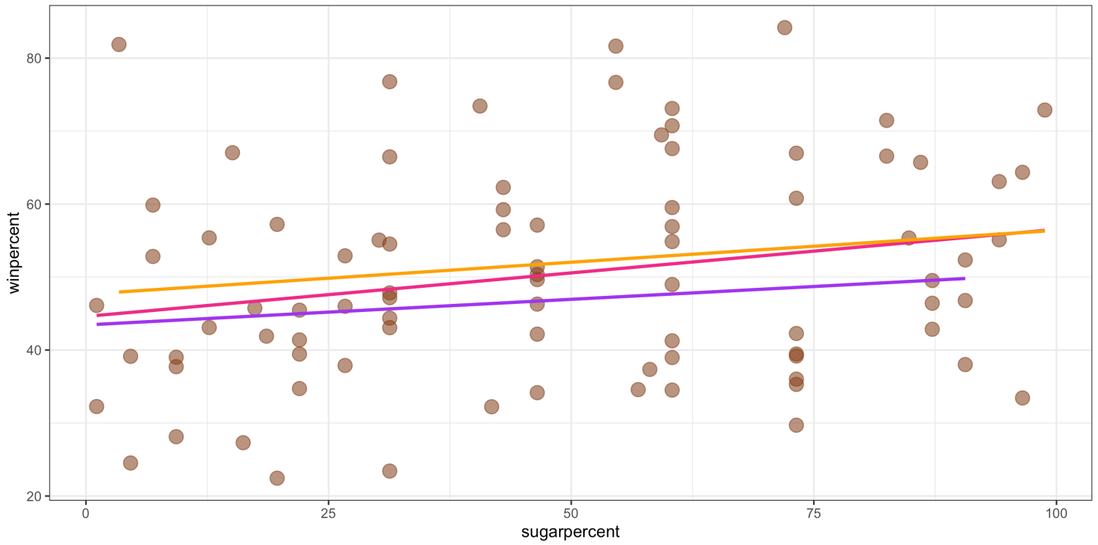


Returning to the Simple Linear Regression model...

Simple Linear Regression

Let's return to the Candy Example.

- A line is a reasonable model form.
- Where should the line be?
  - Slope? Intercept?



Form of the SLR Model

$$y = f(x) + \epsilon$$
$$y = \beta_0 + \beta_1 x + \epsilon$$

- We assume y is not perfectly predicted by $f(x)$, with ϵ representing random error.
- Need to determine the best **estimates** of β_0 and β_1.

Distinguishing between the population and the sample

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Parameters:
  - Based on the **population**
  - Unknown if we don't have data on the whole population
  - EX: β_0 and β_1

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Statistics:
  - Based on the **sample** data
  - Known
  - Usually estimate a population parameter
  - EX: $\hat{\beta}_0$ and $\hat{\beta}_1$

Method of Least Squares

Need two key definitions:

- **Fitted value**: The *estimated* value of the i-th case, given estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- **Residuals**: The *observed* error term for the i-th case, given fitted values

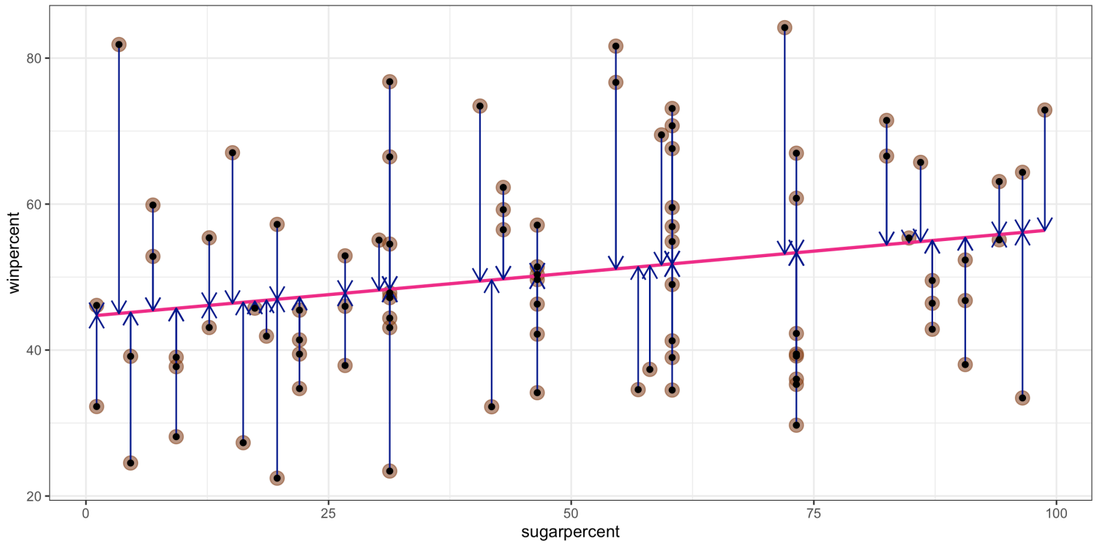
$$e_i = y_i - \hat{y}_i$$

Goal: Pick values for $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the residuals are small!

Method of Least Squares

- Want residuals to be small.
- Minimize a function of the residuals.
- Minimize:

$$\sum_{i=1}^{n} e_i^2$$



Sidenote:

- We could use $\sum_{i=1}^{n} |e_i|$
- But, this is less common, less computationally efficient, and lacks theoretical advantages
- $\sum_{i=1}^{n} e_i^2$ appropriately weights one large residual as “worse” than many small ones

Method of Least Squares

- Suppose n observations of x and y are collected: $(x_1, y_1), \ldots, (x_n, y_n)$.
- It turns out there are specific values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals, given this data:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Method of Least Squares

- Suppose n observations of x and y are collected: $(x_1, y_1), \ldots, (x_n, y_n)$.
- It turns out there are specific values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals, given this data:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_y}{s_x} r$$

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Method of Least Squares

Once we know $\hat{\beta}_0$ and $\hat{\beta}_1$, we can estimate the whole function with:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Called the **least squares line**, **regression line**, or the **line of best fit**.

We need to be precise and careful when interpreting β_0 and β_1 (and their estimates)

- **Intercept:** We [expect/predict] y to be β_0 on average when $x = 0$.
- **Slope**: For a one-unit increase in x, we [expect/predict] y to change by β_1 units on average.

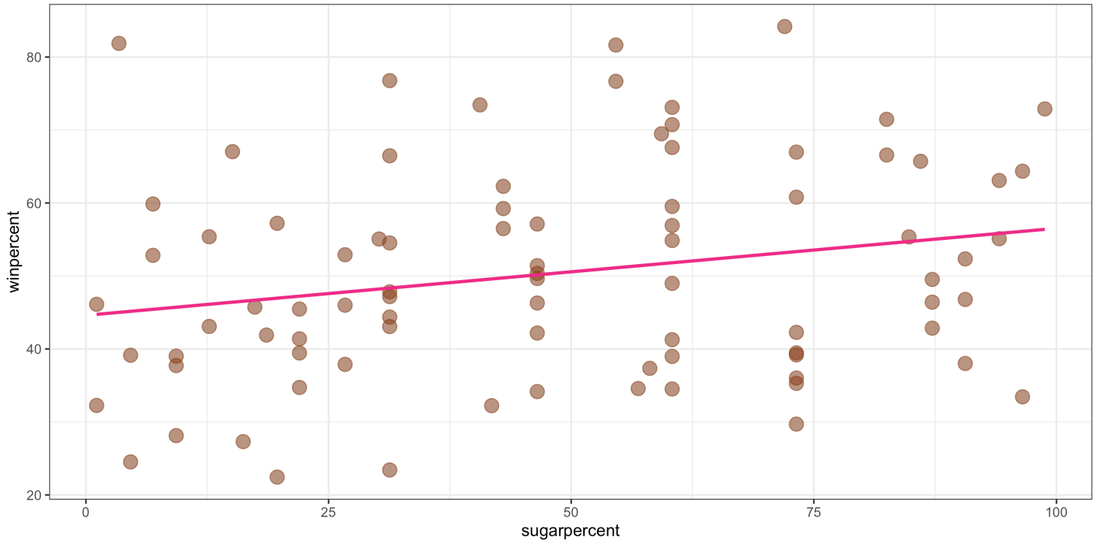
Unless experimental data is involved, avoid causal language.

- Example of causal language: "When x increases by 1 unit, y will increase by β_1"

Method of Least Squares

`ggplot2` will compute the line and add it to your plot using `geom_smooth(method = "lm")`

```
ggplot(data = candy,
       mapping = aes(x = sugarpercent,
                     y = winpercent)) +
  geom_point(alpha = 0.6, size = 4,
             color = "chocolate4") +
  geom_smooth(method = "lm", se = FALSE,
              color = "deeppink2")
```

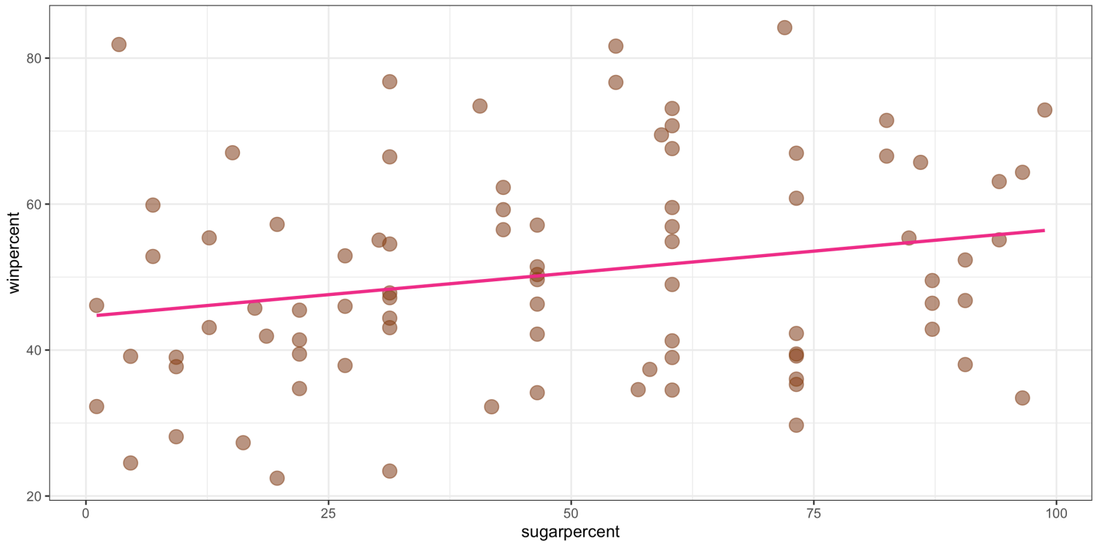


We can calculate the **exact** values of $\hat{\beta}_0$ and $\hat{\beta}_1$ using our formulas, by hand or in R

Method of Least Squares

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

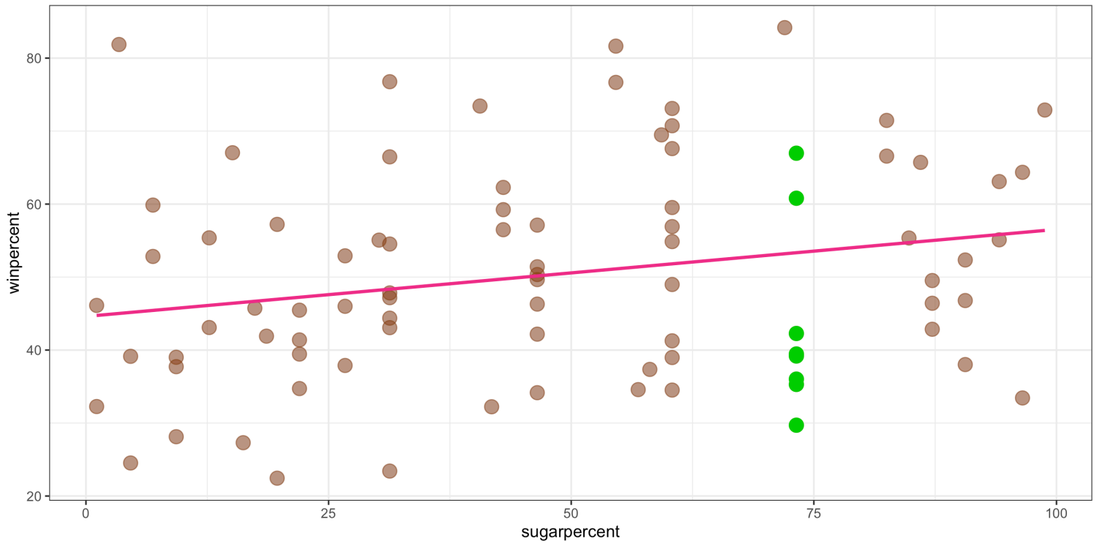
In this case, $\hat{\beta}_0 = 44.6094$ and $\hat{\beta}_1 = 0.1192$.



Method of Least Squares

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

In this case, $\hat{\beta}_0 = 44.6094$ and $\hat{\beta}_1 = 0.1192$.

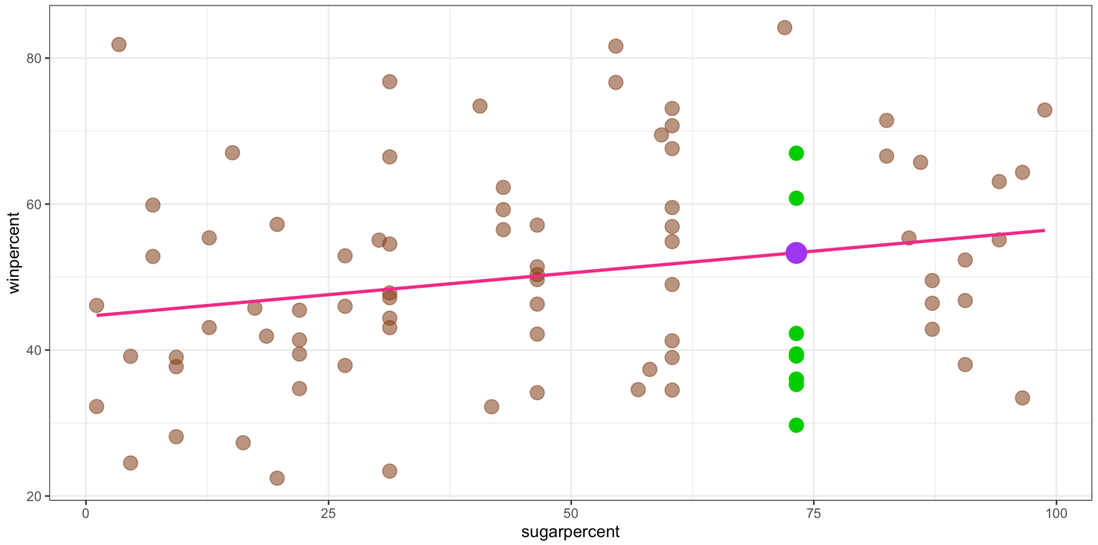


- **Q**: Suppose a new candy bar has `sugarpercent = 73`. What does the model predict for `winpercent`?

Method of Least Squares

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

In this case, $\hat{\beta}_0 = 44.6094$ and $\hat{\beta}_1 = 0.1192$.

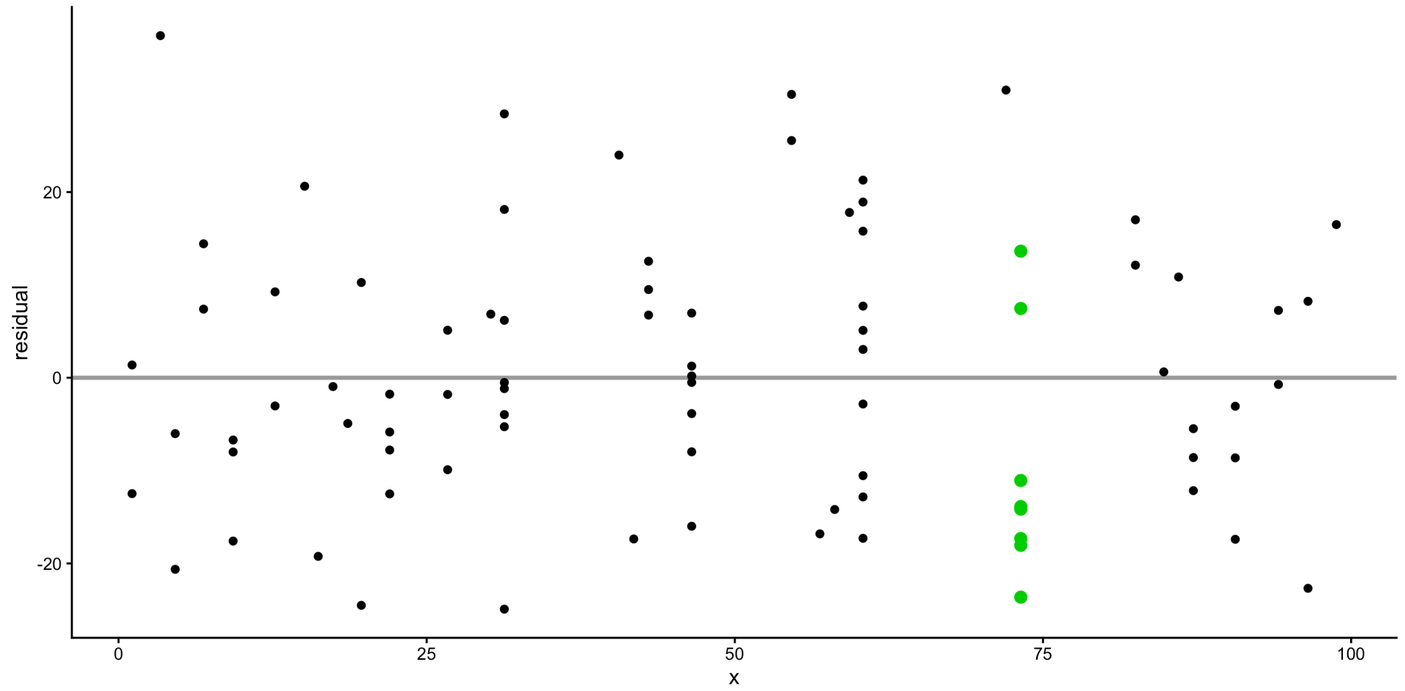


- **Q**: Suppose a new candy bar has `sugarpercent = 73`. What does the model predict for `winpercent`?
- **A**: $\hat{y} = 44.6094 + 0.1192 \cdot 73 = 53.311$

- This is different from the actual `winpercents` we see for candies with `sugarpercent = 73`.
- This isn't unexpected: the line predicts the **average** y for each value of x.

Residual plot preview

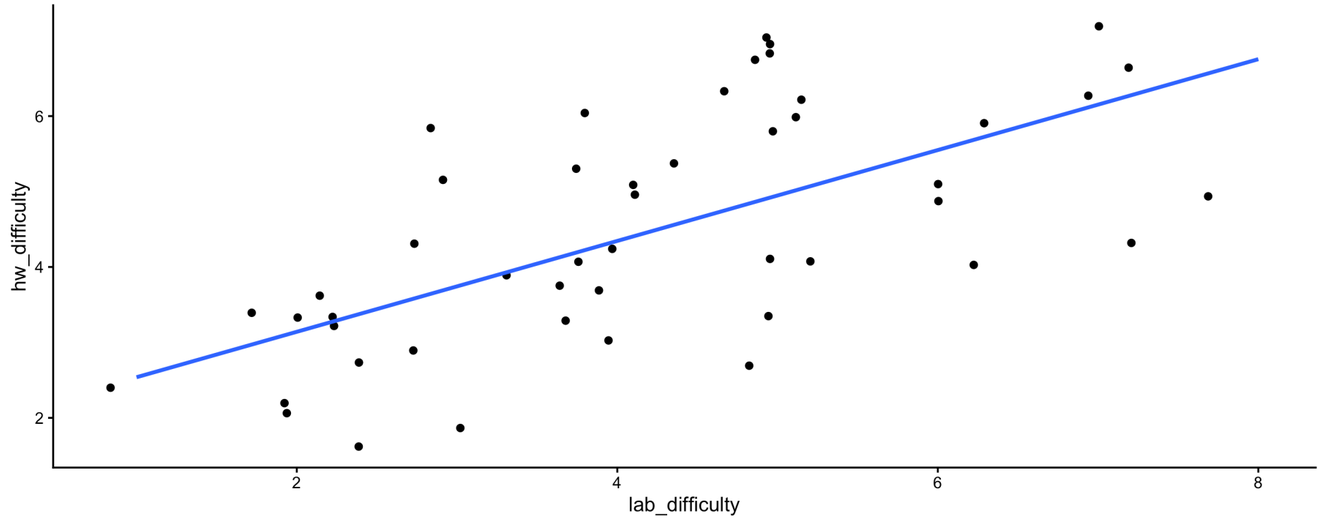
We can visualize the accuracy of a linear model using residual plots:



- We'll learn to use different residual plots to more formally assess linear model assumptions

Activity: Extra practice interpreting coefficients

```r
x <- read.csv("data/qods-anon.csv")
names(x) <- c("lab_difficulty", "hw_difficulty")

set.seed(123)
ggplot(x, aes(x = lab_difficulty, y = hw_difficulty)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_classic()
```



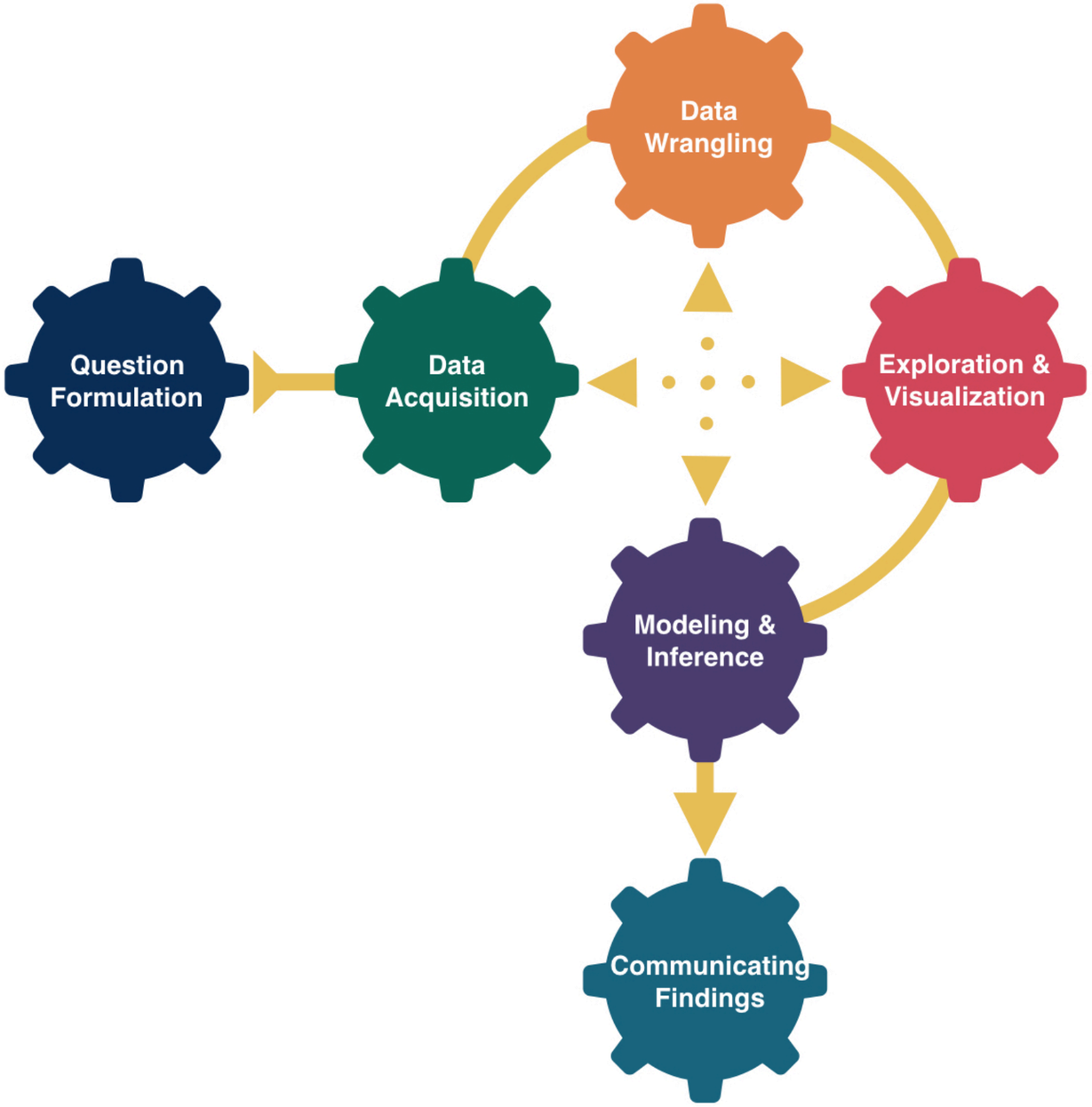
- **Q1**: How should we interpret the intercept?
- **Q2**: How should we interpret the coefficient on `lab_difficulty`?

```r
coef(lm(hw_difficulty ~ lab_difficulty, x))
```

```
(Intercept) lab_difficulty
  1.9365233      0.6022233
```

Next time: more simple linear regression

- We'll use R to compute the least squares line and the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- We'll introduce the formal assumptions we must make when doing linear regression, and diagnostic plots to check these assumptions.



Linear Models II: Accuracy and Assumptions

Announcements

- Fill out the Week 4 Survey

https://tinyurl.com/4wk-survey

Goals for Today

- Recall simple linear regression
- Do simple linear regression in R
- Discuss model assumptions for linear regression
- Assess accuracy of linear regression models
- Discuss “extrapolation”

Form of the Model

$$y = f(x) + \epsilon$$

where ϵ represents an error term.

Goal:

- Determine a reasonable form for $f()$. (Ex: Line, curve, ...)
- Estimate $f()$ with $\widehat{f}()$ using the data.
- Generate predicted values: $\hat{y} = \widehat{f}(x)$.

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \epsilon$$

Consider this model when:

- Response variable (y): quantitative
- Explanatory variable (x): quantitative
  - Have only ONE explanatory variable.
- AND, $f()$ can be approximated by a line.
- Need to determine the best **estimates** of β_0 and β_1.

Distinguishing between the population and the sample

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Parameters:
  - Based on the **population**
  - Unknown then if don't have data on the whole population
  - EX: β_0 and β_1

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Statistics:
  - Based on the **sample** data
  - Known
  - Usually estimate a population parameter
  - EX: $\hat{\beta}_0$ and $\hat{\beta}_1$

Regression goals

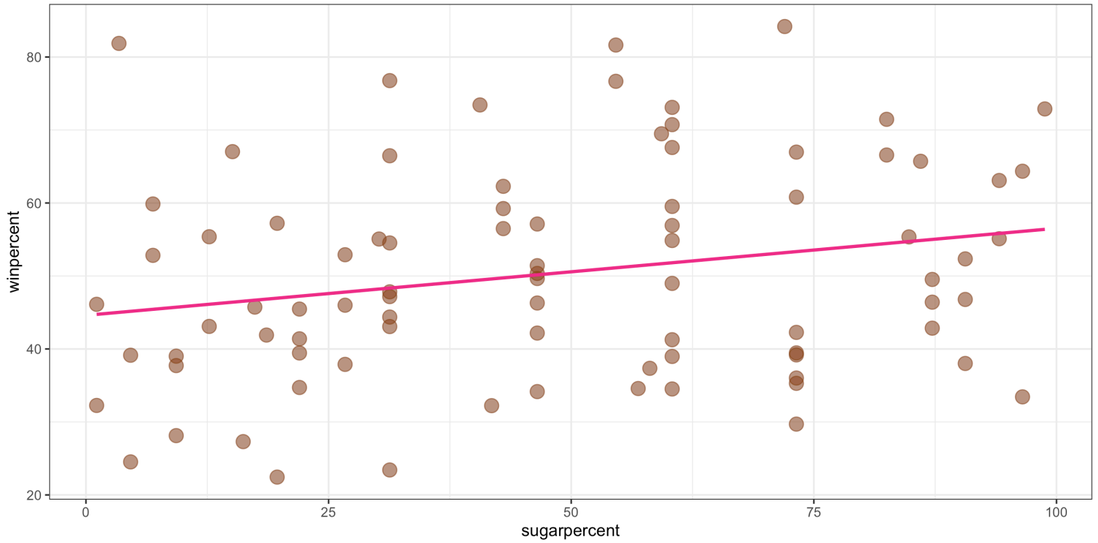
Linear regression is used for 2 main tasks:

- **Explaining** relationships between variables in a data set so that we can **infer** relationships in the population
  - Ex. We observe that a sample of Reed students who sleep more have higher test scores
  - Based on this positive relationship in the sample, we might infer that sleep is associated with higher test scores for similar students **in general**
  - We focus most on **coefficients**
- **Predicting** values of the response variable based on values of the explanatory variable
  - Ex. Using data from 1960-2015, we predict the atmosphere will contain 410 ppm CO_2 in 2025
  - We focus most on **fitted values** and **residuals**

Our Modeling Goal

Recall our modeling goal: **predict win percentage by using the sugar percentage variable.**

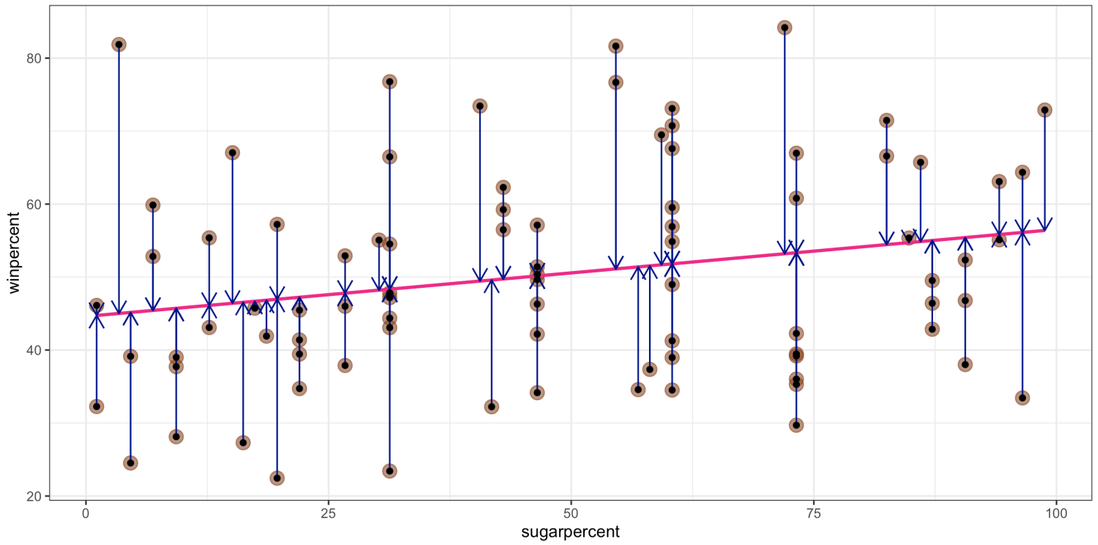
```
candy <- read_csv("https://raw.githubusercontent.com/fi
  mutate(sugarpercent = sugarpercent*100)

ggplot(data = candy,
       mapping = aes(x = sugarpercent,
                     y = winpercent)) +
  geom_point(alpha = 0.6, size = 4,
             color = "chocolate4") +
  geom_smooth(method = "lm", se = FALSE,
              color = "deeppink2")
```



Method of Least Squares

- Want residuals to be small.
- Minimize a function of the residuals.
- Minimize:

$$\sum_{i=1}^{n} e_i^2$$



Method of Least Squares

After minimizing the sum of squared residuals, you get the following equations:

Get the following equations:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Note: A large slope does **not necessarily** indicate strong correlation, and a small slope does **not necessarily** indicate lack of correlation.

Method of Least Squares

Then we can estimate the whole function with:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Called the **least squares line**, **regression line**, or the **line of best fit**.

Constructing the Simple Linear Regression Model in R

We can use the `lm()` function to construct the simple linear regression model in R and the `get_regression_table()` function from `moderndive` to interpret it.

```
1 mod <- lm(winpercent ~ sugarpercent, data = candy)
2 get_regression_table(mod)
```

```
# A tibble: 2 × 7
  term         estimate std_error statistic p_value lower_ci upper_ci
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept      44.609     3.086    14.455   0       38.471   50.748
2 sugarpercent    0.119     0.056     2.145   0.035    0.009    0.23
```

What is the fitted model form?

$$\begin{aligned}\hat{y} &= \widehat{\beta_0} + \widehat{\beta_1} \times x_{sugarpercent} \\ &= 44.6094 + 0.1192 \times x_{sugarpercent}\end{aligned}$$

We can use the `lm()` function to construct the simple linear regression model in R and the `get_regression_table()` function to interpret it.

```
mod <- lm(winpercent ~ sugarpercent, data = candy)
get_regression_table(mod)
```

```
# A tibble: 2 × 7
  term         estimate std_error statistic p_value lower_ci upper_ci
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept      44.609     3.086    14.455   0       38.471   50.748
2 sugarpercent    0.119     0.056     2.145   0.035    0.009    0.23
```

Q: How do we interpret the coefficients?

Coefficient Interpretation

$$\begin{aligned}\hat{y} &= \widehat{\beta_0} + \widehat{\beta_1} \times x_{sugar\,percent} \\ &= 44.6094 + 0.1192 \times x_{sugar\,percent}\end{aligned}$$

We need to be precise and careful when interpreting estimated coefficients!

- **Intercept:** We expect/predict y to be $\hat{\beta}_0$ on average when $x = 0$.
- **Slope:** For a one-unit increase in x, we expect/predict y to change by $\hat{\beta}_1$ units on average.

These interpretations are non-specific to the context of our model, but when we are interpreting coefficients, we always need to interpret the coefficients in context

Coefficient Interpretation: In Context

$$\begin{aligned}\hat{y} &= \widehat{\beta_0} + \widehat{\beta_1} \times x_{sugar\,percent} \\ &= 44.6094 + 0.1192 \times x_{sugar\,percent}\end{aligned}$$

- **Intercept:** We expect/predict a candy's win percentage to be 44.6094 on average when their sugar percentage is 0.
- **Slope:** For a one-unit increase in sugar percentage, we expect/predict the win percentage of a candy to change by 0.1192 units on average.

Prediction, Interpolation, and Extrapolation

```
1 new_cases <- data.frame(sugarpercent = c(22, 77, 150))
2 predict(mod, newdata = new_cases)
```

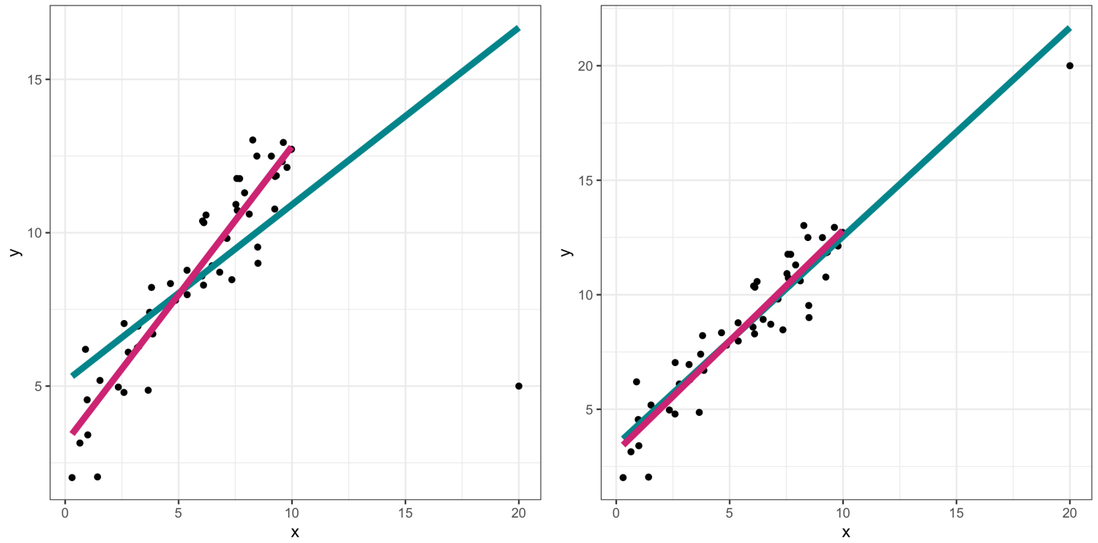
```
       1        2        3 
47.23269 53.79082 62.49524 
```

- We didn't have any treats in our sample with a sugar percentage of 77%. Can we still make this prediction?
  - Called **interpolation**

- We didn't have any treats in our sample with a sugar percentage of 150%. Can we still make this prediction?
  - Called **extrapolation**

Cautions

- Careful to only predict values within the range of x values in the sample.
- Make sure to investigate **outliers**: observations that fall far from the cloud of points.
  - High leverage points: extreme *predictor* value



A closer look at our model

```
1 get_regression_table(mod)
```

```
# A tibble: 2 × 7
  term         estimate std_error statistic p_value lower_ci upper_ci
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept      44.609     3.086    14.455   0       38.471   50.748
2 sugarpercent    0.119     0.056     2.145   0.035    0.009    0.23
```

$$\begin{aligned}\hat{y} &= \widehat{\beta_0} + \widehat{\beta_1} \times x_{sugar\,percent} \\ &= 44.6094 + 0.1192 \times x_{sugar\,percent}\end{aligned}$$

What assumptions have we made?

Linear Regression Assumptions

We can *always* find the line of best fit to explore data, but…

To make accurate predictions or inferences, certain conditions should be met.

To responsibly use linear regression tools for prediction or inference, we require:

1. **Linearity:** The relationship between explanatory and response variables must be approximately linear
    - Check using scatterplot of data, or **residual plot**

2. **Independence:** The observations should be independent of one another.
    - Can check by considering data context, and
    - sometimes by looking at scatterplots/**residual plots**

3. **Normality:** The distribution of residuals should be *approximately* bell-shaped, unimodal, symmetric, and centered at 0 at every "slice" of the explanatory variable
    - Simple check: look at **histogram of residuals**
    - Better to use a **Q-Q plot**

4. **Equal Variability:** Variance of residuals should be roughly constant across data set. Also called "homoscedasticity". Models that violate this assumption are sometimes called "heteroscedastic"
    - Check using **residual plot**.

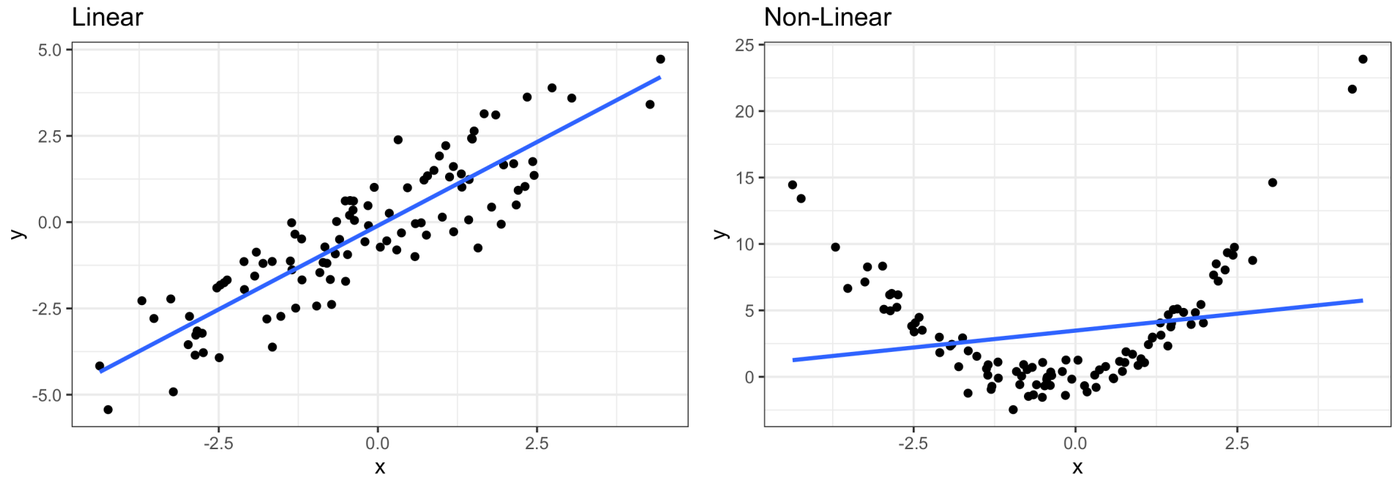
A cute way to remember this: "LINE"

1. **L**inearity
2. **I**ndependence
3. **N**ormality
4. **E**qual Variability

- We assess these using **diagnostic plots** (y vs x scatterplot, residual plot, residual histogram, q-q plot)
- Later in the course, we'll learn why **I**, **N**, and **E** are required for **inference**

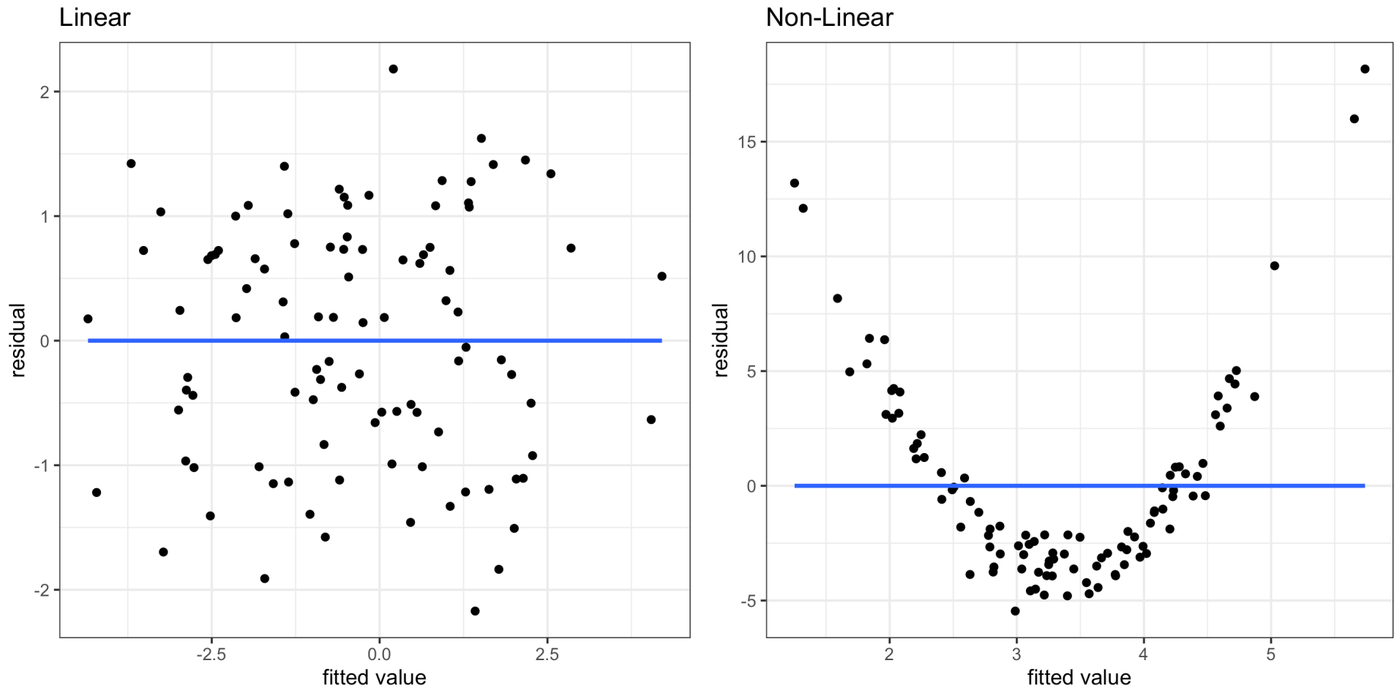
Assessing conditions: Linearity

Linearity: The relationship between explanatory and response variables must be approximately linear



- Points should be evenly distributed above and below the regression line **at each "slice" of x**
- If data is non-linear:
  - Slope does not adequately describe relationship and predictions can be very inaccurate
  - More advanced modeling techniques should be used instead (take STAT 243)

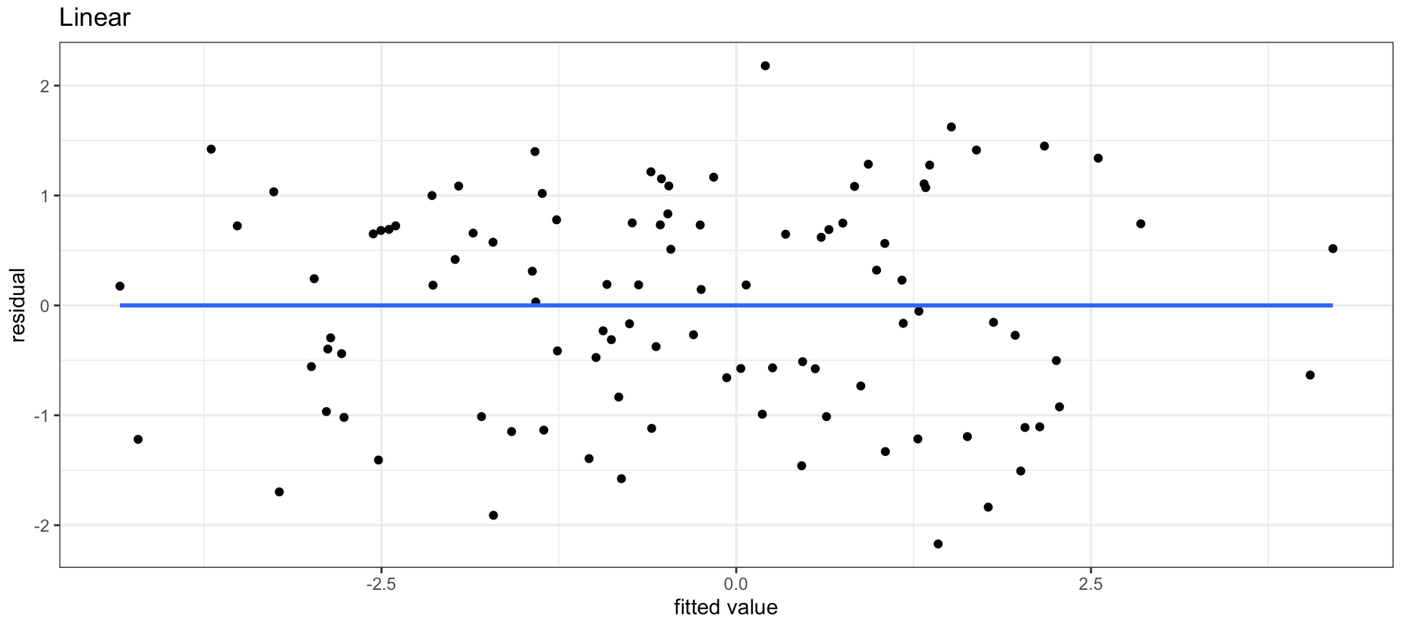
Assessing conditions: Linearity



- **Residual plots** can also be useful for checking linearity
- These plot model **residuals** against the **fitted values**
- We don't want to see any trends here

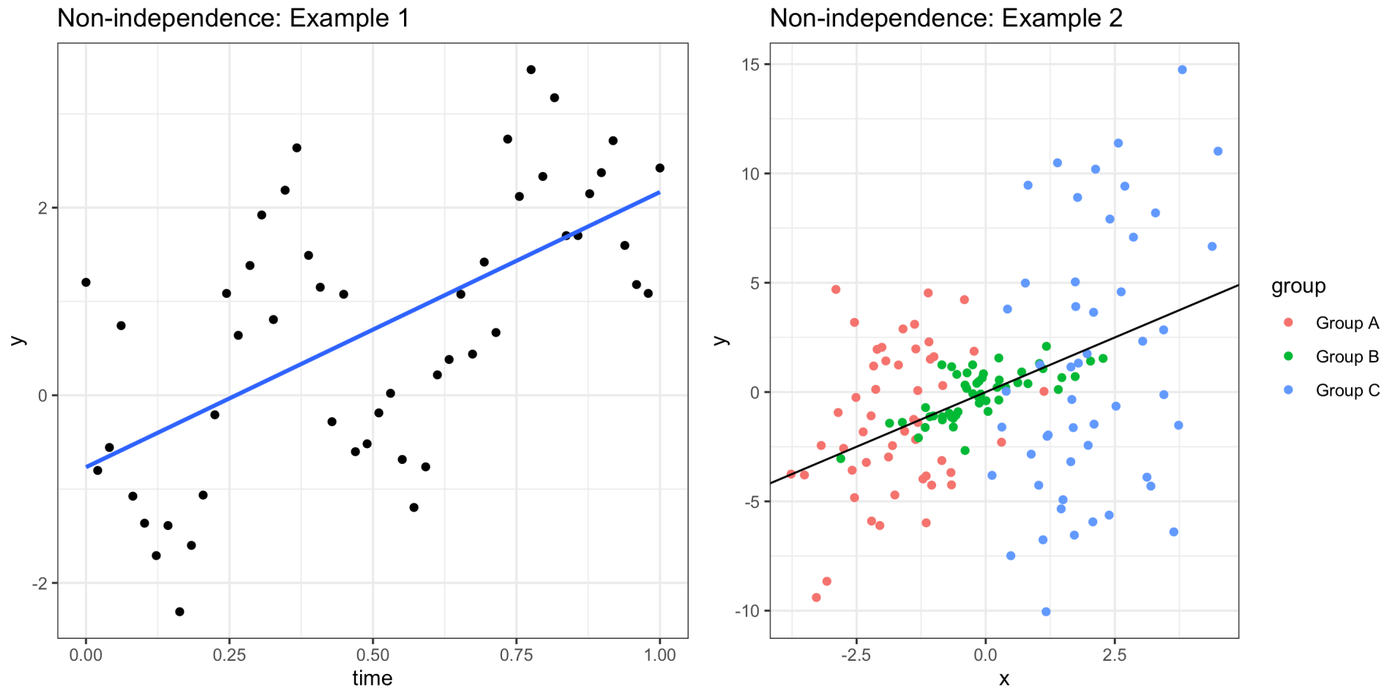
Assessing conditions: Linearity (Code Example)

```r
library(moderndive)

# Creating the model
lm1 <- lm(data = my_df, y1 ~ x)

# *** Pulling out model inputs, residuals, and fitted values ***
res1 <- get_regression_points(lm1)

# Plotting residuals vs fitted values
(g1 <- ggplot(res1, aes(x = y1_hat, y = residual)) +
  geom_point() +
  theme_bw() +
  geom_smooth(method = "lm", se = F) +
  labs(x = "fitted value", y = "residual", title = "Linear"))
```


Linear
residual
fitted value
2
1
0
-1
-2
-2.5
0.0
2.5

Assessing conditions: Independence

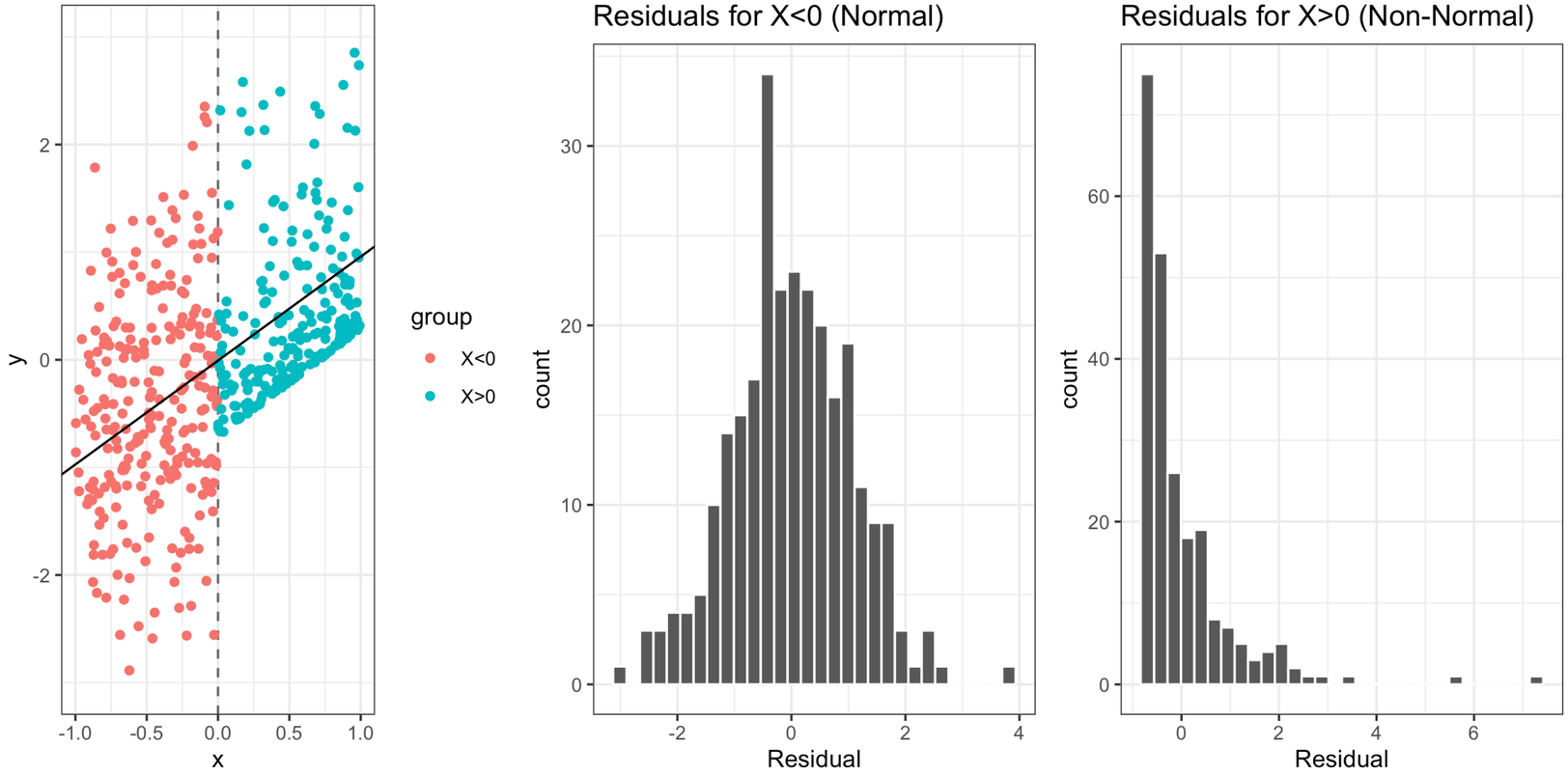
Independence: The observations should be independent of one another



- If observations are not independent, inference about the population may be misleading.
- We usually assess independence qualitatively
  - Are observations inherently connected? (Beyond variables in the model)

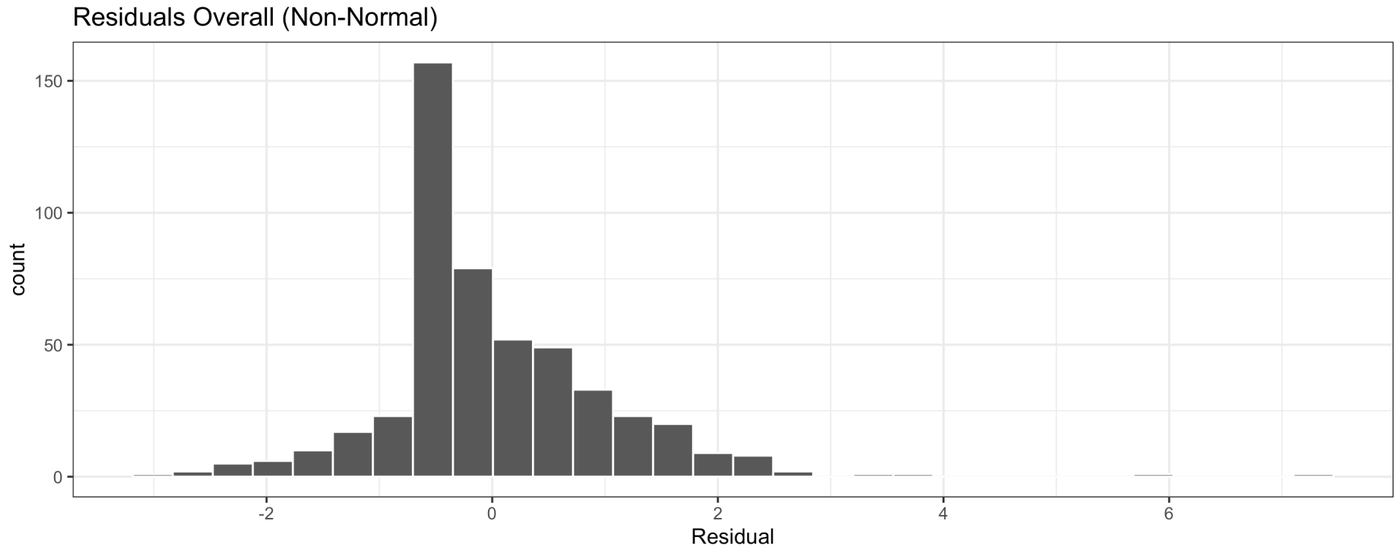
Assessing conditions: Normality (of residuals)

Normality: The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0 at every "slice" of the explanatory variable



Assessing conditions: Normality (of residuals)

Normality: The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0 at every "slice" of the explanatory variable



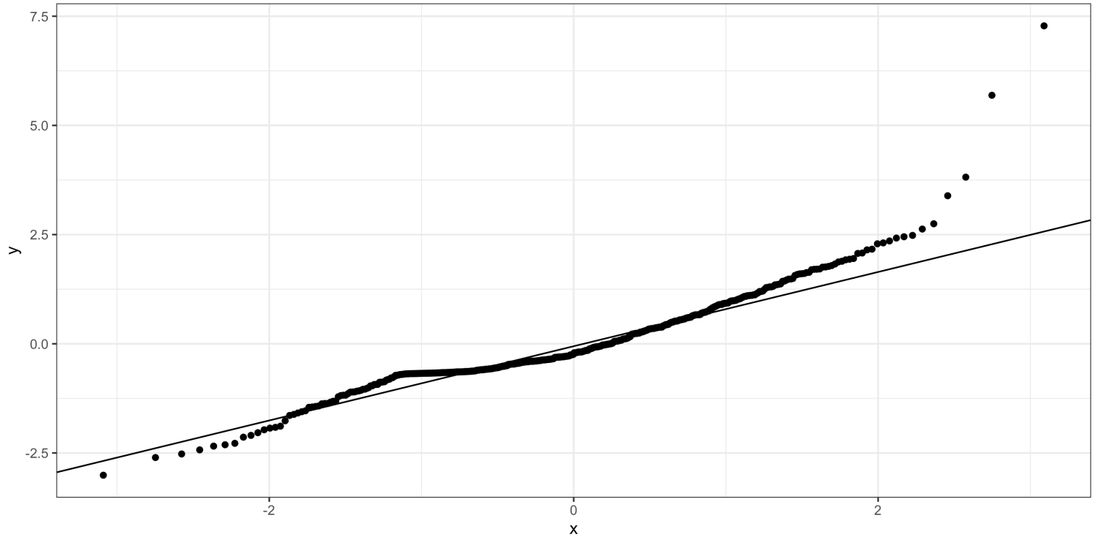
If residuals are non-Normal…

- Some predictions can be very inaccurate
- Inference about the population may be misleading

Assessing conditions: Normality (of residuals)

Normality: The distribution of residuals should be "Normal"

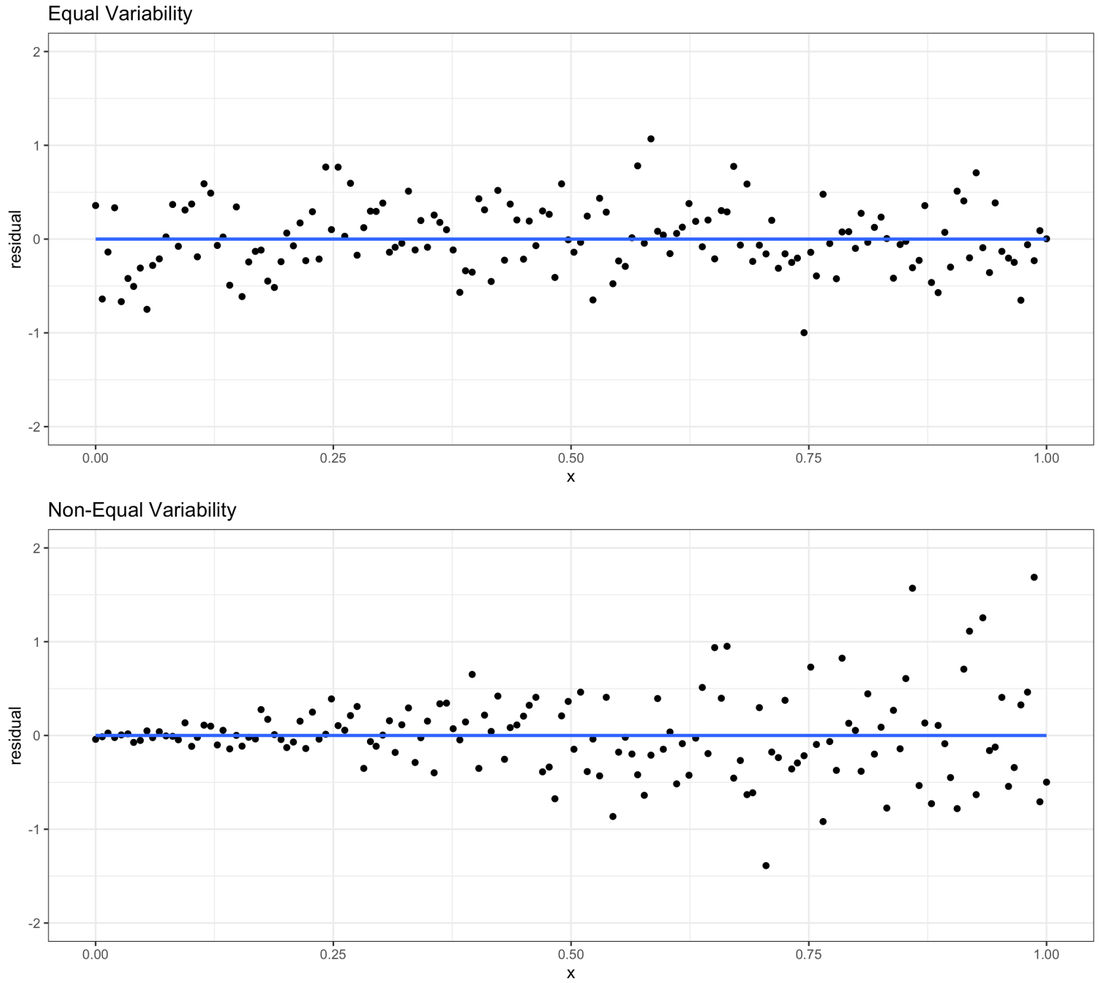
```
# New aes() argument!!
ggplot(my_df, aes(sample = residual)) +
  geom_qq() +
  geom_qq_line() +
  theme_bw()
```



- "Q-Q" plots make it even easier to check Normality - we'll learn some of the theory behind this later.
- We want points to fall as close to the line as possible.

Assessing conditions: Equal variability (of residuals)

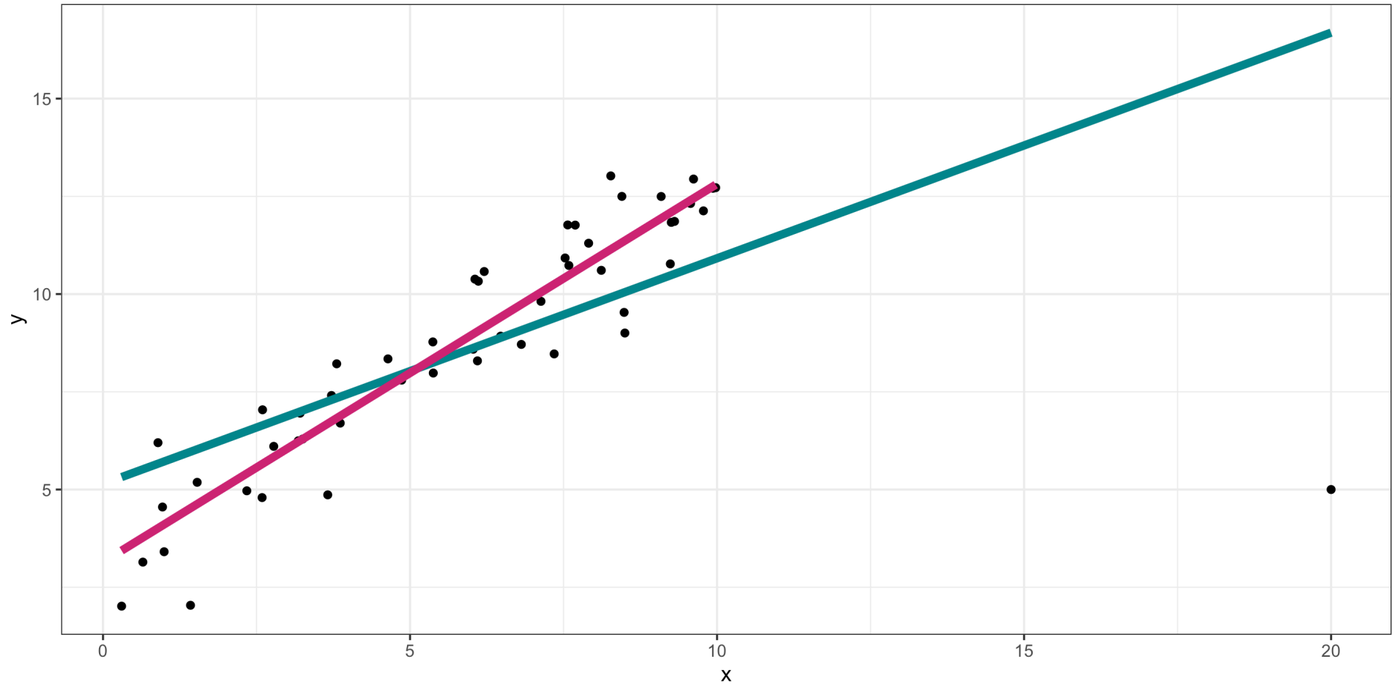
Equal Variability: Variance of residuals should be approximately constant across the data



- **Residual plots** are also very useful for this
- If equal variability isn't met:
  - Inference about the population may be misleading
  - (Potential) outliers in high-variability range are more influential

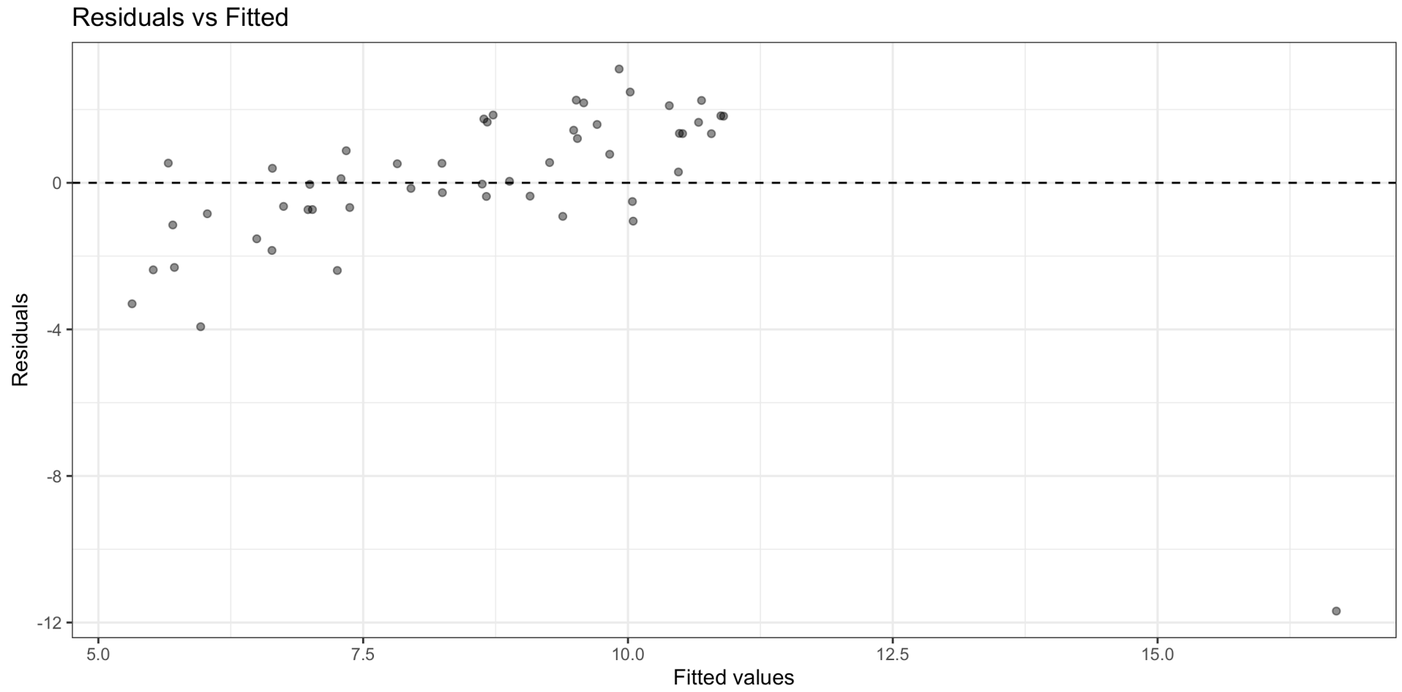
Another example: high leverage points

Remember the outlier example:



What do diagnostics look like when we fit the teal model?

Diagnosing the model



- In this case, can already see the outlier in the residual vs fitted plot.

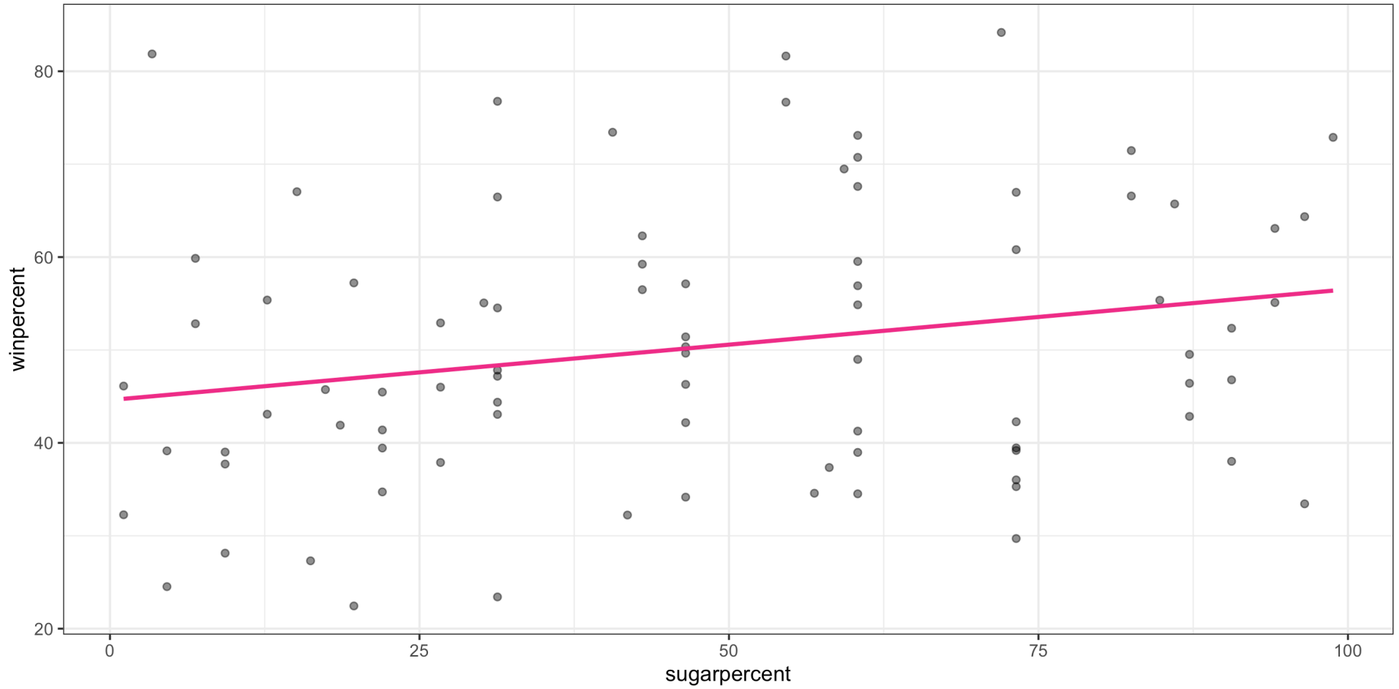
What to do with high leverage points?

- Depends on **context**
- It may be reasonable to remove them and refit the model
  - If we believe the data is the result of an error
  - If we are most focused on **prediction** for non-outliers, or if our population of interest does not include the outlier
- But we shouldn't remove outliers just because we don't like the results that they produce
  - Especially if our goal is **descriptive** or **explanatory**, and the outlier represents a valid data point

```
dat_no_outlier <- dat %>%
  filter(x < 15)

pink_mod <- lm(y ~ x, dat_no_outlier)
```

Recall our `candy` simple linear regression model

```
1 mod <- lm(winpercent ~ sugarpercent, data = candy)
```

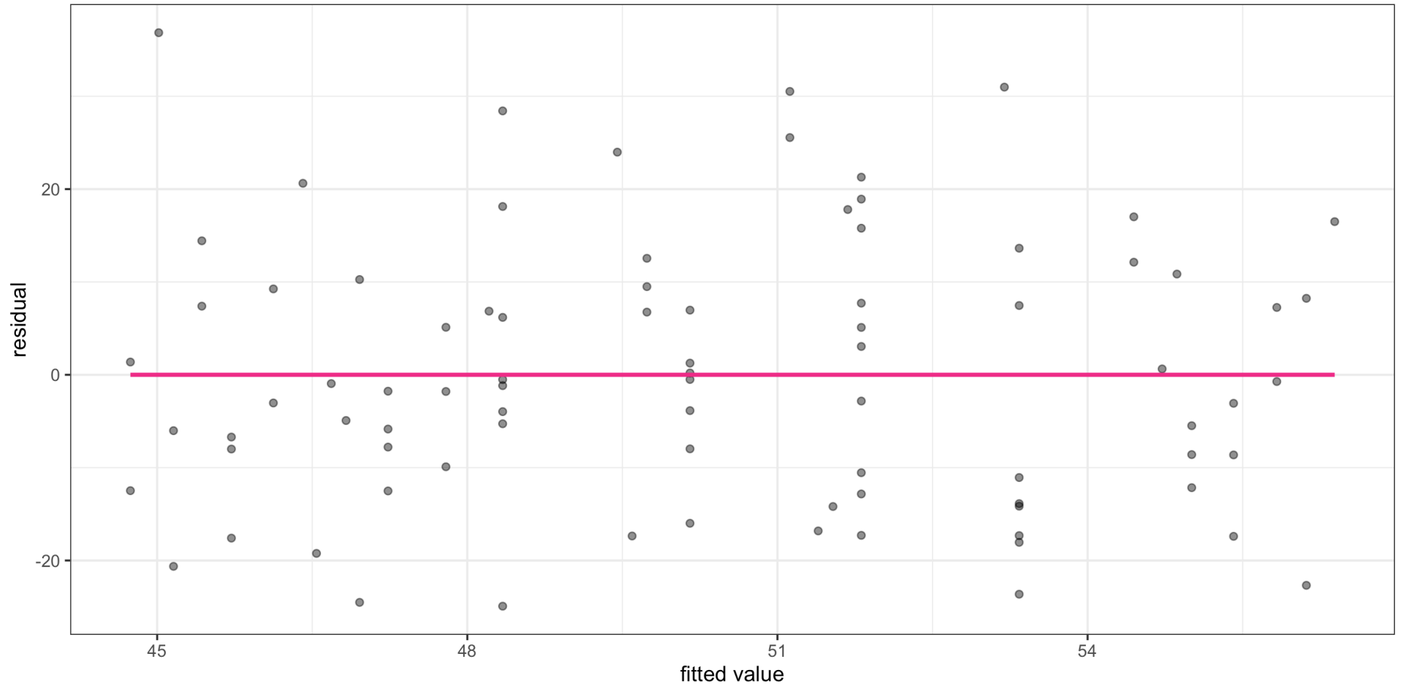


Let's check if this meets the **LINE** assumptions

Q: First, what does this graph tell us about **Linearity**?

Residual vs. fitted plot

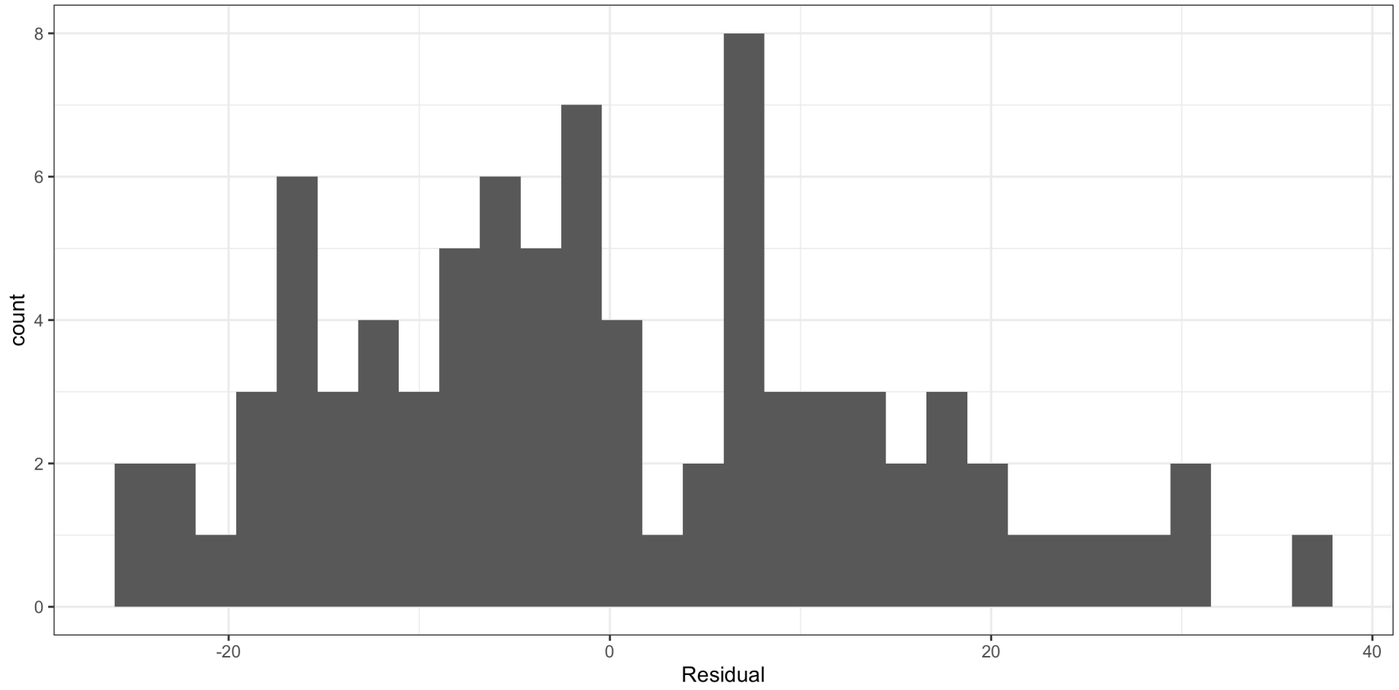
```
1 res <- get_regression_points(mod)
2
3 ggplot(res, aes(x = winpercent_hat, y = residual)) +
4   geom_point(alpha = 0.5) +
5   geom_smooth(method = "lm", se = FALSE, color = "deeppink2") +
6   labs(x = "fitted value", y = "residual")
```



- Linearity: ✅
- Independence: ✅
- Equal Variability: ✅

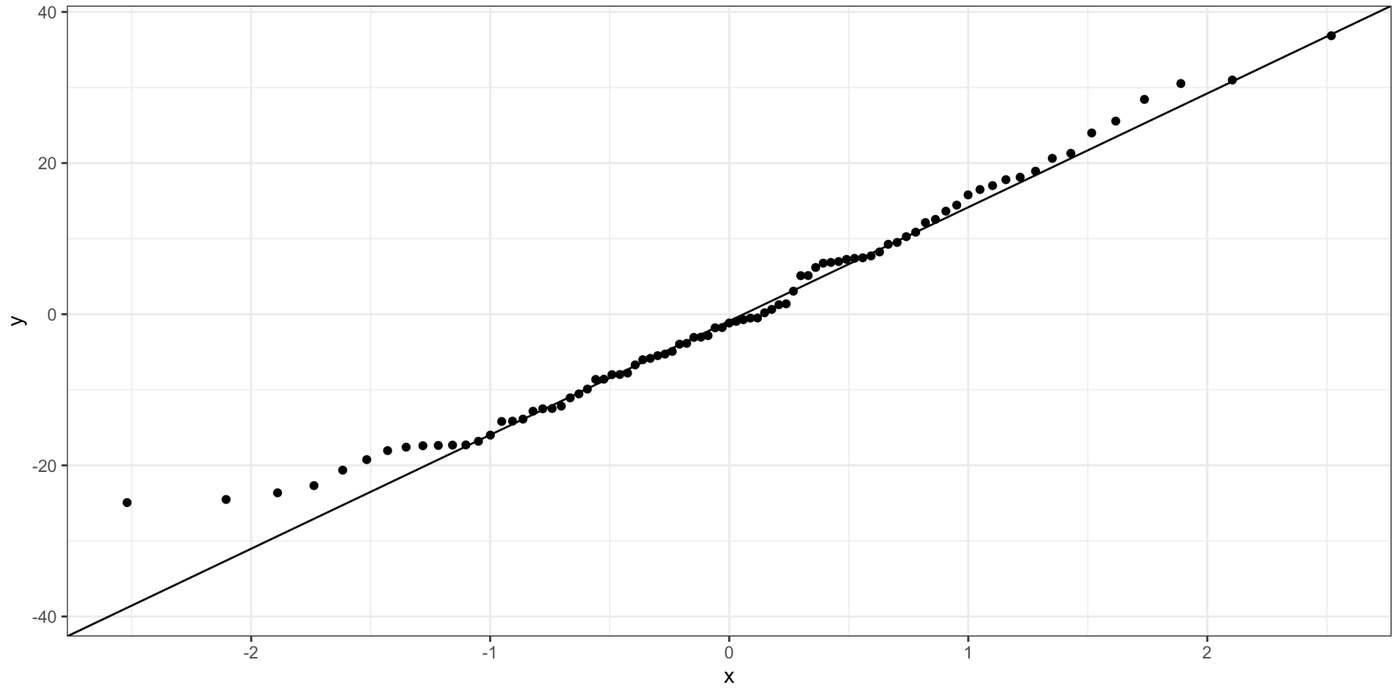
Checking normality: residual histogram

```
ggplot(res, aes(x = residual)) +
  geom_histogram(bins = 30) +
  labs(x = "Residual")
```



Checking normality: Q-Q plot

```
1 ggplot(res, aes(sample = residual)) +
2   geom_qq() +
3   geom_qq_line()
```



- Normality: ❓

Suppose LINE assumptions are met. Is the model good?

Coefficient of Determination, or R^2

A common measure of the **strength** of a linear model is the **coefficient of determination** R^2, (aka "R-squared").

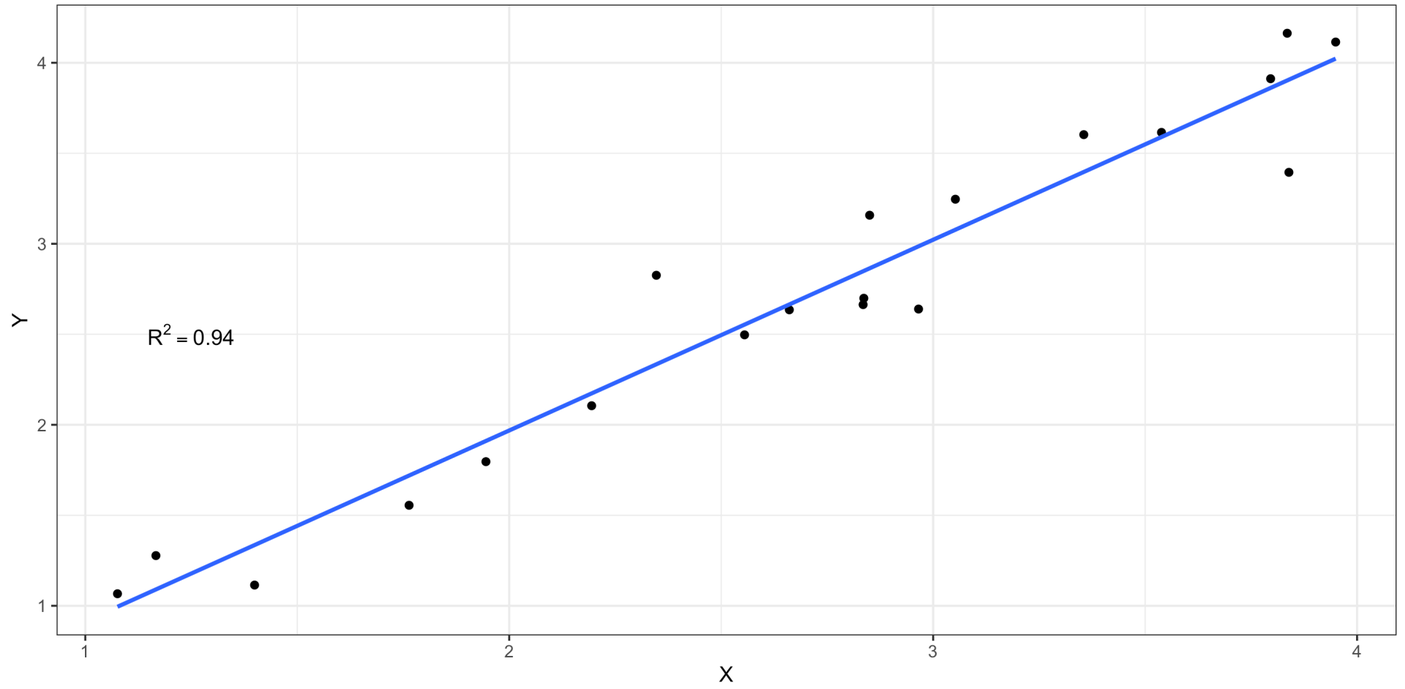
- R^2 is always between 0 and 1 (for the linear models we discuss)
- It measures the **proportion of variation in the response variable y that is explained by the linear model**

$$R^2 = \frac{\overbrace{s_y^2 - s_e^2}^{\text{Variation in Y explained by X}}}{\underbrace{s_y^2}_{\text{Variation in y}}}$$

(Reminder: Variance = (Standard Deviation)^2)

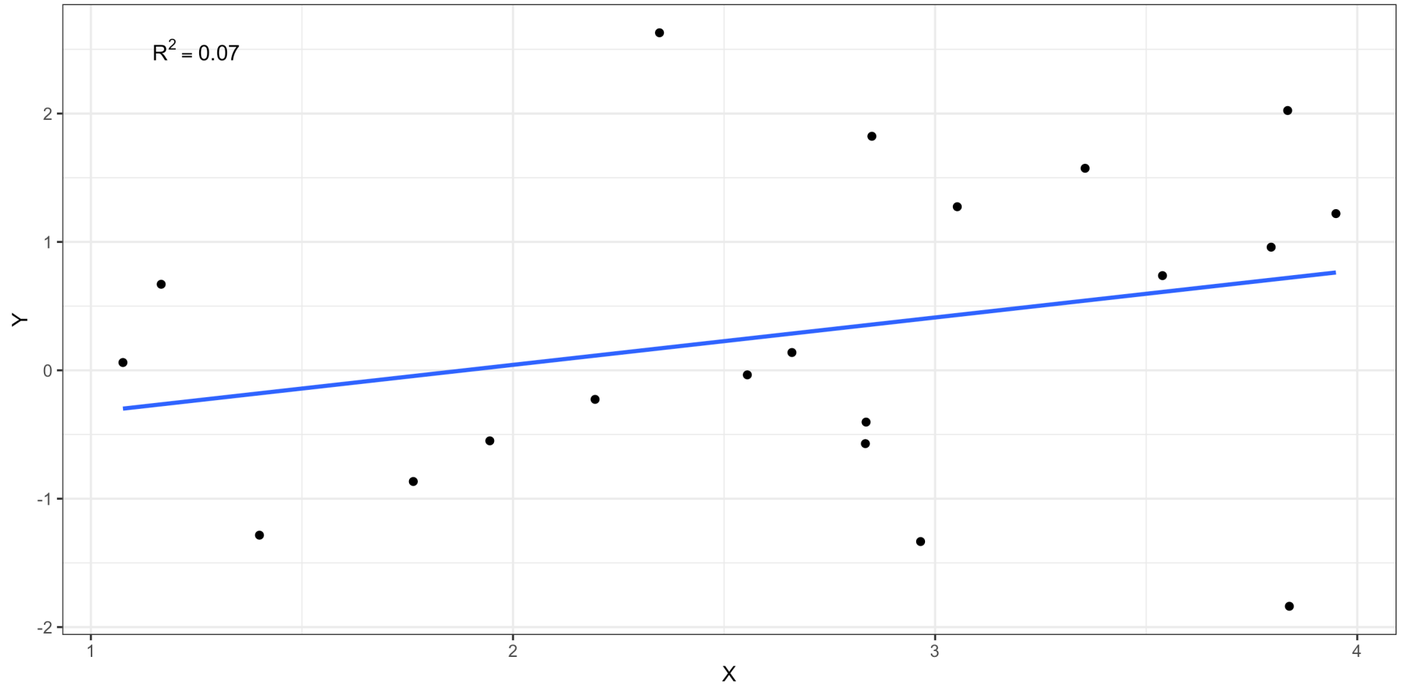
Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is **explained by** variability in the explanatory variable.



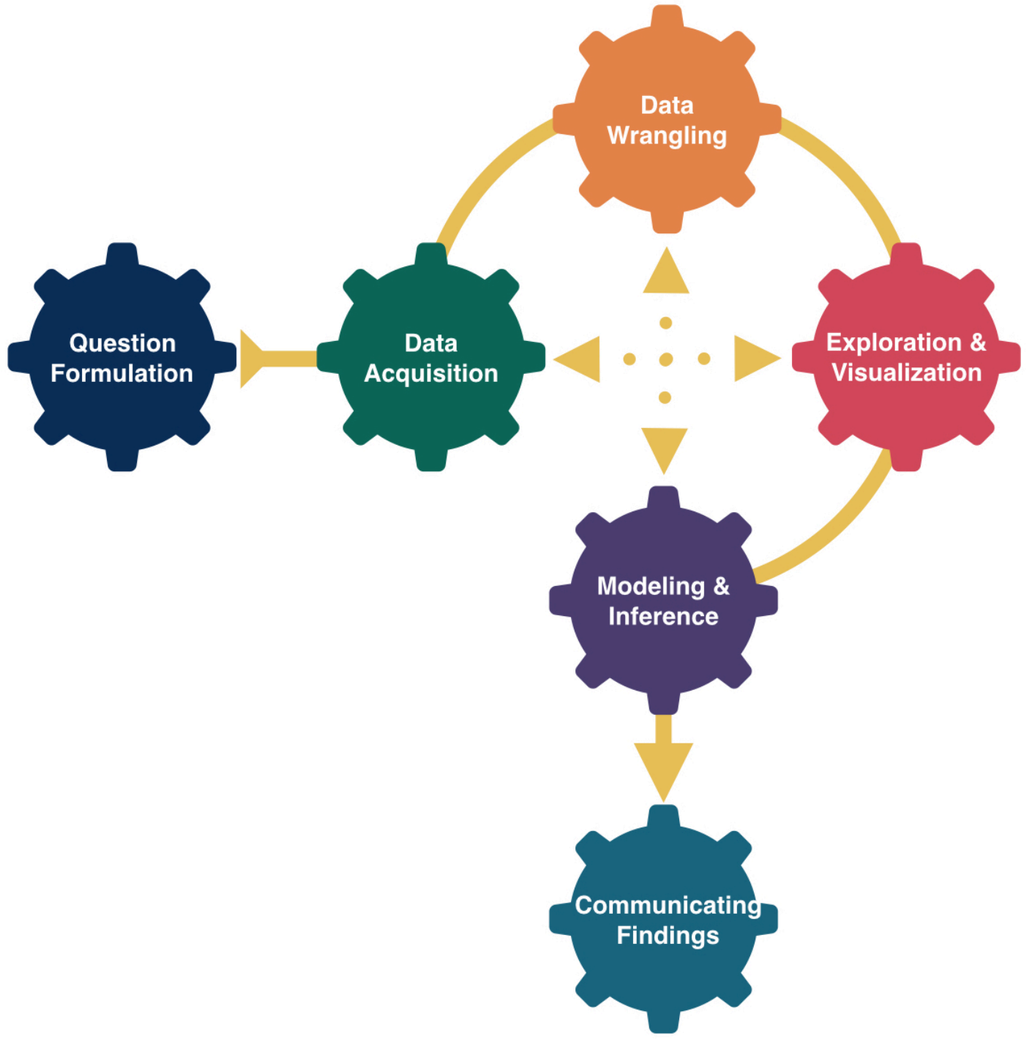
Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is **explained by** variability in the explanatory variable.



Next Time

- Regression with one categorical variable



Linear Models III: Categorical Predictors

Megan Ayers

Math 141 | Spring 2026

Friday, Week 4

Reminders/Announcements

- Please fill out the Week 4 feedback survey (link in Slack)
- Wednesday: we will have a **short** and **completion-based** learning check about coefficient interpretation. Please arrive on time!

Goals for Today

- Recap: Simple linear regression model
- Broadening our idea of linear regression
- Regression with a single, binary categorical explanatory variable
- Regression with a single categorical explanatory variable with more than 2 levels

Simple Linear Regression

So far we've considered this model when:

- Response variable (y): quantitative
- Explanatory variable (x): quantitative
  - Have only ONE explanatory variable.
- AND, $f()$ can be approximated by a line:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Linear Regression

Linear regression is a flexible class of models that allow for:

- Both quantitative and categorical **explanatory** variables.
- **Multiple** explanatory variables.
- Where the **response variable is quantitative**.

Activity: Brunch Wait Times

You're planning when to head to brunch, and want to understand how long you should expect to wait for a table. Info from your previous experiences is below:

- **Q1:** How long did you typically wait for a table?
- **Q2:** You think wait time varies by arrival time. Calculate average wait by arrival time (early or late).
- **Q3:** How much longer did you wait when you arrived late rather than early, on average?
- **Q4:** Q3 can be re-framed as a simple linear regression! What are the explanatory and response variables? How is this different from regressions we've seen so far?

Activity: Brunch Wait Times

The simplest model would predict wait time using a constant value:

$$y = \beta_0 + \epsilon$$

- **Q**: What $\hat{\beta}_0$ would minimize the sum of the squared residuals?
- **A**: $\hat{\beta}_0 = \bar{y}$, the sample mean!

We can make a slightly more complicated model:

$$y = \beta_0 + \beta_1 x_{(\text{arrived late})} + \epsilon$$

where $x_{(\text{arrived late})}$ is either 0 or 1.

$$\widehat{\text{Wait time}} = 5 + 10 * x_{(\text{arrived late})}$$

Linear Models with a Categorical Explanatory Variable with 2 Levels

- Response variable (y): quantitative
- Have 1 categorical explanatory variable (w) with two categories
- y is quantitative: so we need to convert w into a numeric variable. Call this x, taking either the value 0 or 1.

y	w	x
10	level A	0
15	level A	0
25	level B	1
7	level A	0
20	level B	1

- Model form:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- We often refer to categorical variables in linear models as **factors**
- **Think of x like a switch**: our prediction changes depending on whether it's turned on ($x = 1$ when $w = \text{level B}$) or turned off ($x = 0$ when $w = \text{level A}$)

Example: Halloween Candy

```
1 candy <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/candy-data.csv")
2 glimpse(candy)
```

```
Rows: 85
Columns: 13
$ competitorname   <chr> "100 Grand", "3 Musketeers", "One dime", "One quarter…
$ chocolate        <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
$ fruity           <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1,…
$ caramel          <dbl> 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,…
$ peanutyalmondy   <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ nougat           <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
$ crispedricewafer <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ hard             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1,…
$ bar              <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
$ pluribus         <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1,…
$ sugarpercent     <dbl> 0.732, 0.604, 0.011, 0.011, 0.906, 0.465, 0.604, 0.31…
$ pricepercent     <dbl> 0.860, 0.511, 0.116, 0.511, 0.511, 0.767, 0.767, 0.51…
```

What might be a good categorical explanatory variable of `winpercent`?

Exploratory Data Analysis

Before building the model, let's explore and visualize the data!

- **Q**: What `dplyr` functions should we use to find the mean and sd of `winpercent` by the categories of `chocolate`?
- **Q**: What graph should we use to visualize the `winpercent` scores by `chocolate`?

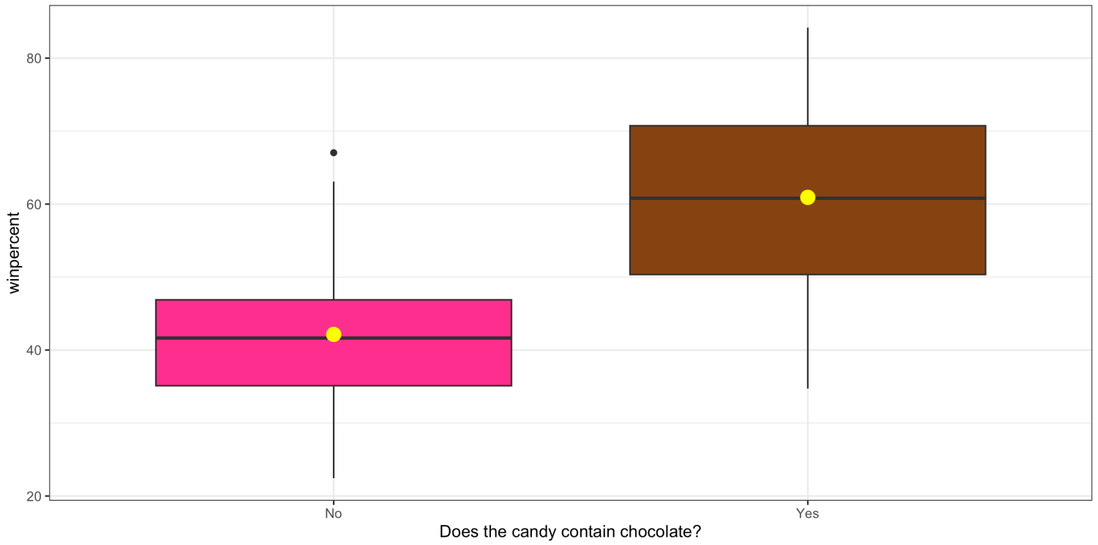
Exploratory Data Analysis

```r
# Summarize
candy %>%
  group_by(chocolate) %>%
  summarize(count = n(),
            mean_win = mean(winpercent),
            sd_win = sd(winpercent))
```

```
# A tibble: 2 × 4
  chocolate count mean_win sd_win
      <dbl> <int>    <dbl>  <dbl>
1         0    48     42.1   10.2
2         1    37     60.9   12.8
```

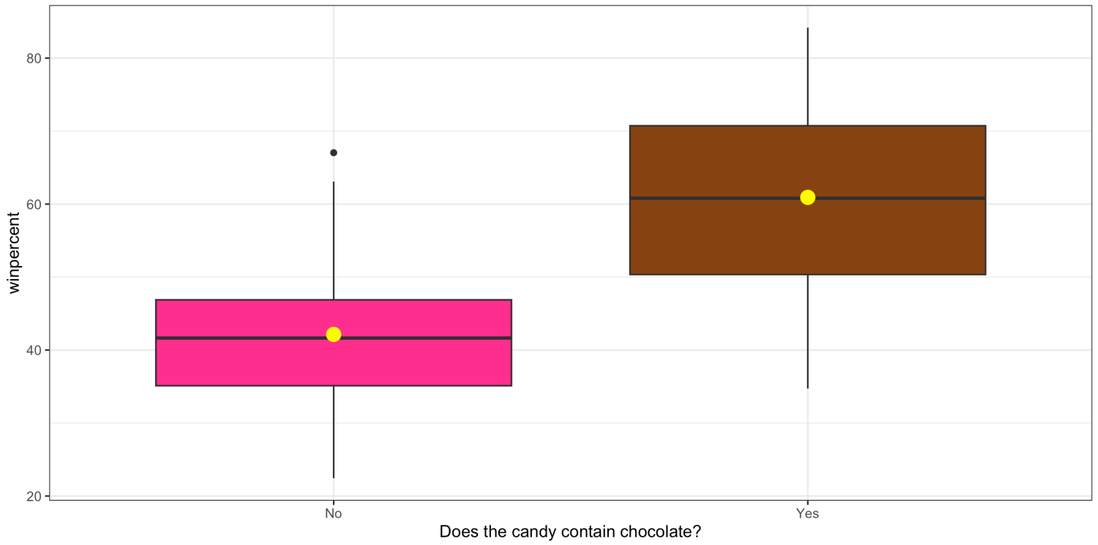
Exploratory Data Analysis

```
ggplot(candy, aes(x = factor(chocolate),
                  y = winpercent,
                 fill = factor(chocolate))) +
  geom_boxplot() +
  stat_summary(fun = mean,
               geom = "point",
               color = "yellow",
               size = 4) +
  guides(fill = "none") +
  scale_fill_manual(values =
                      c("0" = "deeppink",
                        "1" = "chocolate4")) +
  scale_x_discrete(labels = c("No", "Yes"),
                   name =
          "Does the candy contain chocolate?")
```



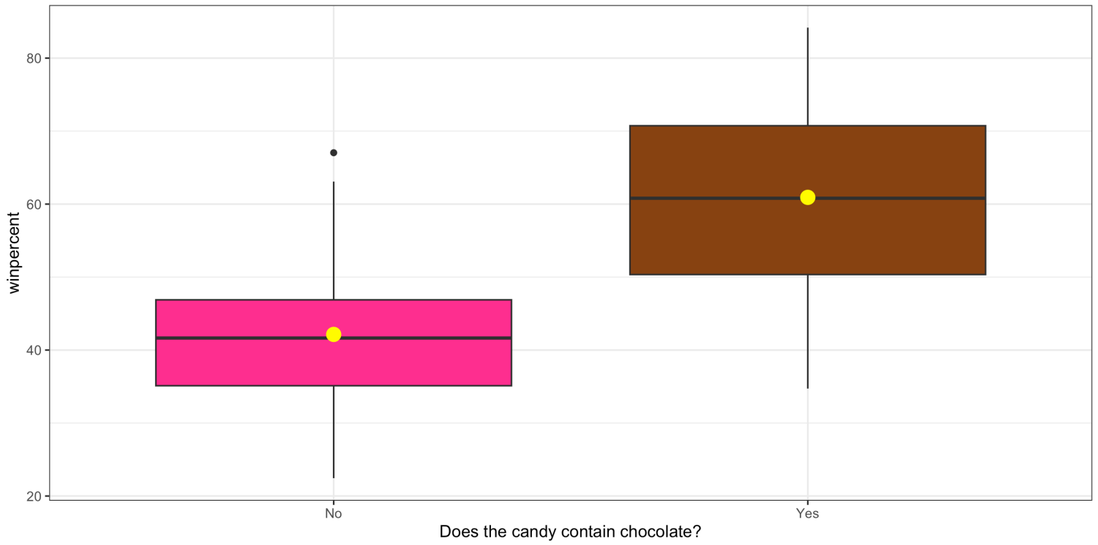
Exploratory Data Analysis

```
ggplot(candy, aes(x = factor(chocolate),
                  y = winpercent,
                 fill = factor(chocolate))) +
  geom_boxplot() +
  stat_summary(fun = mean,
               geom = "point",
               color = "yellow",
               size = 4) +
  guides(fill = "none") +
  scale_fill_manual(values =
                      c("0" = "deeppink",
                        "1" = "chocolate4")) +
  scale_x_discrete(labels = c("No", "Yes"),
                   name =
          "Does the candy contain chocolate?")
```



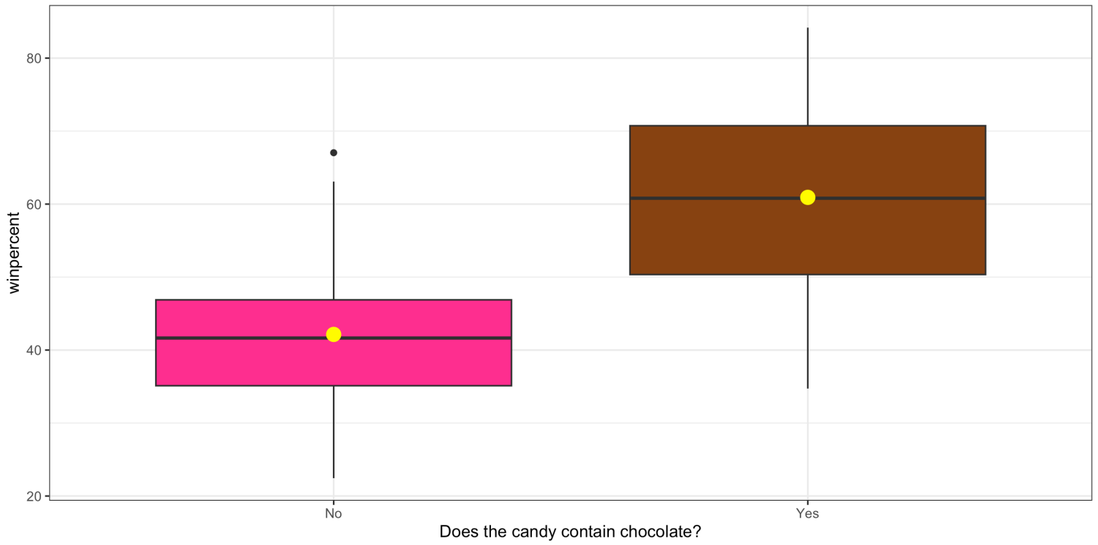
Exploratory Data Analysis

```
ggplot(candy, aes(x = factor(chocolate),
                  y = winpercent,
                 fill = factor(chocolate))) +
  geom_boxplot() +
  stat_summary(fun = mean,
               geom = "point",
               color = "yellow",
               size = 4) +
  guides(fill = "none") +
  scale_fill_manual(values =
                      c("0" = "deeppink",
                        "1" = "chocolate4")) +
  scale_x_discrete(labels = c("No", "Yes"),
                   name =
          "Does the candy contain chocolate?")
```



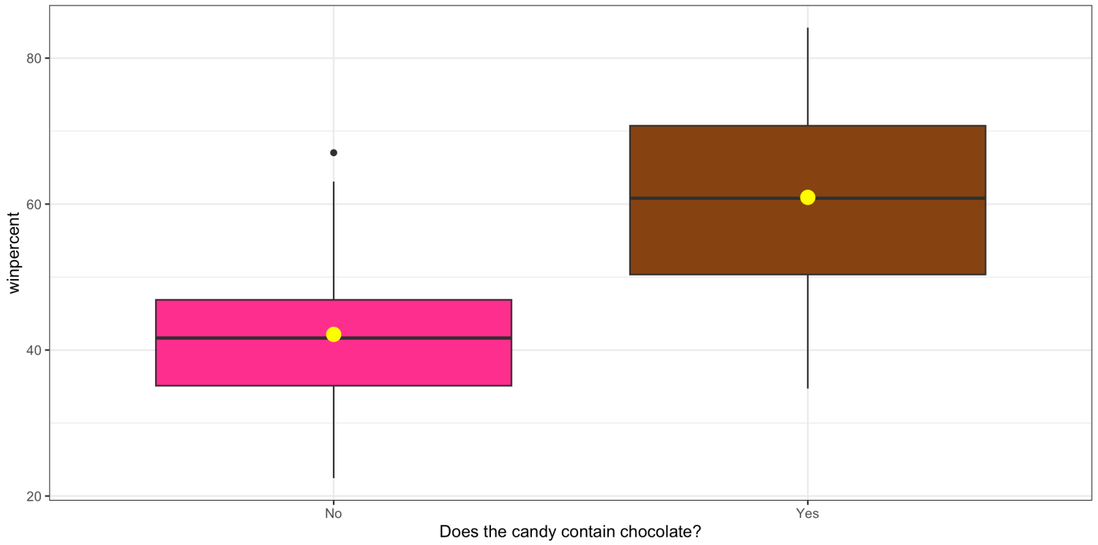
Exploratory Data Analysis

```
ggplot(candy, aes(x = factor(chocolate),
                  y = winpercent,
                 fill = factor(chocolate))) +
  geom_boxplot() +
  stat_summary(fun = mean,
               geom = "point",
               color = "yellow",
               size = 4) +
  guides(fill = "none") +
  scale_fill_manual(values =
                      c("0" = "deeppink",
                        "1" = "chocolate4")) +
  scale_x_discrete(labels = c("No", "Yes"),
                   name =
          "Does the candy contain chocolate?")
```



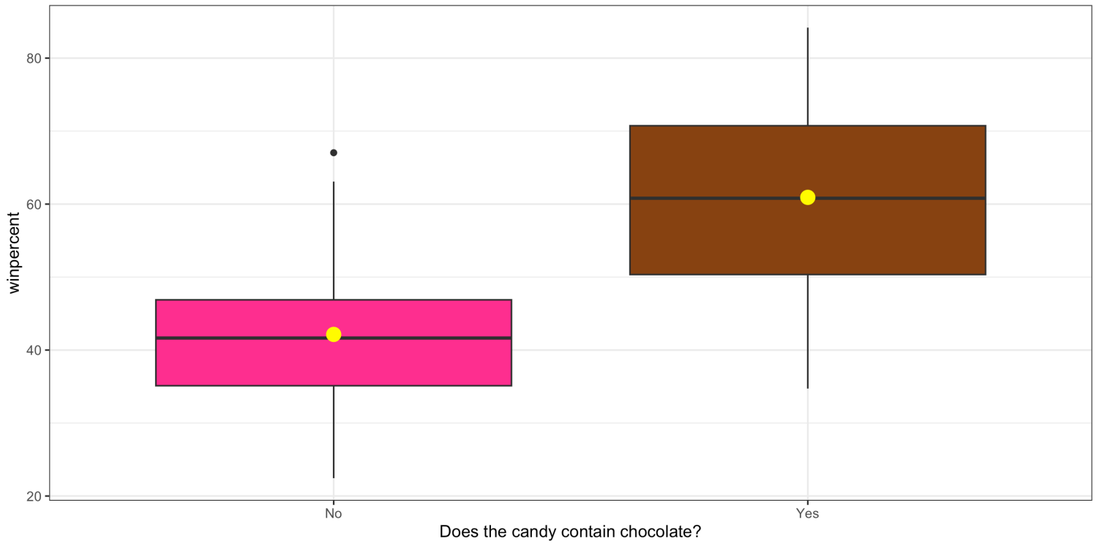
Exploratory Data Analysis

```
ggplot(candy, aes(x = factor(chocolate),
                  y = winpercent,
                 fill = factor(chocolate))) +
  geom_boxplot() +
  stat_summary(fun = mean,
               geom = "point",
               color = "yellow",
               size = 4) +
  guides(fill = "none") +
  scale_fill_manual(values =
                      c("0" = "deeppink",
                        "1" = "chocolate4")) +
  scale_x_discrete(labels = c("No", "Yes"),
                   name =
          "Does the candy contain chocolate?")
```



Exploratory Data Analysis

```
ggplot(candy, aes(x = factor(chocolate),
                  y = winpercent,
                 fill = factor(chocolate))) +
  geom_boxplot() +
  stat_summary(fun = mean,
               geom = "point",
               color = "yellow",
               size = 4) +
  guides(fill = "none") +
  scale_fill_manual(values =
                      c("0" = "deeppink",
                        "1" = "chocolate4")) +
  scale_x_discrete(labels = c("No", "Yes"),
                   name =
          "Does the candy contain chocolate?")
```



Fit the Linear Regression Model

Model Form:

$$y = \beta_0 + \beta_1 x + \epsilon$$

When $x = 0$:

When $x = 1$:

```
mod <- lm(winpercent ~ chocolate, data = candy)
library(moderndive)
get_regression_table(mod)
```

```
# A tibble: 2 × 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept     42.1      1.65     25.6        0     38.9     45.4
2 chocolate     18.8      2.50      7.52       0     13.8     23.7
```

Notes

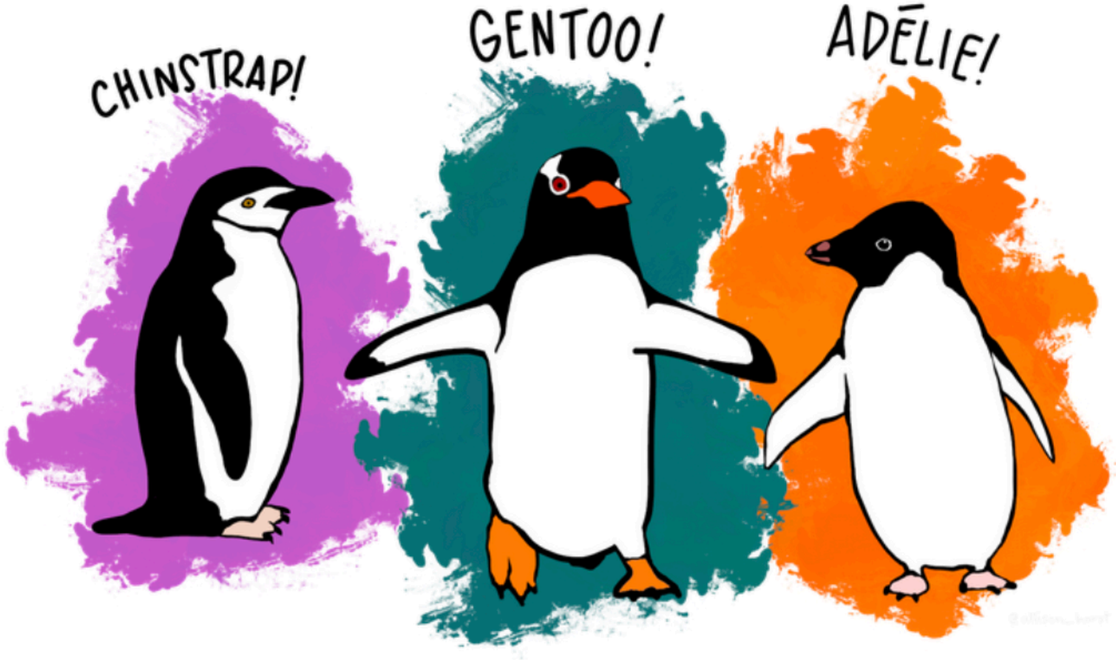
- When the explanatory variable is categorical, β_0 and β_1 no longer represent the intercept and slope.
- Now β_0 represents the (population) mean of the response variable when $x = 0$.
- And, β_1 represents the change in the (population) mean response going from $x = 0$ to $x = 1$.
- Can also do prediction:

```r
new_candy <- data.frame(chocolate = c(0, 1))
predict(mod, newdata = new_candy)
```

```
       1        2
42.14226 60.92153
```

New example: Palmer Penguins

```
1 library(palmerpenguins)
```



The Palmer Archipelago penguins. Artwork by @allison_horst.

Take a look at the data

```
1 glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel…
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse…
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, …
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, …
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186…
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, …
$ sex               <fct> male, female, female, NA, female, male, female, male…
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007…
```

We'd like to predict a penguin's bill length based on their species.

Response variable?

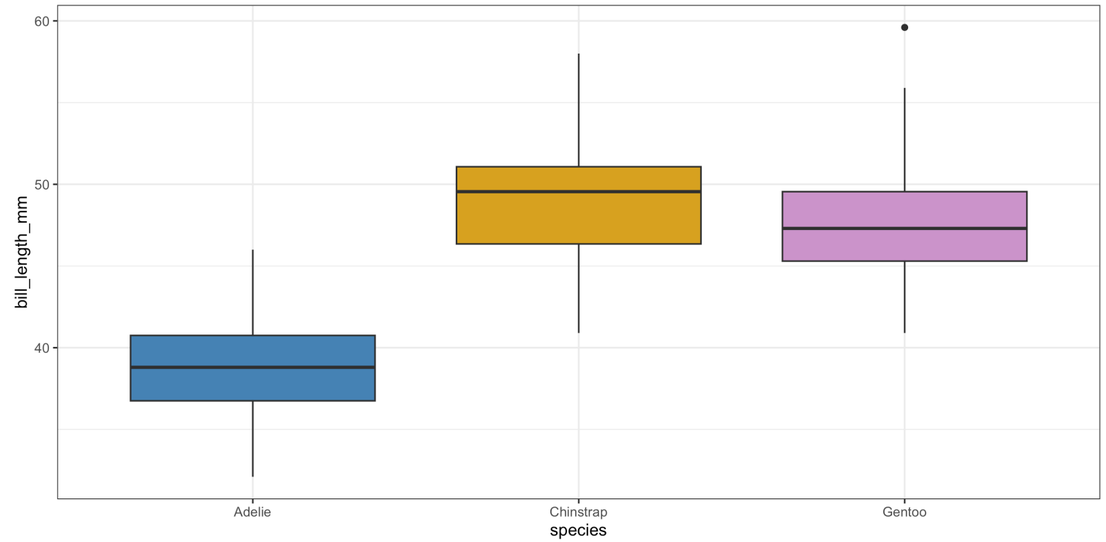
Explanatory variable?

Exploratory data analysis

```r
penguins %>%
  group_by(species) %>%
  summarize(
    avg_bill_length = mean(bill_length_mm,
                           na.rm = TRUE)
    )
```

```
# A tibble: 3 × 2
  species   avg_bill_length
  <fct>               <dbl>
1 Adelie               38.8
2 Chinstrap            48.8
3 Gentoo               47.5
```

```r
ggplot(penguins,
       aes(x = species,
           y = bill_length_mm,
           fill = species)) +
  geom_boxplot() +
  scale_fill_manual(values = c("steelblue",
                               "goldenrod",
                               "plum3")) +
  guides(fill = "none") +
  theme_bw()
```



How do we handle more than 2 groups???

Boardwork

Fit the model in R

$$y = \beta_0 + \beta_1 x_{species:Chinstrap} + \beta_2 x_{species:Gentoo} + \epsilon$$

R automatically makes species indicators for us and chooses a **reference level**.

```
penguin_mod <- lm(bill_length_mm ~ species, penguins)
get_regression_table(penguin_mod)
```

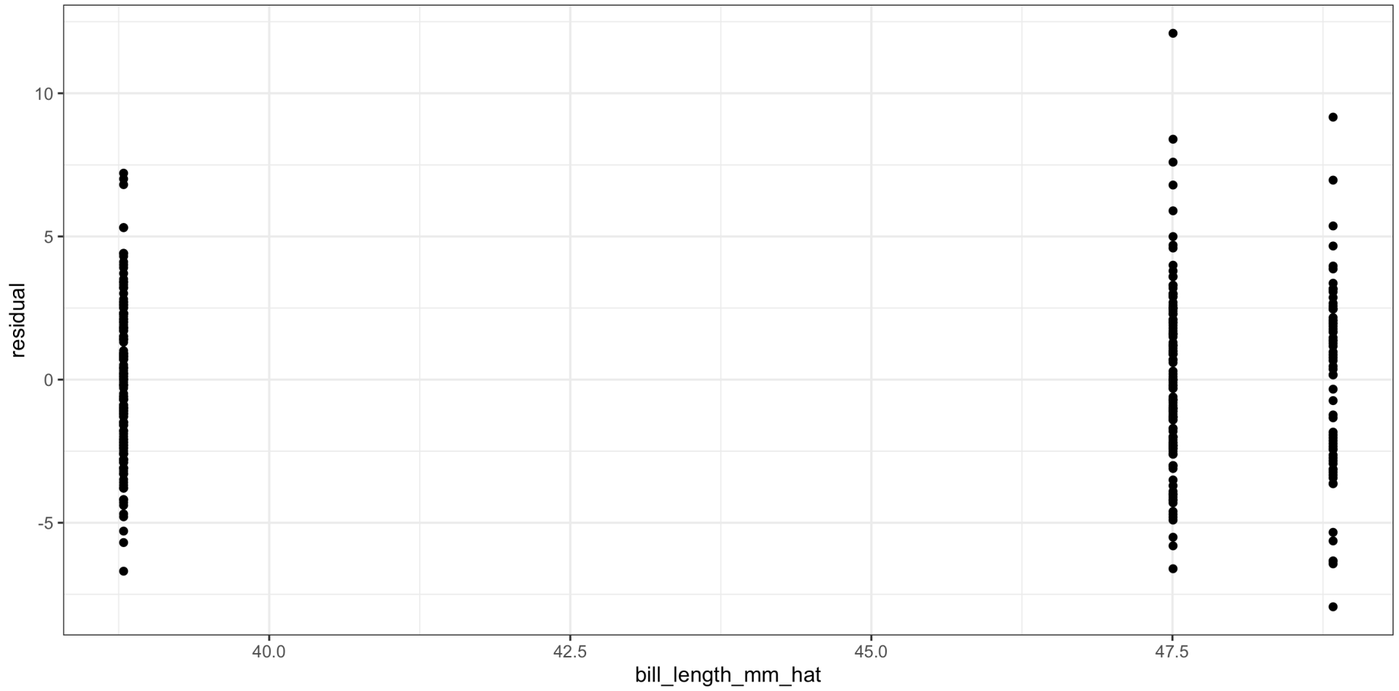
```
# A tibble: 3 × 7
  term              estimate std_error statistic p_value lower_ci upper_ci
  <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept             38.8     0.241     161.        0    38.3     39.3
2 species: Chinstrap    10.0     0.432      23.2       0     9.19    10.9
3 species: Gentoo        8.71    0.36       24.2       0     8.01     9.42
```

- **Q:** What is the equation for the fitted model?

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{species:Chinstrap} + \hat{\beta}_2 \cdot x_{species:Gentoo} \\ &= 38.8 + 10.0 \cdot x_{species:Chinstrap} + 8.71 \cdot x_{species:Gentoo}\end{aligned}$$

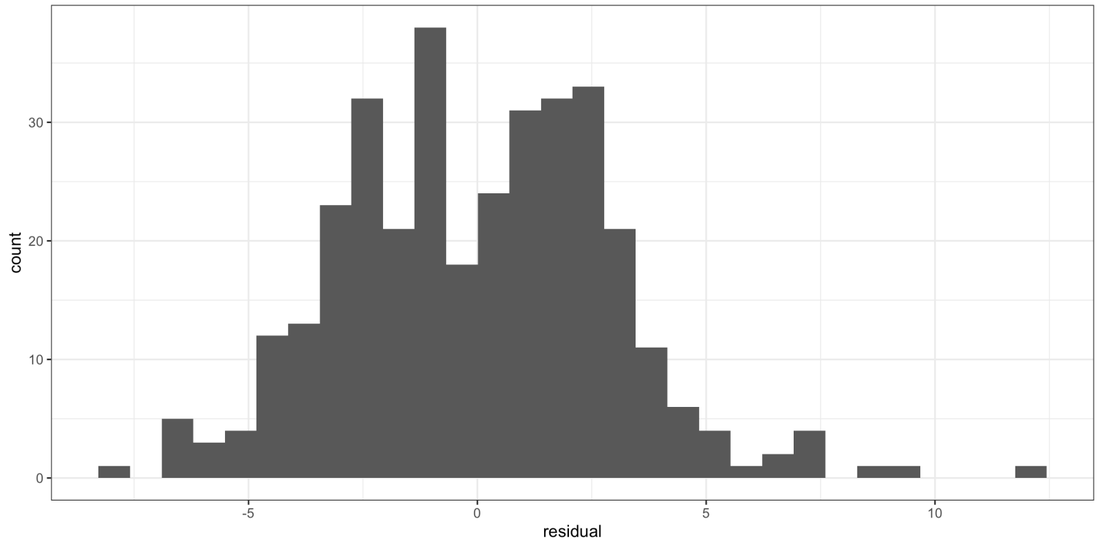
Remember to diagnose your models!

```
1 library(moderndive)
2 res <- get_regression_points(penguin_mod)
3 ggplot(res, aes(x = bill_length_mm_hat, y = residual)) +
4   geom_point()
```

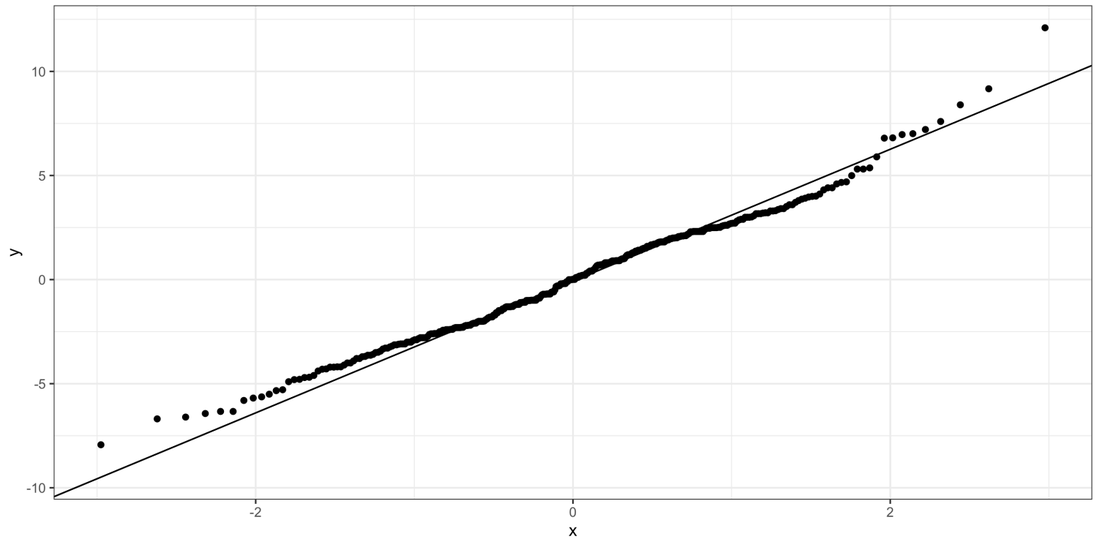


Remember to diagnose your models!

```
1 ggplot(res, aes(x = residual)) +
2   geom_histogram()
```




```
1 ggplot(res, aes(sample = residual)) +
2   geom_qq() +
3   geom_qq_line()
```



Multiple Linear Regression: A peak into next week

Recall our penguin model

$$y = \beta_0 + \beta_1 x_{species:Chinstrap} + \beta_2 x_{species:Gentoo} + \epsilon$$

Even though we are using one predictor (species), we now have β_0, β_1, **and** β_2!

- We **recoded** the species predictor into two binary predictors
- We are actually doing **multiple** linear regression now
- **Next week**: We'll formalize and extend multiple linear regression

- Categorical explanatory variables are tricky! Revisit these examples, more practice with HW 04 Exercise 3

Activity: Changing Reference Level

We can change the **reference level** if we want.

```
1 penguins$species <- relevel(penguins$species, ref = "Chinstrap")
2 penguin_mod <- lm(bill_length_mm ~ species, penguins)
3 get_regression_table(penguin_mod)
```

```
# A tibble: 3 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept          48.8      0.359    136.     0        48.1    49.5
2 species: Adelie   -10.0      0.432    -23.2    0       -10.9    -9.19
3 species: Gentoo    -1.33     0.447     -2.97   0.003    -2.21   -0.449
```

Get in groups of ~5 and find a spot on one of the boards.

- **Q1:** Write down the equation for this fitted model.
- **Q2:** Draw a few rows of the data frame with the variables used by the model.
- **Q3:** How is the interpretation of β_0 and $\hat{\beta}_0$ different now?
- **Q4:** How should we interpret the other two coefficients?

Activity: Changing Reference Level (Answers)

- **Q1:** Write down the equation for this fitted model.

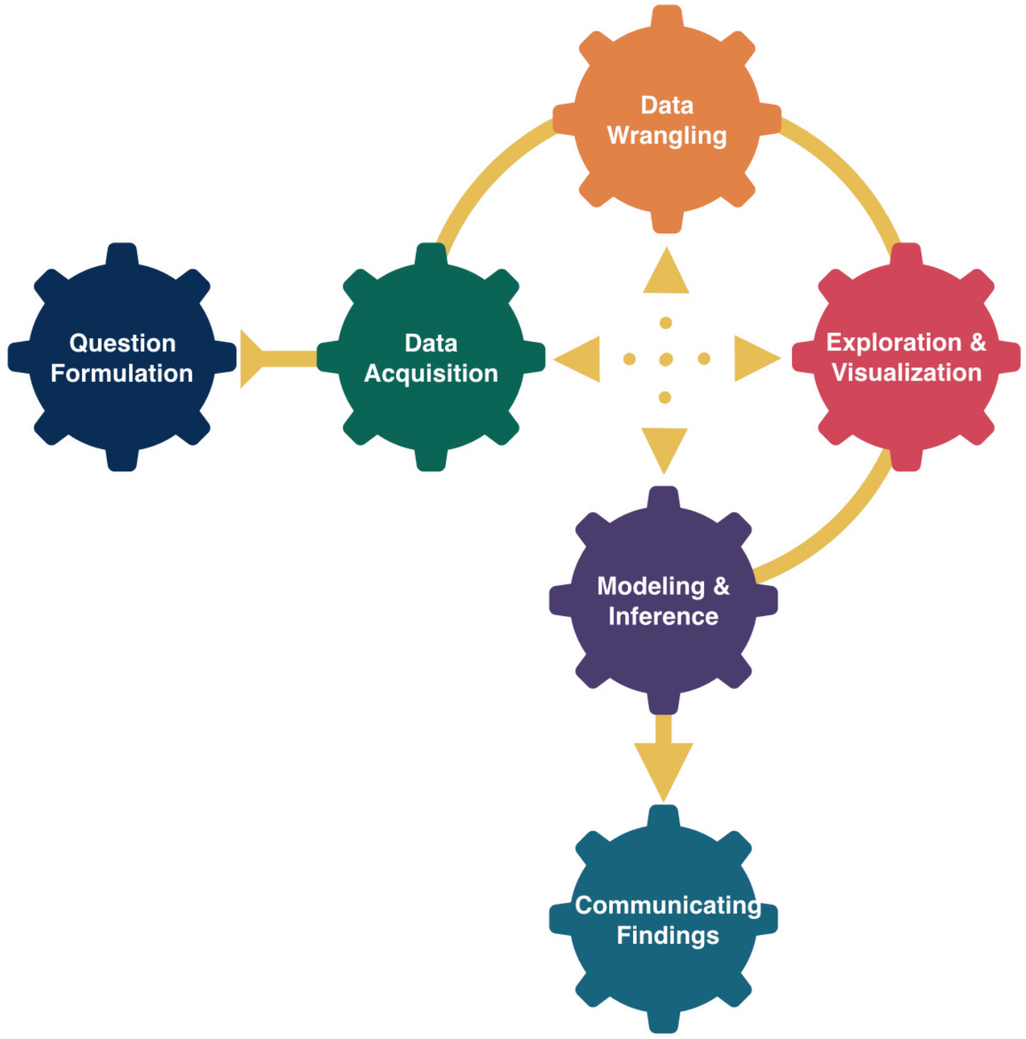
$$\widehat{Y} = 48.8 - 10.0x_{species:Adelie} - 1.3x_{species:Gentoo}$$

- **Q2:** Draw a few rows of the data frame with the variables used by the model.

```
# A tibble: 6 × 4
# Groups:   species [3]
  bill_length_mm species   x_adelie x_gentoo
           <dbl> <fct>        <dbl>    <dbl>
1           50.3 Chinstrap        0        0
2           49.8 Chinstrap        0        0
3           33.1 Adelie           1        0
4           34.5 Adelie           1        0
5           49   Gentoo           0        1
6           48.1 Gentoo           0        1
```

- **Q3:** How is the interpretation of β_0 and $\hat{\beta}_0$ different now?
  - β_0 ($\hat{\beta}_0$) now represents (estimates) the population mean bill length for Chinstrap penguins, not Adelie penguins

- **Q4:** How should we interpret the other two coefficients?
  - They now represent the predicted change in bill lengths for Adelie and Gentoo penguins on average, relative to the predicted bill length for Chinstrap penguins.



Linear Models IV: Multiple Regression

Megan Ayers

Math 141 | Spring 2026

Monday, Week 5

Announcements

- Week 4 assessments
- Office hours changes

Goals for Today

- Handling categorical and quantitative explanatory variables at the same time.
- Compare parallel slopes and interaction models for multiple regression
- Transformations

Multiple Linear Regression

Form of the Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

How does extending to more predictors change our process?

- What **doesn't** change:
  - Still use **Method of Least Squares** to estimate coefficients
  - Still use `lm()` to fit the model and `predict()` for prediction
- What **does** change:
  - Meaning of the coefficients are more complicated and depend on other variables in the model
  - Need to decide which variables to include and how (linear term, squared term…)

Multiple Linear Regression

- We are going to see a few examples of multiple linear regression this week.
- We will need to return to modeling later in the course once we have learned about statistical inference (i.e., confidence intervals and p-values).

Example

Meadowfoam is a plant that grows in the Pacific Northwest and is harvested for its seed oil. In a randomized experiment, researchers at Oregon State University looked at how two light-related factors influenced the **number of flowers per meadowfoam plant**, the primary measure of productivity for this plant. The two light measures were **light intensity** (in mmol/ m^2 /sec) and the **timing of onset of the light** (early or late in terms of photo periodic floral induction).

Response variable?

Explanatory variables?

Model Form?

Data Loading and Wrangling

```
library(tidyverse)
library(Sleuth3)
data(case0901)

# Check out the timing variable
count(case0901, Time)
```

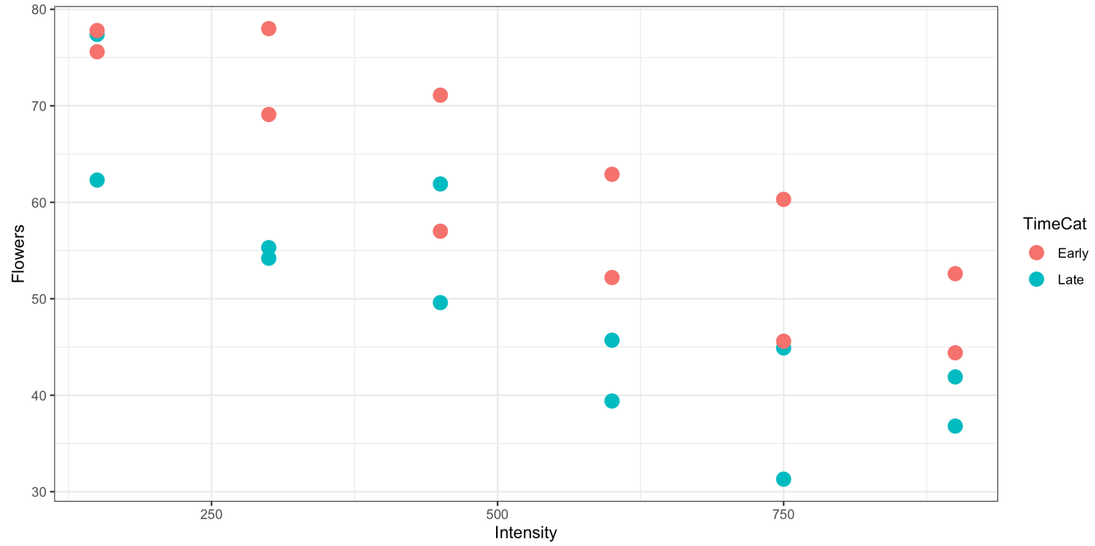
```
  Time  n
1    1 12
2    2 12
```

```
# Recode the timing variable
case0901 <- case0901 %>%
  mutate(TimeCat = case_when(Time == 1 ~ "Late",
                             Time == 2 ~ "Early"))
count(case0901, TimeCat)
```

```
  TimeCat  n
1   Early 12
2    Late 12
```

Visualizing the Data

```
ggplot(case0901,
       aes(x = Intensity,
           y = Flowers,
           color = TimeCat)) +
  geom_point(size = 4)
```



- **Q**: How does this plot help us intuit what $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ might be?
- Why don't I have to include `data =` and `mapping =` in my `ggplot()` layer?

Building the Linear Regression Model

Full model form:

```
1 modFlowers <- lm(Flowers ~ Intensity + TimeCat, data = case0901)
2
3 library(moderndive)
4 get_regression_table(modFlowers)
```

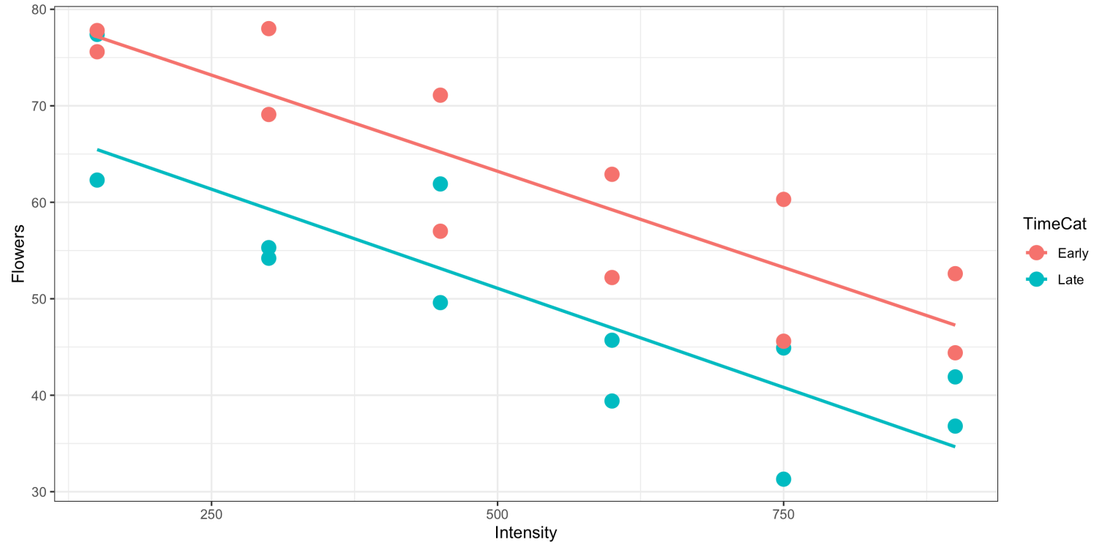
```
# A tibble: 3 × 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept         83.5      3.27      25.5        0   76.7      90.3
2 Intensity         -0.04     0.005     -7.89       0   -0.051    -0.03
3 TimeCat: Late    -12.2      2.63      -4.62       0  -17.6      -6.69
```

- Estimated regression line for **TimeCat = "Late"**

- Estimated regression line for **TimeCat = "Early"**:

Appropriateness of Model Form

```
ggplot(case0901,
       aes(x = Intensity,
           y = Flowers,
           color = TimeCat)) +
  geom_point(size = 4) +
  geom_smooth(method = "lm", se = FALSE)
```



Is the assumption of **equal slopes** reasonable here?

Prediction

```r
flowersNew <- data.frame(Intensity = c(700, 700), TimeCat = c("Early", "Late"))
flowersNew
```

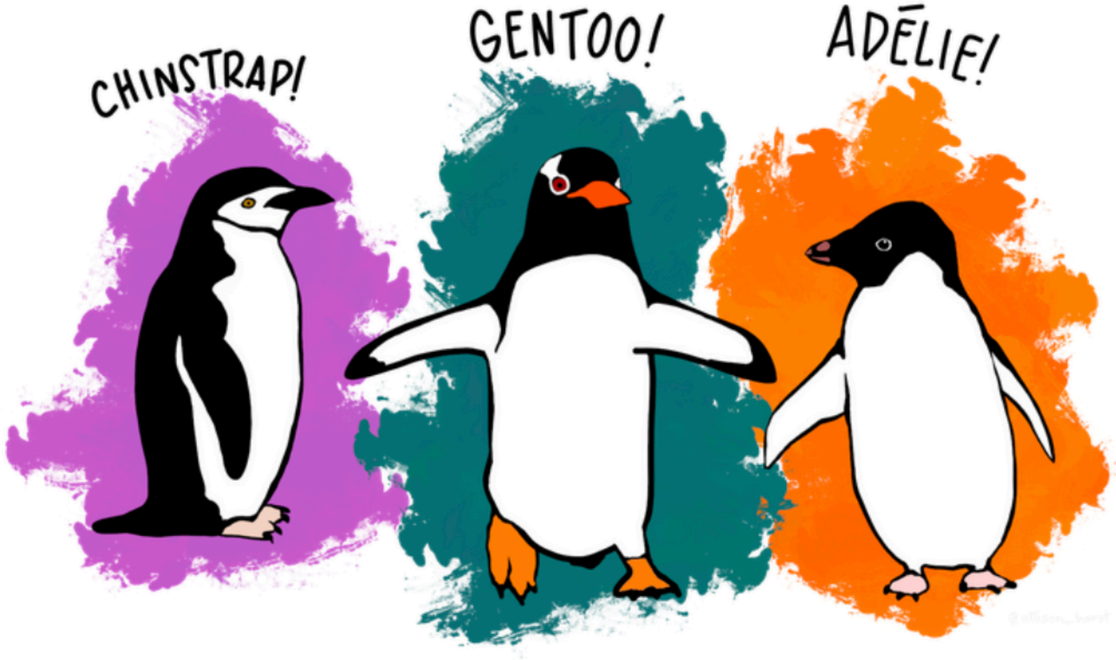
```
  Intensity TimeCat
1       700   Early
2       700    Late
```

```r
predict(modFlowers, newdata = flowersNew)
```

```
       1        2
55.13417 42.97583
```

Returning to the Palmer Penguins

```
1 library(palmerpenguins)
```

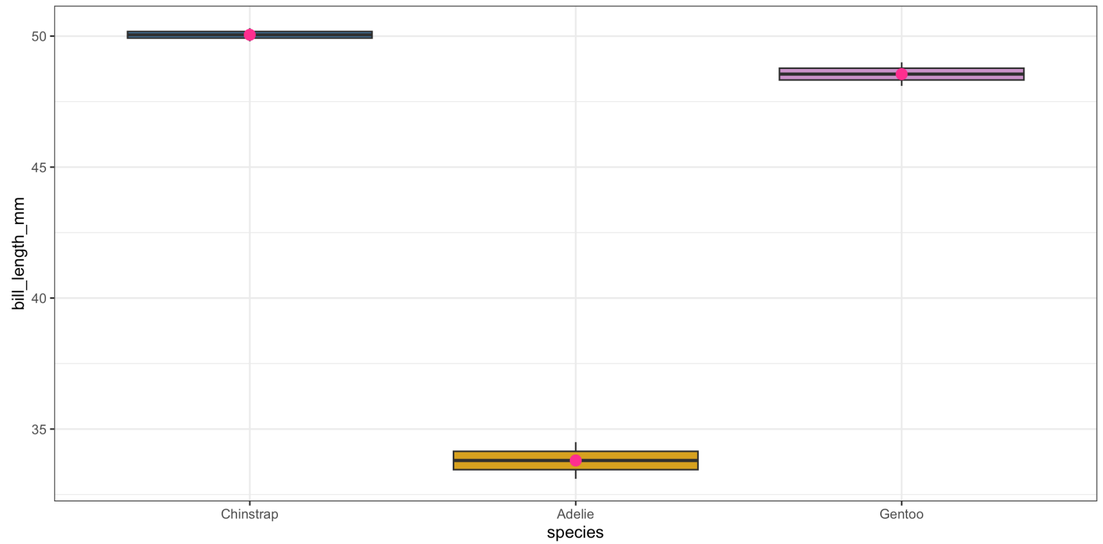


The Palmer Archipelago penguins. Artwork by @allison_horst.

Last time:

We predicted a penguin's bill length based on their species.

```
ggplot(penguins, aes(x = species, y = bill_length_mm, f
  geom_boxplot() +
  scale_fill_manual(values = c("steelblue",
                               "goldenrod",
                               "plum3")) +
  stat_summary(fun = mean,
               geom = "point",
               size = 3,
               color = "deeppink") +
  guides(fill = "none") +
  theme_bw()
```

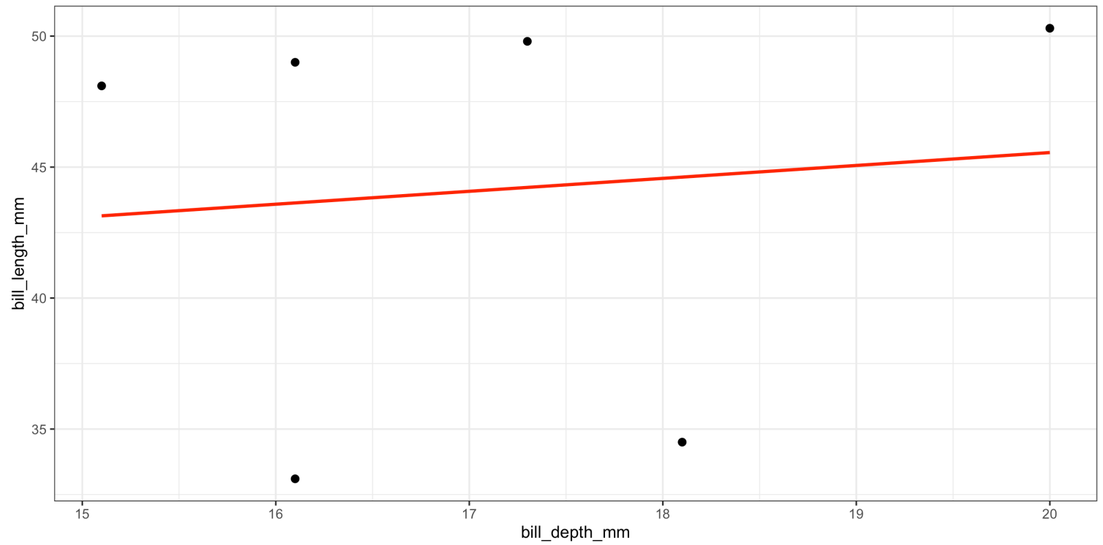


- Pink dots represent the mean value in each group.
- For the single categorical variable model, those pink dots are the predicted values for each group.

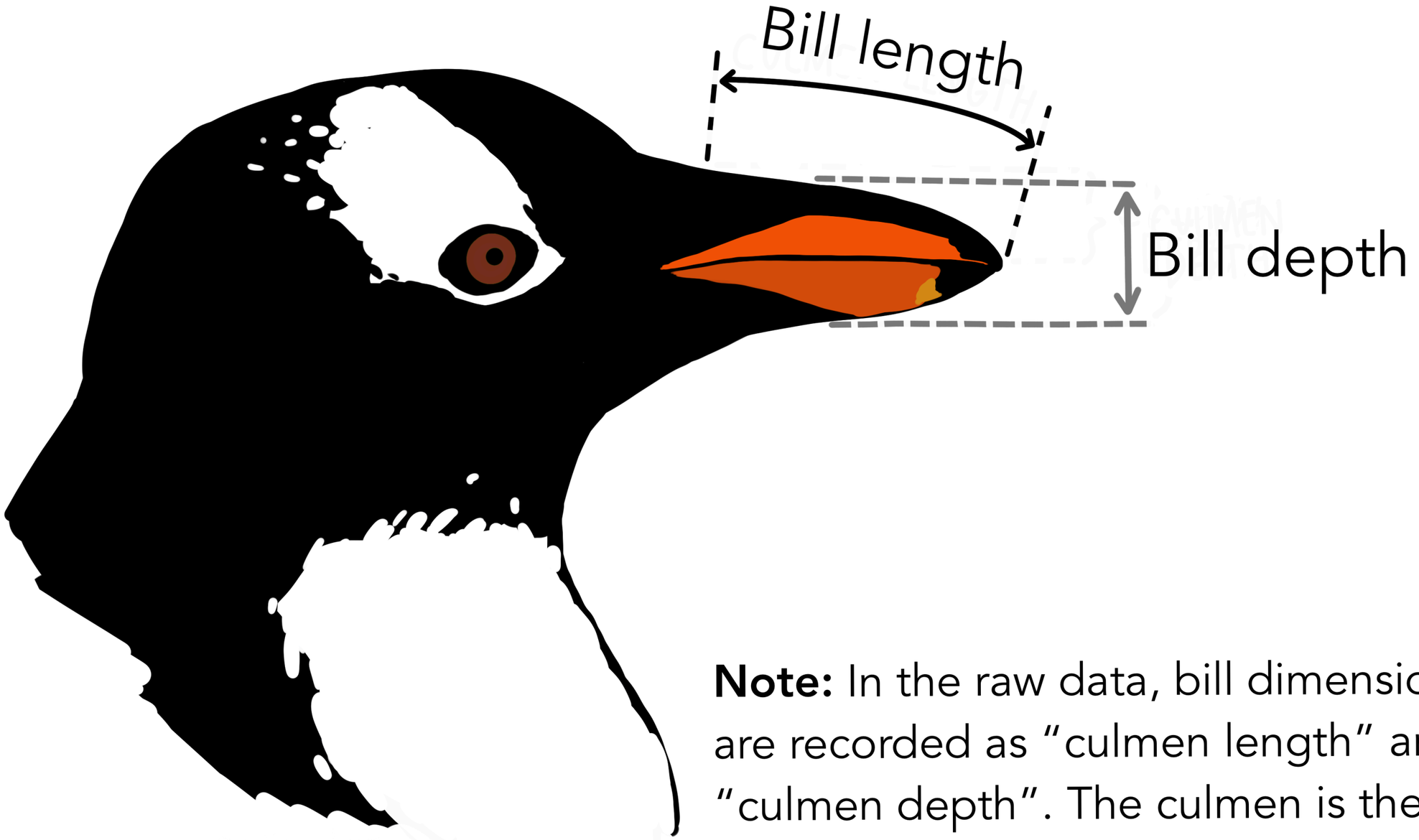
This time:

We'll incorporate bill depth for prediction!

```
1 ggplot(penguins, aes(x = bill_depth_mm, y = bill_length_
2   geom_point(size = 2) +
3   geom_smooth(method = "lm", se = FALSE, color = "red")
4   theme_bw()
```



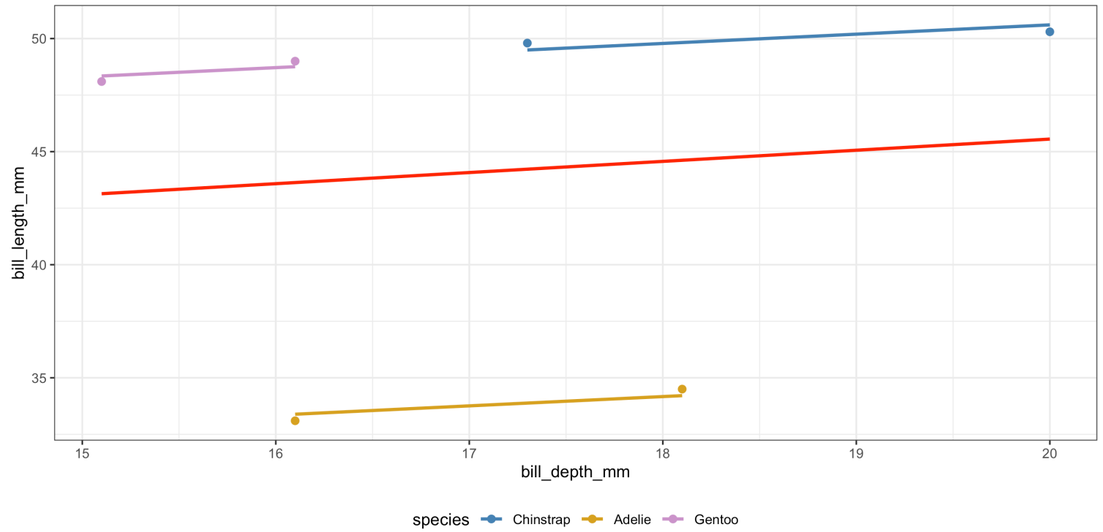
- A moderate negative relationship between bill length and bill depth!
- Does this make sense?



Note: In the raw data, bill dimensions are recorded as “culmen length” and “culmen depth”. The culmen is the dorsal ridge atop the bill.

What if we include both explanatory variables?

```
ggplot(penguins, aes(x = bill_depth_mm,
                     y = bill_length_mm,
                     color = species)) +
  geom_point(size = 2) +
  geom_smooth(inherit.aes = FALSE,
              mapping = aes(x = bill_depth_mm,
                            y = bill_length_mm),
              method = "lm", se = FALSE,
              color = "red") +
  geom_parallel_slopes(se = FALSE) +
  scale_color_manual(values = c("steelblue",
                                "goldenrod",
                                "plum3")) +
  theme_bw() +
  theme(legend.position = "bottom")
```



- **Negative** relationships between bill depth and bill length overall.
- **Positive** relationships between bill depth and bill length when accounting for species!
- **What is going on here??**
- This is a case of **Simpson's Paradox**.

Three candidate models

Explanatory variable: species

```
species_mod <- lm(bill_length_mm ~ species, penguins)
get_regression_table(species_mod) %>%
  select(term, estimate)
```

```
# A tibble: 3 × 2
  term            estimate
  <chr>              <dbl>
1 intercept           50.0
2 species: Adelie    -16.2
3 species: Gentoo     -1.5
```

Explanatory variable: bill_depth_mm

```
depth_mod <- lm(bill_length_mm ~ bill_depth_mm, penguins
get_regression_table(depth_mod) %>%
  select(term, estimate)
```

```
# A tibble: 2 × 2
  term          estimate
  <chr>            <dbl>
1 intercept       35.7
2 bill_depth_mm    0.493
```

Explanatory variables: species and bill_depth_mm (equal slope)

```
both_mod <- lm(bill_length_mm ~ bill_depth_mm + species
get_regression_table(both_mod) %>%
  select(term, estimate)
```

```
# A tibble: 4 × 2
  term            estimate
  <chr>              <dbl>
1 intercept          42.4
2 bill_depth_mm       0.411
3 species: Adelie   -15.6
4 species: Gentoo    -0.247
```

- **Coefficient interpretations?**

Interpreting the Multiple Regression Equation

```
1 get_regression_table(both_mod)
```

```
# A tibble: 4 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept         42.4       3.67      11.5    0.007   26.6      58.2
2 bill_depth_mm      0.411     0.196      2.10   0.171   -0.433     1.25
3 species: Adelie  -15.6       0.573    -27.2    0.001  -18.1     -13.1
4 species: Gentoo   -0.247     0.771     -0.32   0.779   -3.56      3.07
```

$$\hat{y} = 13.2 + 1.39 \cdot x_{\text{Bill Depth}} + 9.94 \cdot x_{\text{Species:Chinstrap}} + 13.4 \cdot x_{\text{Species:Gentoo}}$$

Interpreting the Multiple Regression Equation

```
1 get_regression_table(both_mod)
```

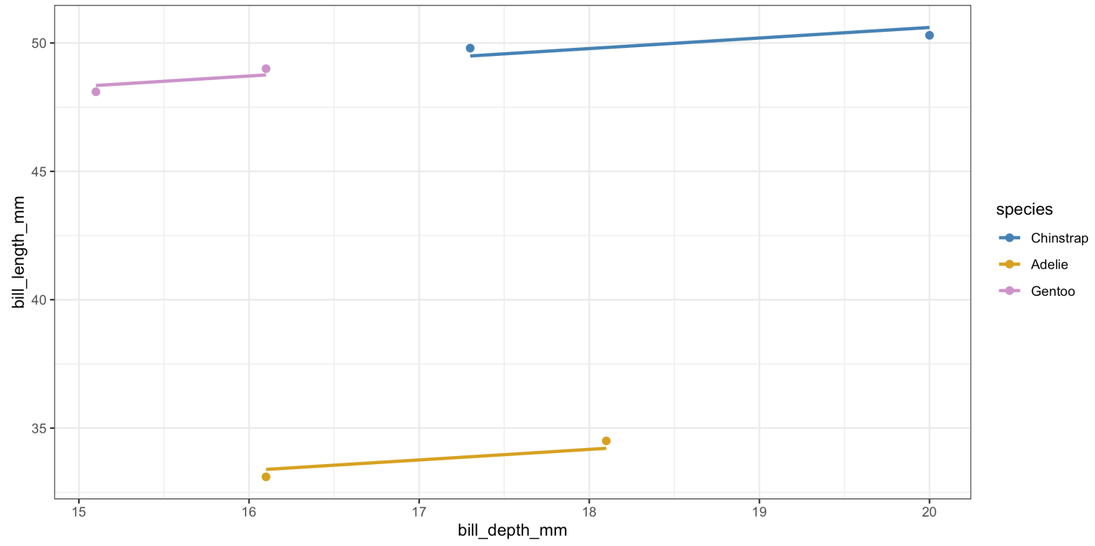
```
# A tibble: 4 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept         42.4       3.67      11.5    0.007   26.6      58.2
2 bill_depth_mm      0.411     0.196      2.10   0.171   -0.433     1.25
3 species: Adelie  -15.6       0.573    -27.2    0.001  -18.1     -13.1
4 species: Gentoo   -0.247     0.771     -0.32   0.779   -3.56      3.07
```

$$\hat{y} = 13.2 + 1.39 \cdot x_{\text{Bill Depth}} + 9.94 \cdot x_{\text{Species:Chinstrap}} + 13.4 \cdot x_{\text{Species:Gentoo}}$$

- **Intercept Coefficient:** The *expected/predicted* value of the response, *on average*, when all explanatory variables are *set to 0*.
  - What does setting species to 0 mean in this context?

Returning to our MLR model

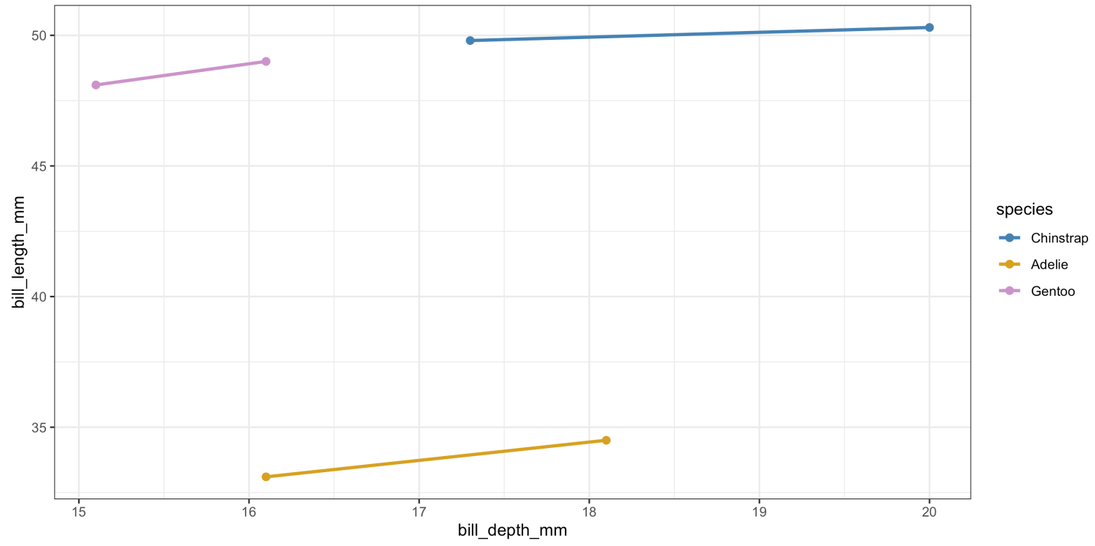
```
ggplot(penguins, aes(x = bill_depth_mm, y = bill_length_
  geom_point(size = 2) +
  geom_parallel_slopes(se = FALSE) +
  scale_color_manual(values = c("steelblue",
                                "goldenrod",
                                "plum3")) +
  theme_bw()
```



Are equal slopes a reasonable assumption here?

Different Slopes Model

```
ggplot(penguins, aes(x = bill_depth_mm, y = bill_length_
  geom_point(size = 2) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(values = c("steelblue",
                                "goldenrod",
                                "plum3")) +
  theme_bw()
```



- **Equal slopes** models force the relationship between quantitative predictors and the response variable to be the same for each group in the model.
- In contrast, **different slopes** models allow for **different relationships** between quantitative predictors and the response variable for each group in the model.
- How can we allow our model to have different slopes?

Contrasting model forms

Recall the equal slopes model:

$$y = \beta_0 + \beta_1 \cdot x_{\text{Bill Depth}} + \beta_2 \cdot x_{\text{Species:Chinstrap}} + \beta_3 \cdot x_{\text{Species:Gentoo}} + \epsilon$$

How can we allow the slopes to vary?

$$y = \beta_0 + \beta_1 \cdot x_{\text{Bill Depth}} + \beta_2 \cdot x_{\text{Species:Chinstrap}} + \beta_3 \cdot x_{\text{Species:Gentoo}} + \beta_4 \cdot x_{\text{Bill Depth}} \cdot x_{\text{Species:Chinstrap}} + \beta_5 \cdot x_{\text{Bill Depth}} \cdot x_{\text{Species:Gentoo}} + \epsilon$$

- **Coefficient interpretation?**

Different Slopes Model in R

```r
same_slope <- lm(bill_length_mm ~ bill_depth_mm + species, penguins)

diff_slope <- lm(bill_length_mm ~ bill_depth_mm * species, penguins)
```

```r
get_regression_table(same_slope)
```

```
# A tibble: 4 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept         42.4       3.67      11.5    0.007   26.6      58.2
2 bill_depth_mm      0.411     0.196      2.10   0.171   -0.433     1.25
3 species: Adelie  -15.6       0.573    -27.2    0.001  -18.1     -13.1
4 species: Gentoo   -0.247     0.771     -0.32   0.779   -3.56      3.07
```

```r
get_regression_table(diff_slope)
```

```
# A tibble: 6 × 7
  term                    estimate std_error statistic p_value lower_ci upper_ci
  <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept                 46.6         NaN       NaN     NaN      NaN      NaN
2 bill_depth_mm              0.185       NaN       NaN     NaN      NaN      NaN
3 species: Adelie          -24.8         NaN       NaN     NaN      NaN      NaN
4 species: Gentoo          -12.1         NaN       NaN     NaN      NaN      NaN
5 bill_depth_mm:speciesA…    0.515       NaN       NaN     NaN      NaN      NaN
6 bill_depth_mm:speciesG…    0.715       NaN       NaN     NaN      NaN      NaN
```

Detour: handling curved relationships

New data context: movie ratings

```
1 library(tidyverse)
2 movies <- read_csv("https://www.lock5stat.com/datasets2e/HollywoodMovies.csv")
3
4 # Restrict our attention to dramas, horrors, and actions
5 movies2 <- movies %>%
6   filter(Genre %in% c("Drama", "Horror", "Action")) %>%
7   drop_na(Genre, AudienceScore, RottenTomatoes)
8 glimpse(movies2)
```

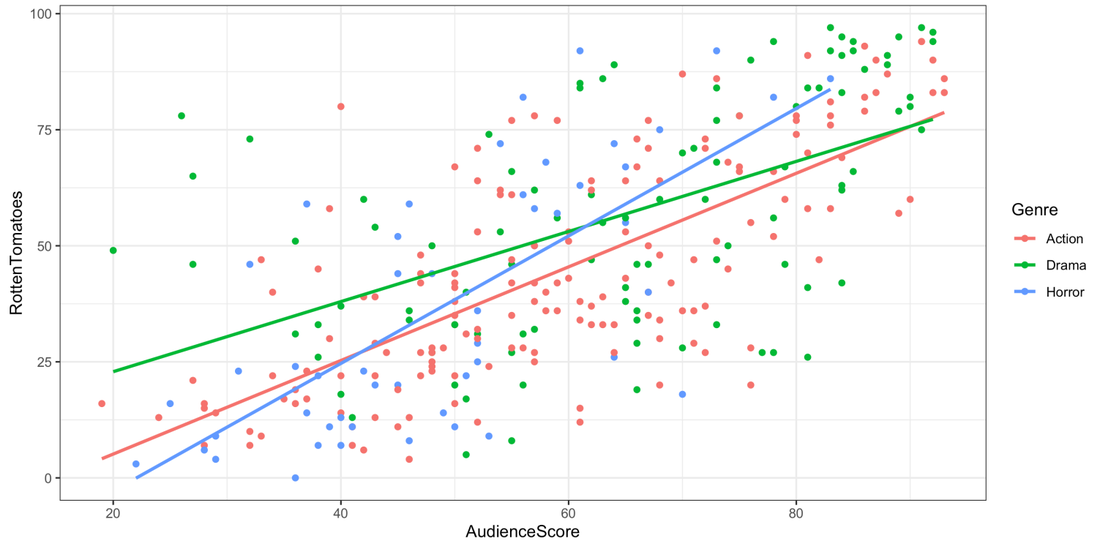
```
Rows: 313
Columns: 16
$ Movie            <chr> "Spider-Man 3", "Transformers", "Pirates of the Carib…
$ LeadStudio       <chr> "Sony", "Paramount", "Disney", "Warner Bros", "Warner…
$ RottenTomatoes   <dbl> 61, 57, 45, 60, 20, 79, 35, 28, 41, 71, 95, 42, 18, 2…
$ AudienceScore    <dbl> 54, 89, 74, 90, 68, 86, 55, 56, 81, 52, 84, 55, 70, 6…
$ Story            <chr> "Metamorphosis", "Monster Force", "Rescue", "Sacrific…
$ Genre            <chr> "Action", "Action", "Action", "Action", "Action", "Ac…
$ TheatersOpenWeek <dbl> 4252, 4011, 4362, 3103, 3778, 3408, 3959, 3619, 2911,…
$ OpeningWeekend   <dbl> 151.1, 70.5, 114.7, 70.9, 49.1, 33.4, 58.0, 45.3, 19.…
$ BOAvgOpenWeekend <dbl> 35540, 17577, 26302, 22844, 12996, 9791, 14663, 12541…
$ DomesticGross    <dbl> 336.53, 319.25, 309.42, 210.61, 140.13, 134.53, 131.9…
$ ForeignGross     <dbl> 554.34, 390.46, 654.00, 245.45, 117.90, 249.00, 157.1…
$ WorldGross       <dbl> 890.87, 709.71, 963.42, 456.07, 258.02, 383.53, 289.0…
```

Response variable: `RottenTomatoes`

Explanatory variables: `AudienceScore`

Exploring the Data

```
ggplot(data = movies2,
       mapping = aes(x = AudienceScore,
                     y = RottenTomatoes,
                     color = Genre)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

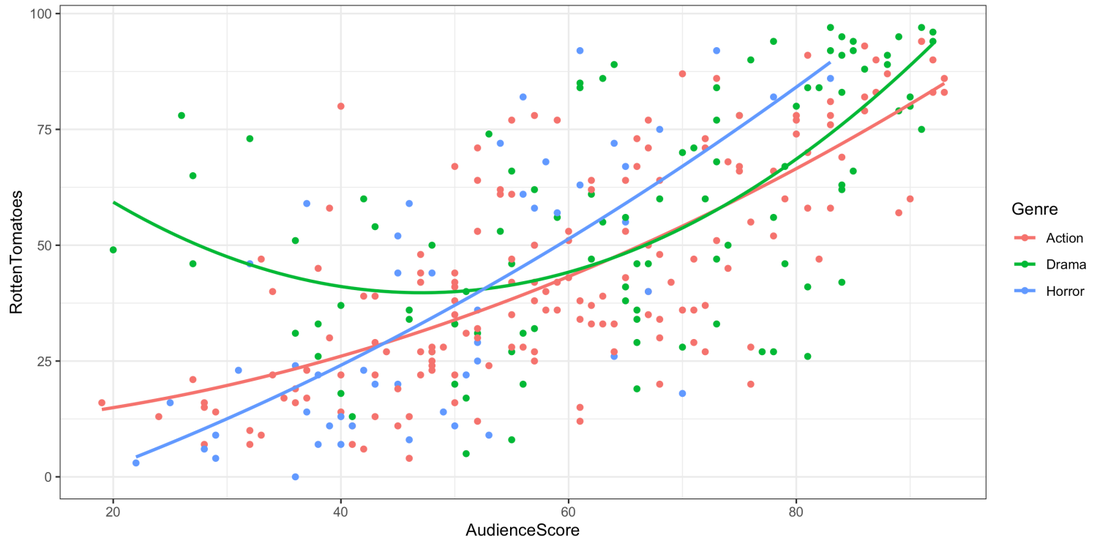


- **Q:** What do the trends suggest, would it make sense to include interaction terms in the model?

- **Q:** Does anyone spot a curved relationship for one of the genres?

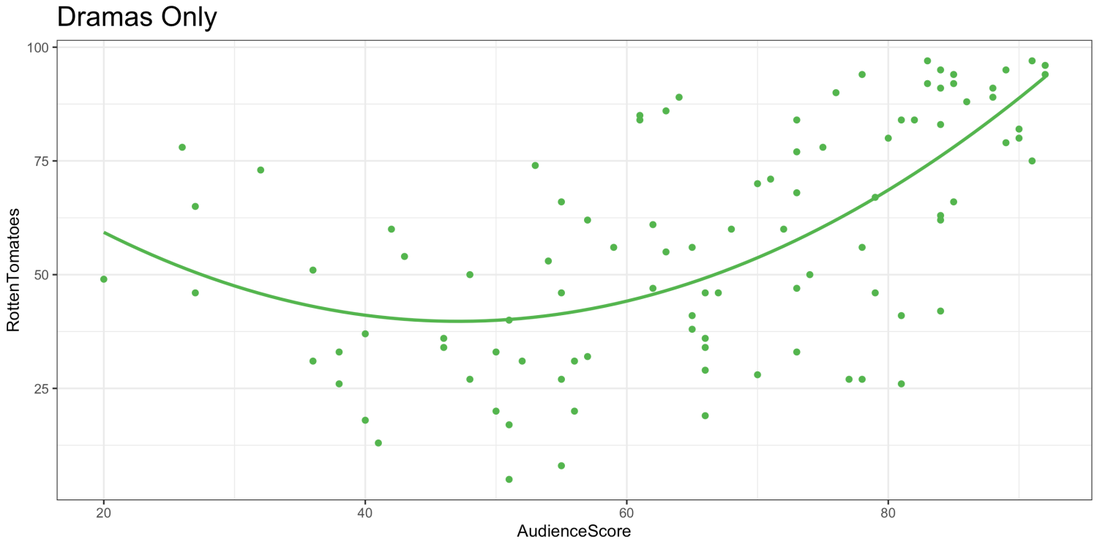
Adding a Curve to your Scatterplot

```
ggplot(data = movies2,
       mapping = aes(x = AudienceScore,
                     y = RottenTomatoes,
                     color = Genre)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE,
              formula = y ~ poly(x, degree = 2))
```



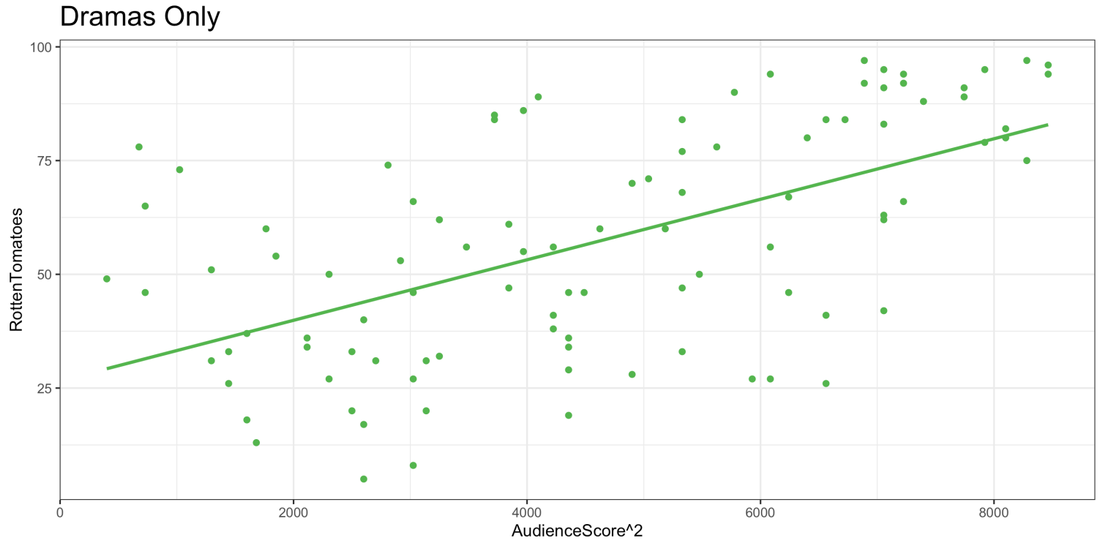
Adding a Curve to your Scatterplot

```
ggplot(data = movies2 %>% filter(Genre == "Drama"),
       mapping = aes(x = AudienceScore,
                     y = RottenTomatoes)) +
  geom_point(color = "#55B84F") +
  geom_smooth(method = "lm", se = FALSE, color = "#55B8
              formula = y ~ poly(x, degree = 2)) +
  ggtitle("Dramas Only") +
  theme(plot.title = element_text(size = 18))
```



Using Transformations to Account for Non-Linear Relationships

```
ggplot(data = movies2 %>% filter(Genre == "Drama"),
       mapping = aes(x = AudienceScore^2,
                     y = RottenTomatoes)) +
  geom_point(color = "#55B84F") +
  geom_smooth(method = "lm", se = FALSE, color = "#55B8
  ggtitle("Dramas Only") +
  theme(plot.title = element_text(size = 18))
```



- Notice that the relationship between `AudienceScore`^{2} and `RottenTomatoes` is approximately linear.
- `AudienceScore`^{2} is a **transformation** of `AudienceScore`

Fitting a Model with a Transformed Variable

```
dramas <- movies2 %>% filter(Genre == "Drama")
mod2 <- lm(RottenTomatoes ~ poly(AudienceScore, degree = 2, raw = TRUE),
           data = dramas)
get_regression_table(mod2)
```

```
# A tibble: 3 × 7
  term                    estimate std_error statistic p_value lower_ci upper_ci
  <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept                  98.8      21.3       4.64   0       56.6    141.
2 poly(AudienceScore, de…   -2.51      0.722     -3.48   0.001   -3.94    -1.08
3 poly(AudienceScore, de…    0.027     0.006      4.58   0        0.015    0.038
```

- We'll practice a very common transformation in lab this week - the **log transformation**
- **Q**: Why is it still called a linear regression if the model also handles curved relationship?
- Transformations are super powerful: we can now use linear models even if the *untransformed* variables have non-linear relationships.

Model Building Guidance

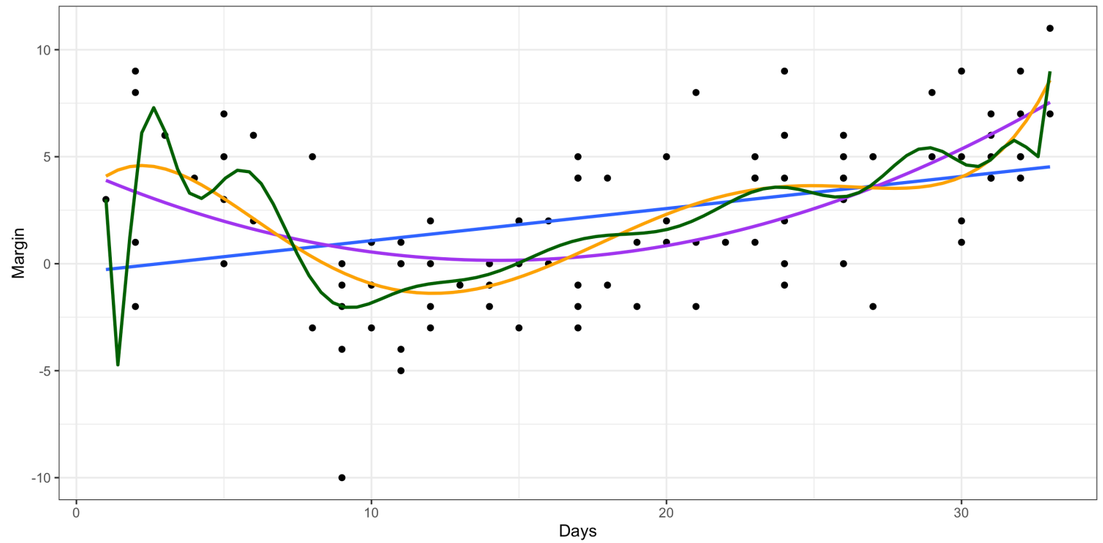
What degree of polynomial should I include in my model?

Guiding Principle: Capture the general trend, not the noise.

$$y = f(x) + \epsilon$$

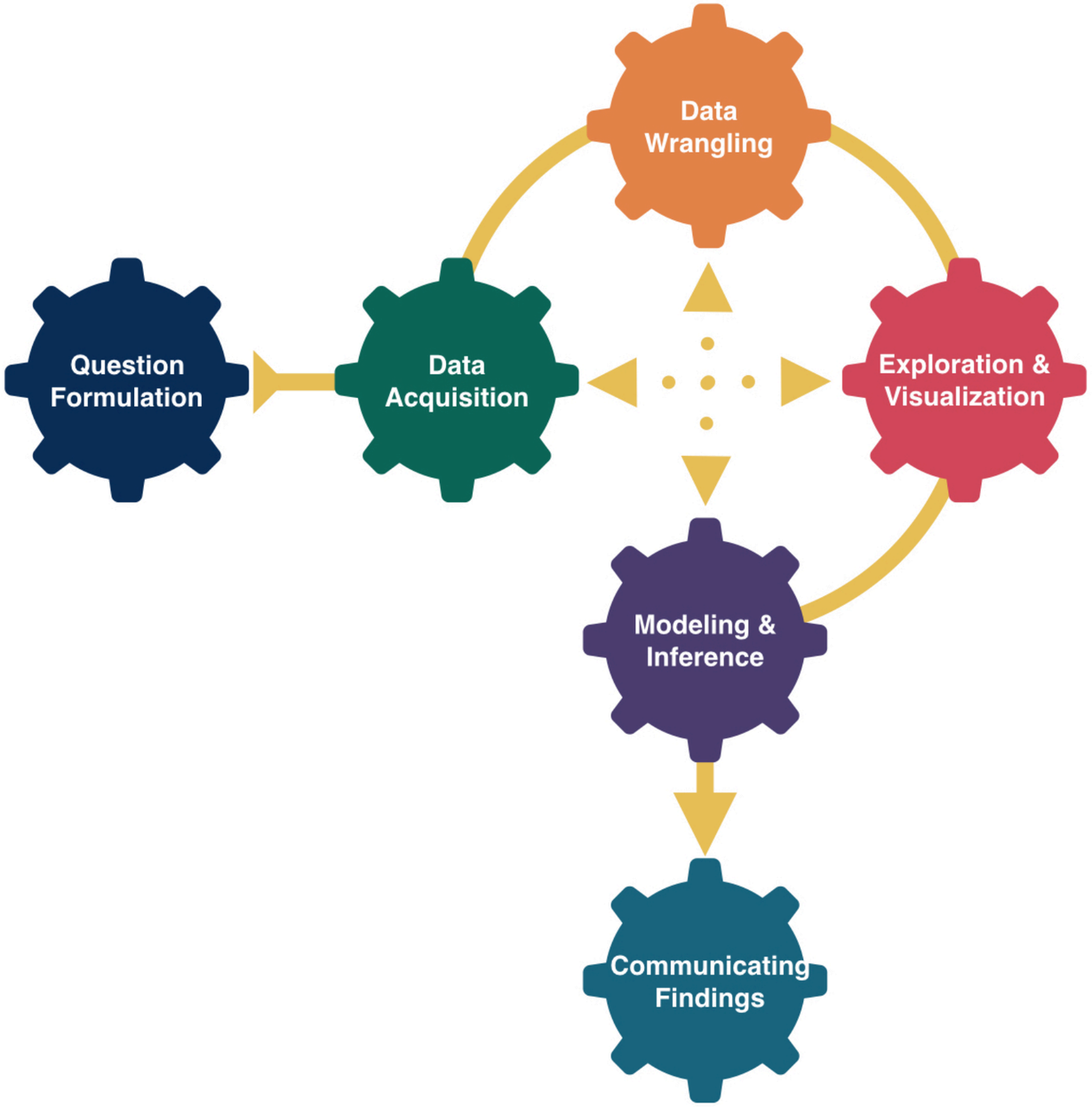
$$y = \text{TREND} + \text{NOISE}$$

2008 Election Polls Example:



Next time

- Linear regression with many variables
- Multicollinearity
- Model selection



Linear Models V

Megan Ayers

Math 141 | Spring 2026

Wednesday, Week 5

Goals for Today

- Learning Check
- Regression with many variables
- Adjusted R^2 and model selection
- Multicollinearity
- Application of Multiple Linear Regression

Learning check handout

Linear Regression

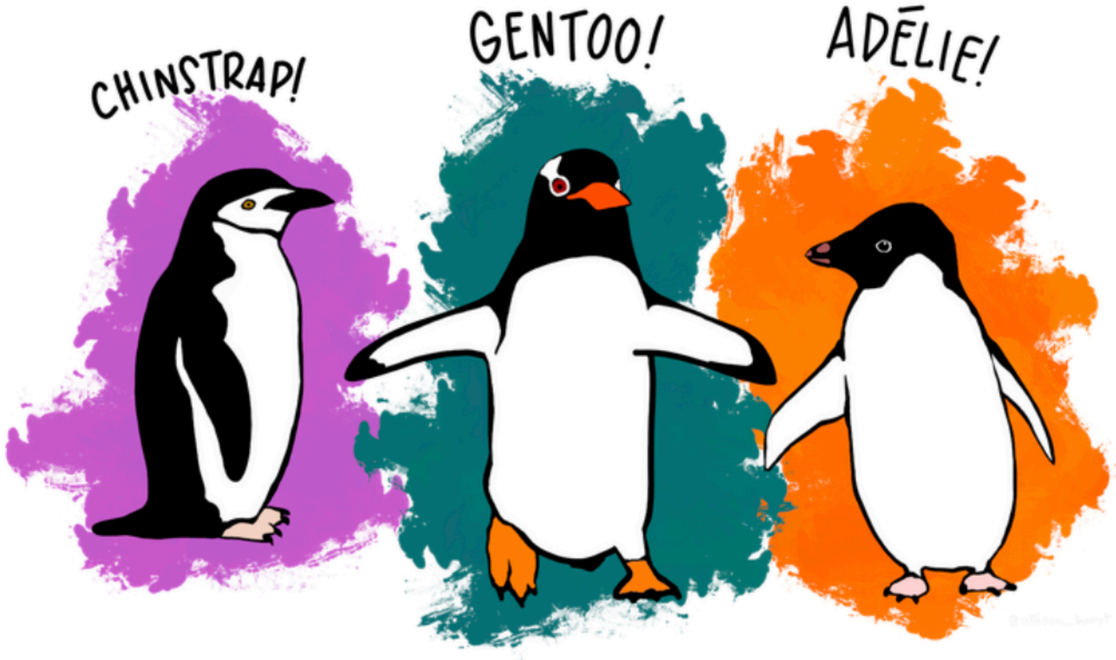
Model Form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Linear regression is a flexible class of models that allow for:

- Both quantitative and categorical **explanatory** variables.
- **Multiple** explanatory variables.
- Transformed explanatory variables.
- BUT the **response variable is quantitative**.

More time with the `palmerpenguins`!



The Palmer Archipelago penguins. Artwork by @allison_horst.

Recap: Regression with the penguins so far

One Quantitative Variable

```
library(palmerpenguins)
library(moderndive)
mod <- lm(bill_length_mm ~ bill_depth_mm,
          data = penguins)
get_regression_table(mod) %>% select(term, estimate)
```

```
# A tibble: 2 × 2
  term          estimate
  <chr>            <dbl>
1 intercept       35.7
2 bill_depth_mm    0.493
```

One Categorical Variable

```
mod <- lm(bill_length_mm ~ species,
          data = penguins)
get_regression_table(mod) %>% select(term, estimate)
```

```
# A tibble: 3 × 2
  term             estimate
  <chr>               <dbl>
1 intercept           50.0
2 species: Adelie    -16.2
3 species: Gentoo     -1.5
```

Equal Slopes Model

```
mod <- lm(bill_length_mm ~ species + bill_depth_mm,
          data = penguins)
get_regression_table(mod) %>% select(term, estimate)
```

```
# A tibble: 4 × 2
  term             estimate
  <chr>               <dbl>
1 intercept          42.4
2 species: Adelie   -15.6
3 species: Gentoo    -0.247
4 bill_depth_mm       0.411
```

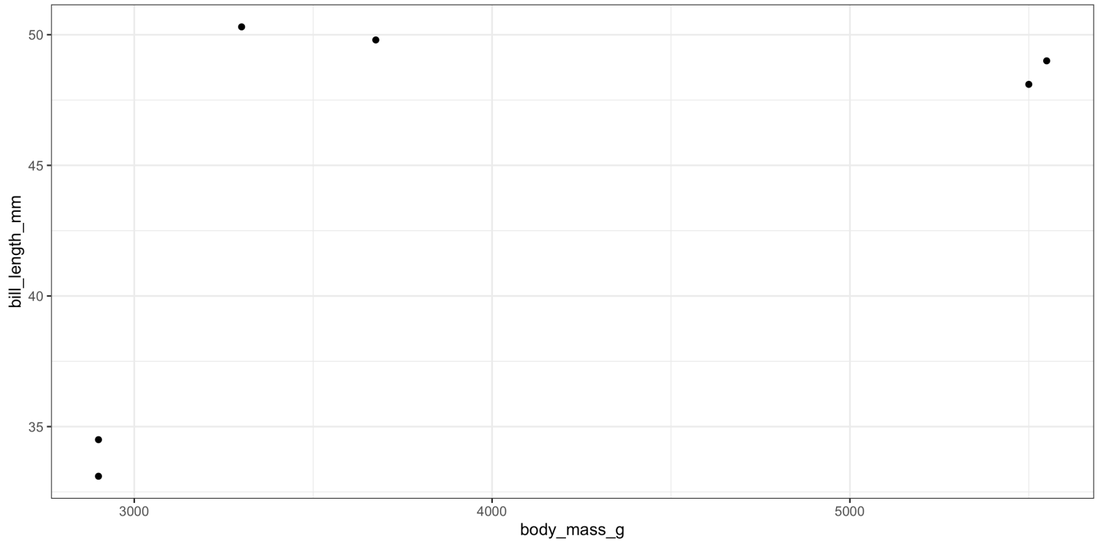
Different Slopes Model

```
mod <- lm(bill_length_mm ~ species*bill_depth_mm,
          data = penguins)
get_regression_table(mod) %>% select(term, estimate)
```

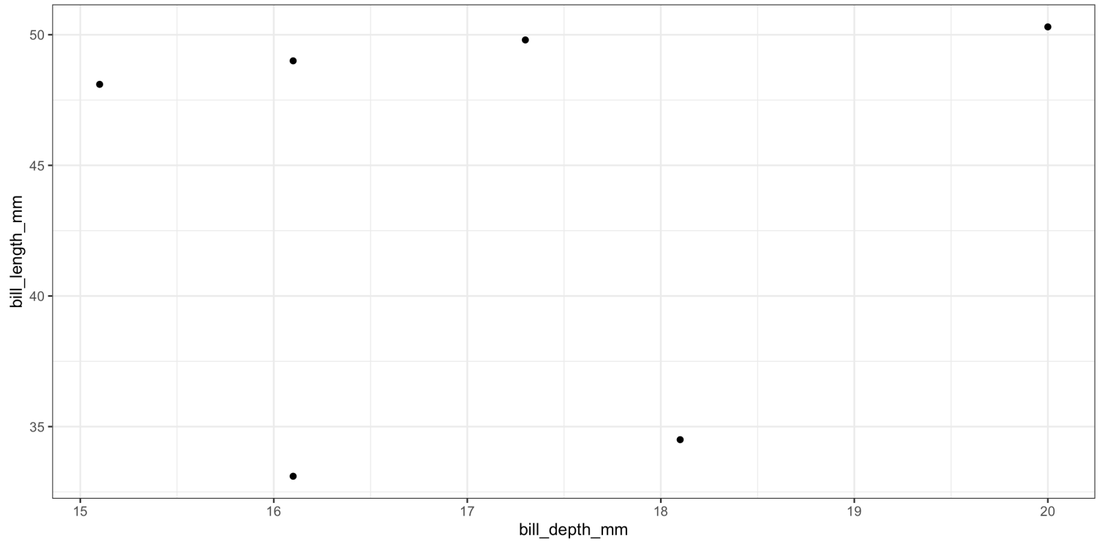
```
# A tibble: 6 × 2
  term                           estimate
  <chr>                             <dbl>
1 intercept                          46.6
2 species: Adelie                   -24.8
3 species: Gentoo                   -12.1
4 bill_depth_mm                       0.185
5 species: Adelie:bill_depth_mm       0.515
6 species: Gentoo:bill_depth_mm       0.715
```

This time: Regression with multiple quantitative predictors

```r
ggplot(penguins, aes(x = body_mass_g, y = bill_length_mm
  geom_point()
```




```r
ggplot(penguins, aes(x = bill_depth_mm, y = bill_length_
  geom_point()
```



$$y = \beta_0 + \beta_1 x_{\text{Bill Depth}} + \beta_2 x_{\text{Body Mass}} + \epsilon$$

This time: Regression with multiple quantitative predictors

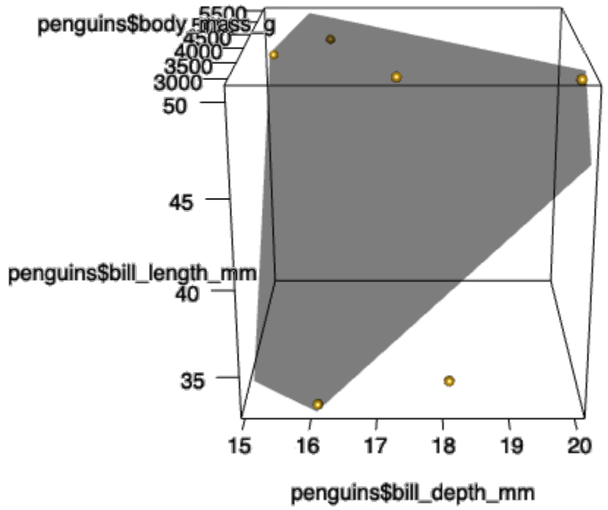
In R:

```r
mod <- lm(bill_length_mm ~ bill_depth_mm + body_mass_g, penguins)
get_regression_table(mod)
```

```
# A tibble: 3 × 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept      -42.9      36.9       -1.16   0.329 -160.      74.5
2 bill_depth_mm    3.48      1.75       1.99   0.141   -2.10     9.04
3 body_mass_g      0.007     0.002      2.79   0.068   -0.001    0.015
```

Visualizing the best fit plane

```
options(rgl.useNULL = TRUE)
options(rgl.printRglwidget = TRUE)
library(rgl)
plot3d(x = penguins$bill_depth_mm, y = penguins$body_mass_g,
       z = penguins$bill_length_mm, type = "s", col = "goldenrod", size = 1)
planes3d(a = coef(mod)[2], b = coef(mod)[3], c=-1, d = coef(mod)[1], alpha = .5)
```



Least Squares and Prediction

- Still minimizing the sum of squared residuals (from the plane)
- Still predict like usual

```
pred_dat <- data.frame(bill_depth_mm = c(16, 18),
                       body_mass_g = c(4000, 5000))
predict(mod, pred_dat)
```

```
       1        2 
40.45496 54.34147 
```

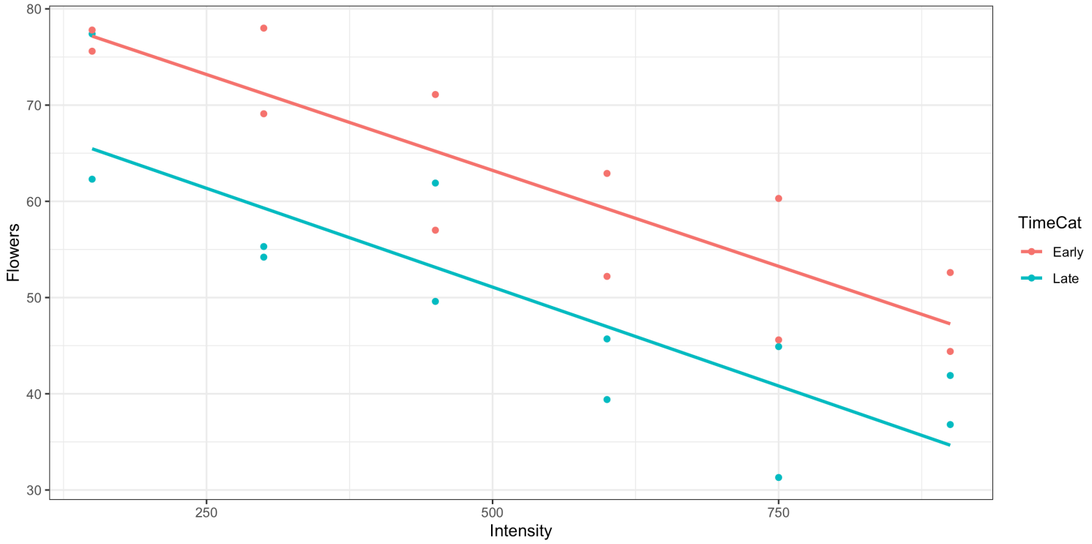
Model Building Guidance

Should we always include an interaction term? How many variables should I include?

Guiding Principle: Occam's Razor for Modeling

> "All other things being equal, simpler models are to be preferred over complex ones." – ModernDive

Guiding Principle: Consider your modeling goals.



- The equal slopes model allows us to control for the intensity of the light and then see the impact of being in the early or late timing groups on the number of flowers.
- Later in the course will learn statistical procedures for determining whether or not a particular term should be included in the model.

What if I want to include many explanatory variables??

Model Building Guidance

We often have several potential explanatory variables. How do we determine which to include in the model and in what form?

Guiding Principle: Include explanatory variables that attempt to explain **different** aspects of the variation in the response variable.

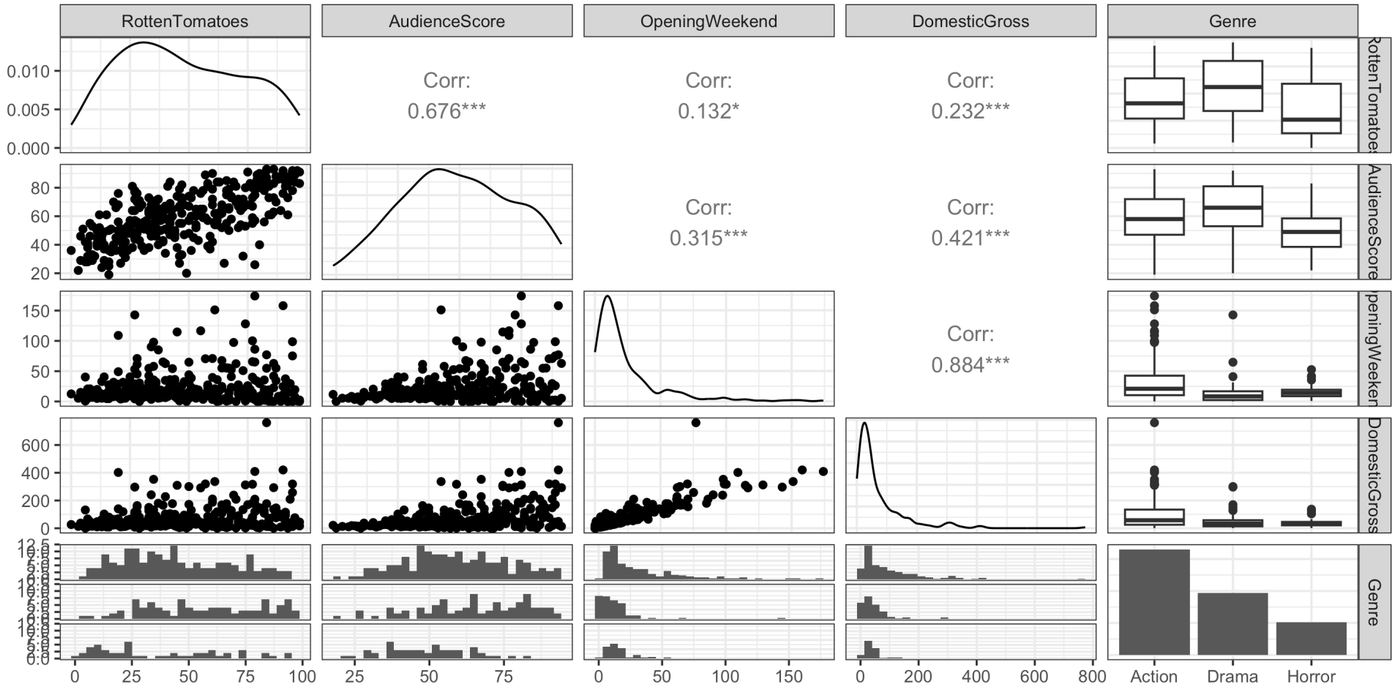
Example: Movie Ratings

```
1 library(tidyverse)
2 library(moderndive)
3 movies <- read_csv("https://www.lock5stat.com/datasets2e/HollywoodMovies.csv")
4
5 # Restrict our attention to dramas, horrors, and actions
6 movies2 <- movies %>%
7   filter(Genre %in% c("Drama", "Horror", "Action")) %>%
8   drop_na(Genre, AudienceScore, RottenTomatoes)
9 glimpse(movies2)
```

```
Rows: 313
Columns: 16
$ Movie            <chr> "Spider-Man 3", "Transformers", "Pirates of the Carib…
$ LeadStudio       <chr> "Sony", "Paramount", "Disney", "Warner Bros", "Warner…
$ RottenTomatoes   <dbl> 61, 57, 45, 60, 20, 79, 35, 28, 41, 71, 95, 42, 18, 2…
$ AudienceScore    <dbl> 54, 89, 74, 90, 68, 86, 55, 56, 81, 52, 84, 55, 70, 6…
$ Story            <chr> "Metamorphosis", "Monster Force", "Rescue", "Sacrific…
$ Genre            <chr> "Action", "Action", "Action", "Action", "Action", "Ac…
$ TheatersOpenWeek <dbl> 4252, 4011, 4362, 3103, 3778, 3408, 3959, 3619, 2911,…
$ OpeningWeekend   <dbl> 151.1, 70.5, 114.7, 70.9, 49.1, 33.4, 58.0, 45.3, 19.…
$ BOAvgOpenWeekend <dbl> 35540, 17577, 26302, 22844, 12996, 9791, 14663, 12541…
$ DomesticGross    <dbl> 336.53, 319.25, 309.42, 210.61, 140.13, 134.53, 131.9…
$ ForeignGross     <dbl> 554.34, 390.46, 654.00, 245.45, 117.90, 249.00, 157.1…
$ WorldGross       <dbl> 890.87, 709.71, 963.42, 456.07, 258.02, 383.53, 289.0…
```

Example: Movie Ratings

```
1 library(GGally)
2 movies2 %>%
3   select(RottenTomatoes, AudienceScore, OpeningWeekend, DomesticGross, Genre) %>%
4   ggpairs()
```



Model Building Guidance

We often have several potential explanatory variables. How do we determine which to include in the model and in what form?

Guiding Principle: Include explanatory variables that attempt to explain **different** aspects of the variation in the response variable.

```
1 mod_movies <- lm(RottenTomatoes ~ AudienceScore + Genre + DomesticGross, data = movies2)
2 get_regression_table(mod_movies)
```

```
# A tibble: 5 × 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept      -12.5       4.08     -3.06    0.002  -20.5     -4.44
2 AudienceScore    0.975     0.072    13.6     0        0.834    1.12
3 Genre: Drama     6.12      2.64      2.31    0.021    0.916   11.3
4 Genre: Horror    2.06      3.14      0.655   0.513   -4.12     8.24
5 DomesticGross   -0.006     0.015    -0.431   0.667   -0.035    0.023
```

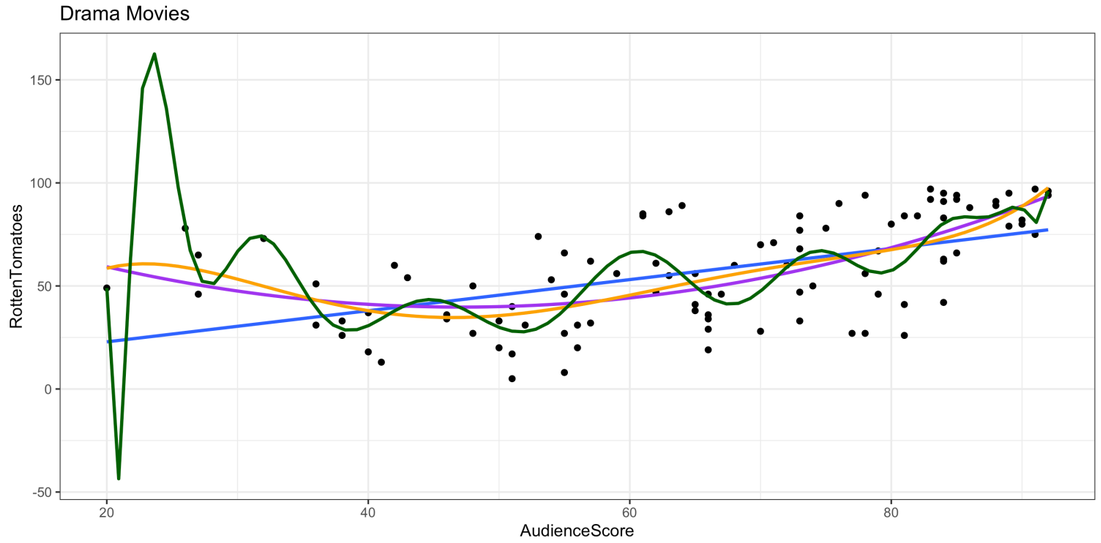
Transformations: Model Building Guidance

What degree of polynomial should I include in my model?

Guiding Principle: Capture the general trend, not the noise.

$$y = f(x) + \epsilon$$

$$y = \text{TREND} + \text{NOISE}$$



Model Building Guidance

Suppose I built 3 different models. **Which is best?**

```
mod1 <- lm(RottenTomatoes ~ AudienceScore, data = movies2)
mod2 <- lm(RottenTomatoes ~ AudienceScore + Genre, data = movies2)
mod3 <- lm(RottenTomatoes ~ AudienceScore + Genre + DomesticGross, data = movies2)
```

- Big question! Take **Math 243: Statistical Learning** to learn systematic model selection techniques.
- We will explore one approach. (But there are many possible approaches!)

Comparing Models with R^2

Strategy: Compute the R^2 value for each model and pick the one with the highest R^2.

```
mod1 <- lm(RottenTomatoes ~ AudienceScore, data = movies2)
mod2 <- lm(RottenTomatoes ~ AudienceScore + Genre, data = movies2)
mod3 <- lm(RottenTomatoes ~ AudienceScore + Genre + DomesticGross, data = movies2)

get_regression_summaries(mod1) %>% select(r_squared)
```

```
# A tibble: 1 × 1
  r_squared
      <dbl>
1     0.457
```

```
get_regression_summaries(mod2) %>% select(r_squared)
```

```
# A tibble: 1 × 1
  r_squared
      <dbl>
1     0.469
```

```
get_regression_summaries(mod3) %>% select(r_squared)
```

```
# A tibble: 1 × 1
  r_squared
      <dbl>
1     0.469
```

Strategy: Compute the R^2 value for each model and pick the one with the highest R^2.

```
  1 get_regression_summaries(mod1) %>% select(r_squared)
```

```
# A tibble: 1 × 1
  r_squared
      <dbl>
1     0.457
```

```
  1 get_regression_summaries(mod2) %>% select(r_squared)
```

```
# A tibble: 1 × 1
  r_squared
      <dbl>
1     0.469
```

```
  1 get_regression_summaries(mod3) %>% select(r_squared)
```

```
# A tibble: 1 × 1
  r_squared
      <dbl>
1     0.469
```

Problem: As we add predictors, the R^2 value will only increase.

Guiding Principle: Occam's Razor for Modeling

> "All other things being equal, simpler models are to be preferred over complex ones." – ModernDive

Comparing Models with the Adjusted R^2

New Measure of quality: Adjusted R^2 (Coefficient of Determination)

$$\text{adj}R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}\left(\frac{n-1}{n-p-1}\right)$$

where p is the number of explanatory variables in the model.

- Now we will penalize larger models.
- **Strategy**: Compute the adjusted R^2 value for each model and pick the one with the highest adjusted R^2.

Strategy: Compute the **adjusted** R^2 value for each model and pick the one with the highest adjusted R^2.

```
mod1 <- lm(RottenTomatoes ~ AudienceScore, data = movies2)
mod2 <- lm(RottenTomatoes ~ AudienceScore + Genre, data = movies2)
mod3 <- lm(RottenTomatoes ~ AudienceScore + Genre + DomesticGross, data = movies2)

get_regression_summaries(mod1) %>% select(r_squared, adj_r_squared)
```

```
# A tibble: 1 × 2
  r_squared adj_r_squared
      <dbl>         <dbl>
1     0.457         0.455
```

```
get_regression_summaries(mod2) %>% select(r_squared, adj_r_squared)
```

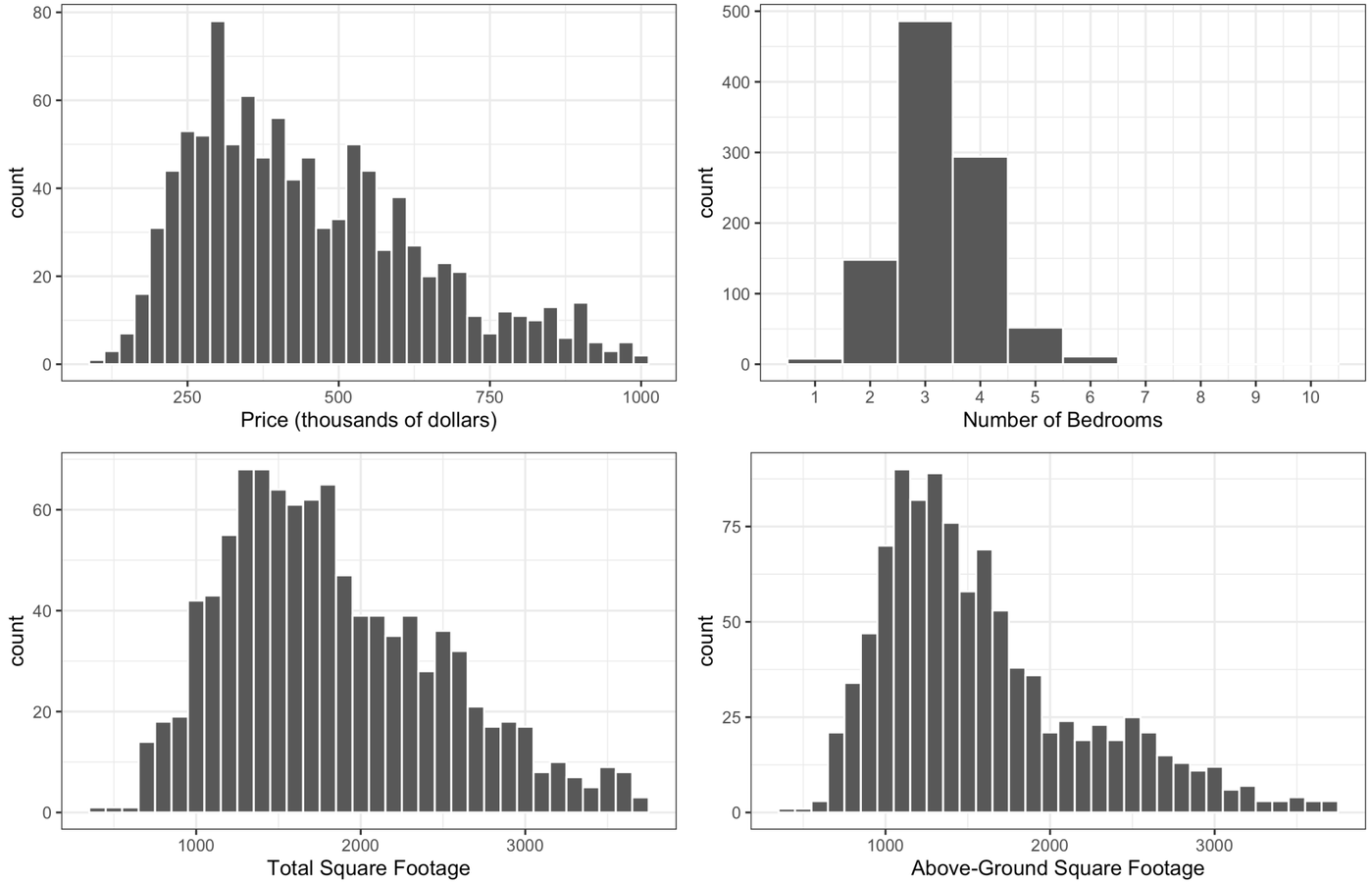
```
# A tibble: 1 × 2
  r_squared adj_r_squared
      <dbl>         <dbl>
1     0.469         0.464
```

```
get_regression_summaries(mod3) %>% select(r_squared, adj_r_squared)
```

```
# A tibble: 1 × 2
  r_squared adj_r_squared
      <dbl>         <dbl>
1     0.469         0.462
```

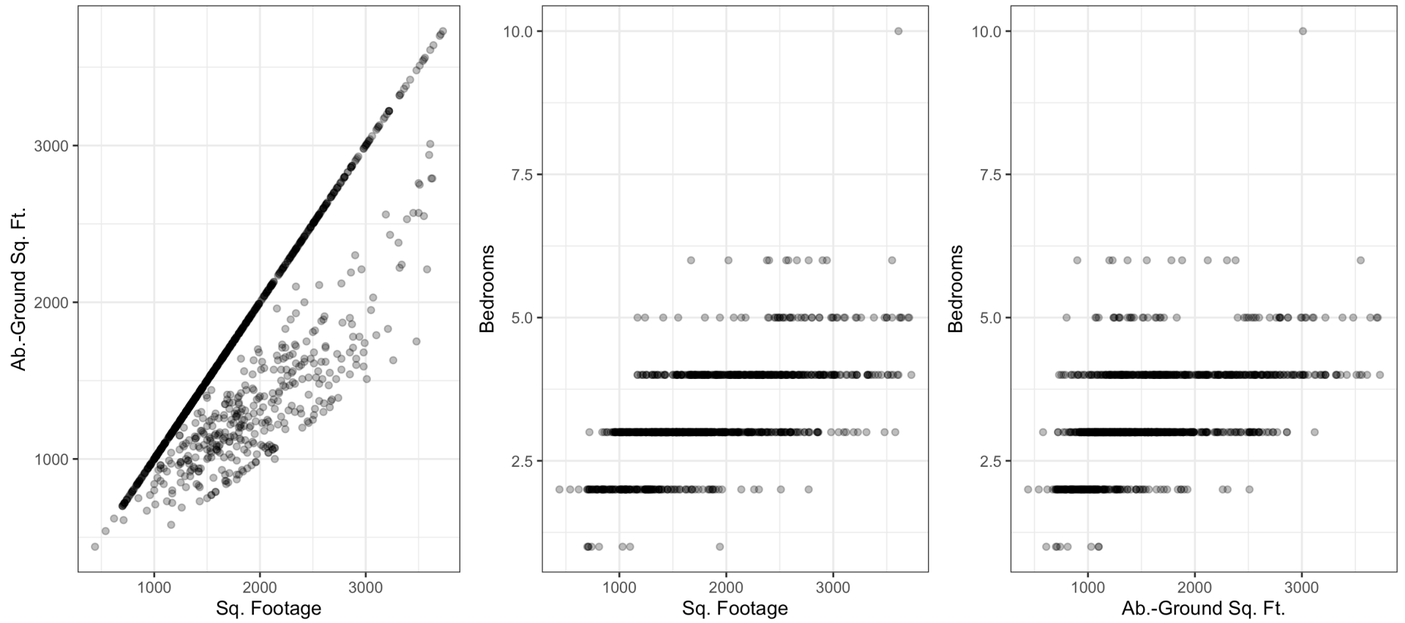
Case Study: Home Prices in King County, WA

We'll consider : Sale price (**response**); and square footage, above-ground square footage, and number of bedrooms (potential explanatory variables). Check out the exploratory plots below:



- **Q**: How would you describe the distribution of `price`?
- **Q**: What do you expect the relationship between `price` and `bedrooms` to be like?

Exploration, Multicollinearity



- **Multicollinearity**: when explanatory variables are **highly** correlated with one another. Multicollinearity often results in coefficients that are distorted in erroneous ways! We want to avoid this (*some* correlation is okay!)
- **Q**: Which pair of explanatory variables suffers *most* from multicollinearity?

Model Estimation

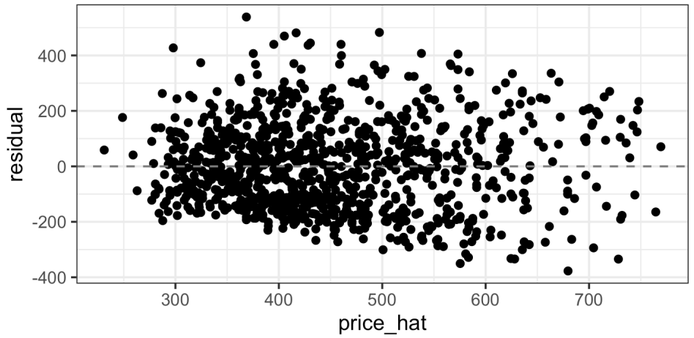
```
# A tibble: 3 × 7
  term        estimate std_error statistic p_value lower_ci upper_ci
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    197.        19.4      10.1    0      159.     235.
2 sqft_living    0.177      0.01     18.4    0        0.158    0.196
3 bedrooms     -21.6        7.27     -2.97   0.003  -35.9     -7.34
```

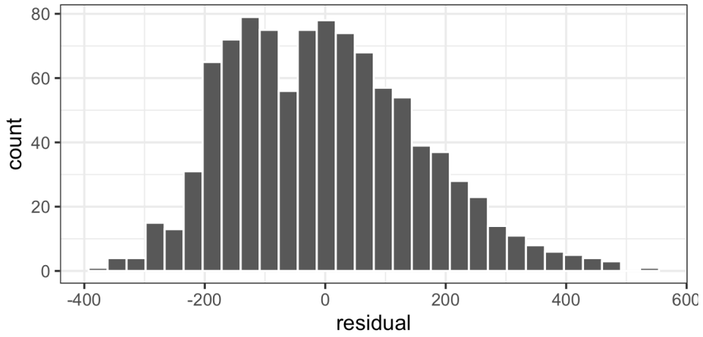
- **Q**: What is the equation, including coefficient values, for the estimated model?

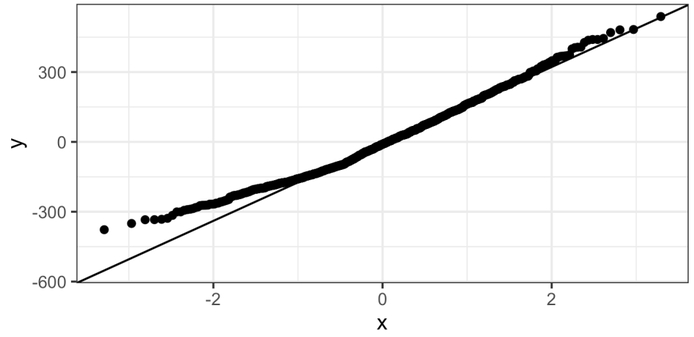
$$\widehat{\text{price}} = 197 + 0.177 * \text{sqrt\_living} - 21.6 * \text{bedrooms}$$

- **Q**: How should interpret the **intercept** coefficient in our model?
  - When `sqrt_living` and `bedrooms` are both 0, we expect/predict price to be $197,000 on average.
- **Q**: The coefficient for **Bedrooms**? Does this match your expectation?
  - When the number of `bedrooms` increases by 1, we expect/predict price to **decrease** by $21,600 on average, **holding `sqft_living` constant**.

LINE Assumptions







Adjusted R^2

What's the adjusted R^2 of our model?

```
get_regression_summaries(mod)[, c("r_squared", "adj_r_squared")]
```

```
# A tibble: 1 × 2
  r_squared adj_r_squared
      <dbl>         <dbl>
1     0.314         0.313
```

Not super close to 1. Compare to a simpler model:

```
small_mod <- lm(price ~ bedrooms, data = house)
get_regression_summaries(small_mod)[, c("r_squared", "adj_r_squared")]
```

```
# A tibble: 1 × 2
  r_squared adj_r_squared
      <dbl>         <dbl>
1     0.082         0.081
```

Compare to a much bigger model with additional variables:

```
big_mod <- lm(price ~ bedrooms + bathrooms + sqft_living + sqft_above + sqft_lot +
                  view + condition + yr_built, data = house)
get_regression_summaries(big_mod)[, c("r_squared", "adj_r_squared")]
```

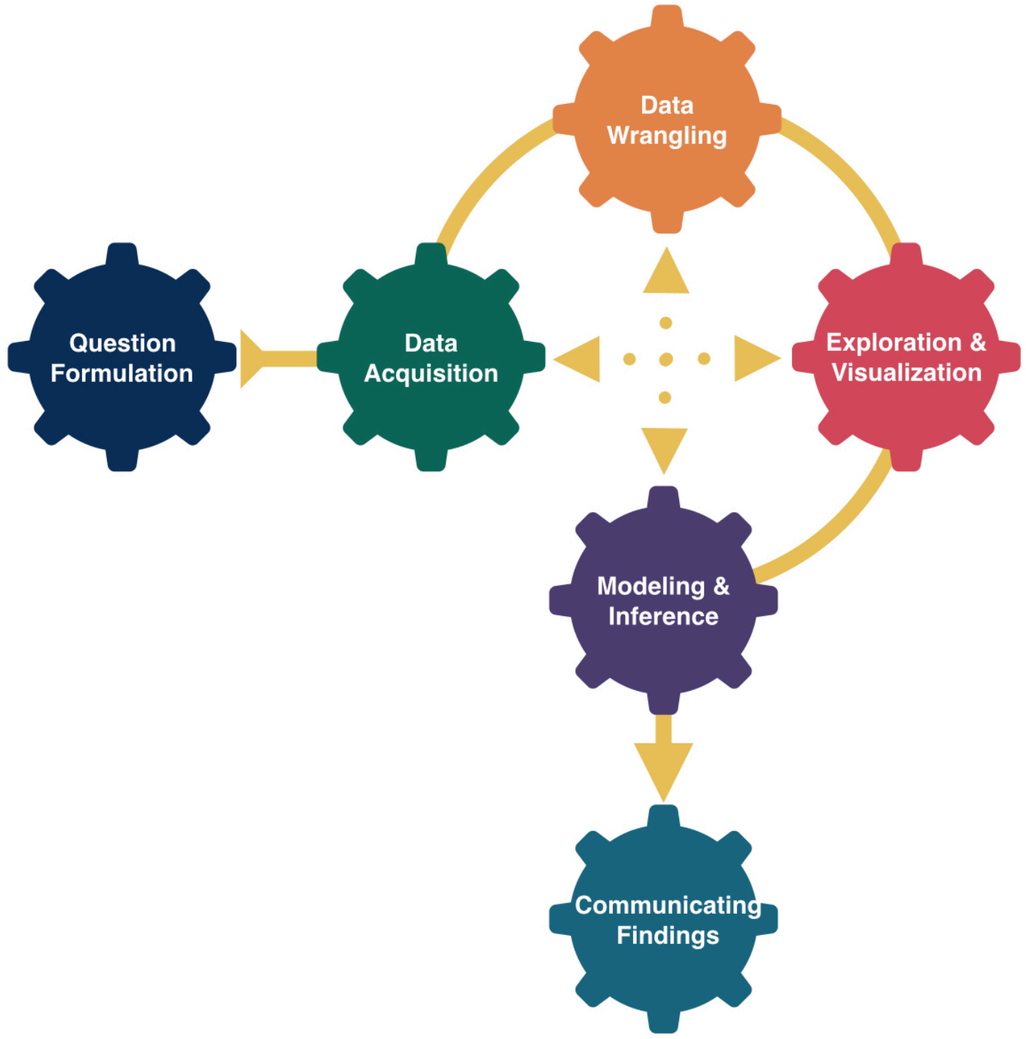
```
# A tibble: 1 × 2
  r_squared adj_r_squared
      <dbl>         <dbl>
1     0.434         0.429
```

Developing Linear Models

- Key components that we've learned about over the last 5 lectures:
  - Determining the **response** variable and the potential **explanatory** variable(s)
  - Writing out the **model form** and understanding what the terms represent **in context**
  - **Building** and **visualizing** linear regression models in R
  - **Validating** model assumptions with diagnostic plots
  - **Comparing** different potential models

Next time

- Sampling distributions!



Sampling Distributions I

Megan Ayers

Math 141 | Spring 2026

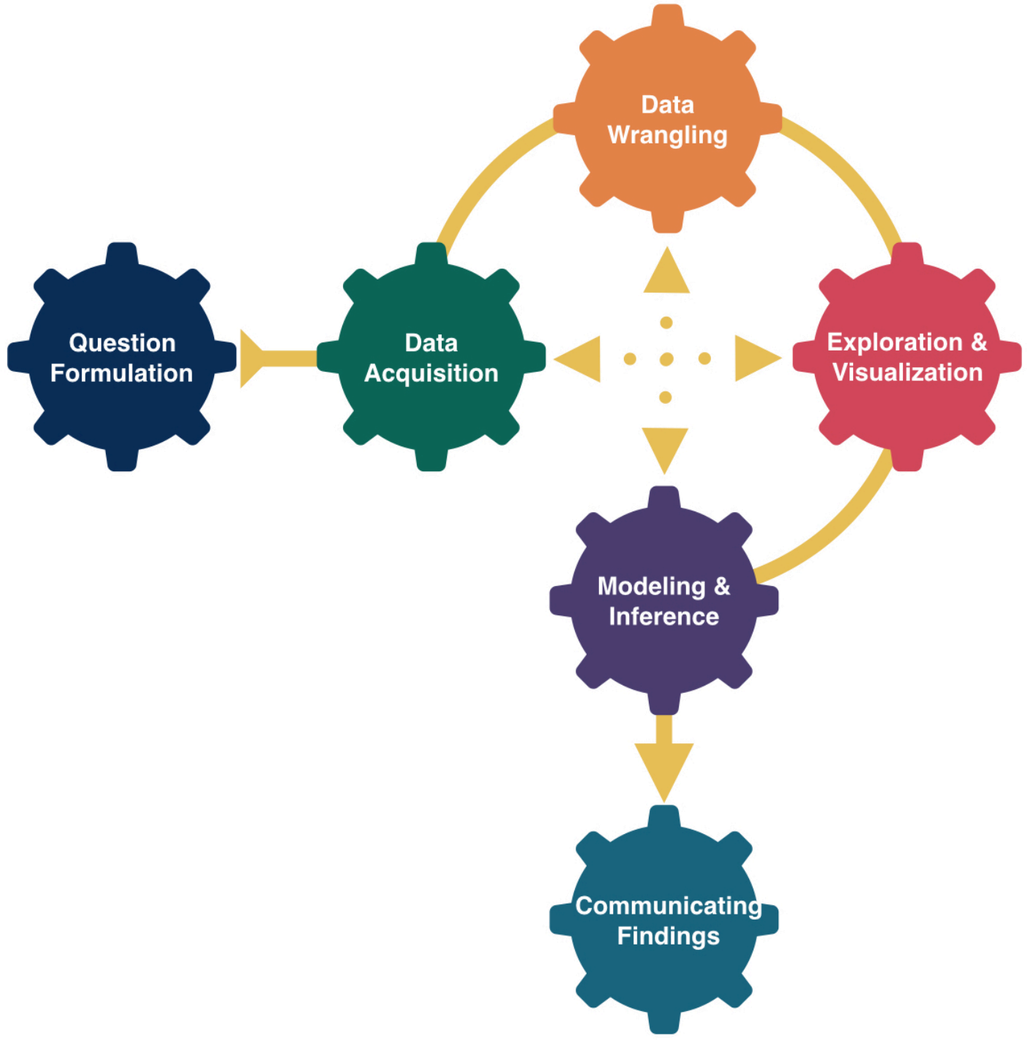
Friday, Week 5

Announcements/Reminders

- New member of the course assistant team
- Updated office hours
- HW 4 due today
- Extra time for HW 5: due **Monday** March 9
- Final exam slots finalized
- Reminder to pick up your learning assessment for feedback

Goals for Today

- Start learning the foundations of inference
- Perform a group sampling activity
- Discuss random sampling: the heart of statistics!



Distinguishing between the population and the sample

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

- Parameters:
  - Based on the **population**
  - Unknown then if don't have data on the whole population
  - EX: $\beta_0, \beta_1, \ldots, \beta_p$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

- Statistics:
  - Based on the **sample** data
  - Known
  - Usually estimate a population parameter
  - EX: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$

General Definitions: Parameters and Statistics

- **Parameter:** Numerical characteristic of a population (e.g., average of a variable in a population)
- **Statistic:** Estimate of the population parameter using the sample (e.g., average of the same variable *in the sample*)
- Researchers often wish to investigate the value of a **parameter** in a population.
  - The proportion of U.S. voters who plan to vote for a particular presidential candidate.
  - The mean life-time earnings for Reed college graduates
- But it is often not feasible to collect complete information on the population.
- Instead, researchers collect a sample and measure a **statistic**, which *estimates* the population parameter
  - The proportion of voters in a sample of size 500 who plan to vote for the candidate.
  - The mean life-time earnings for 100 randomly chosen Reed graduates.

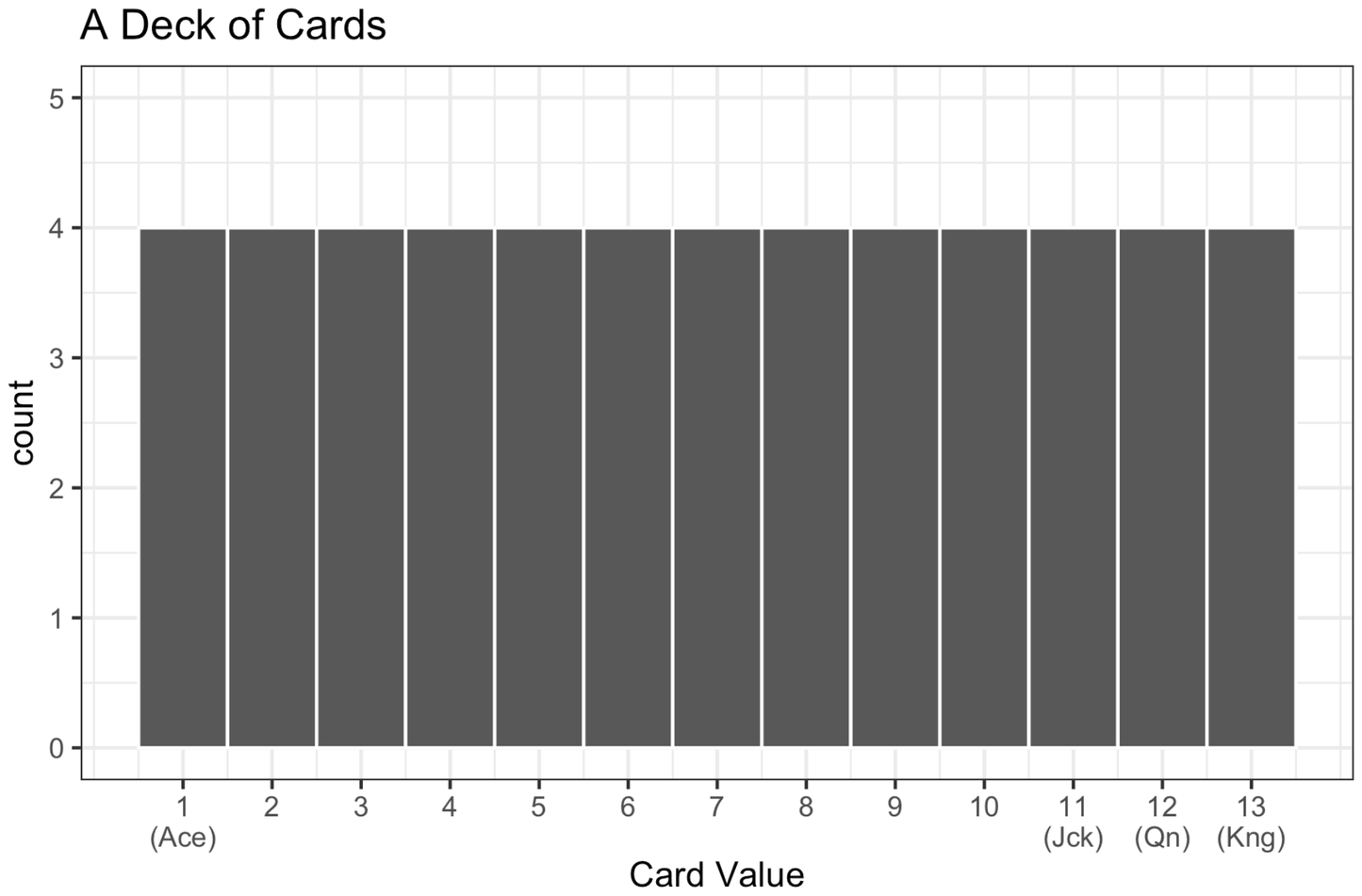
Sampling Activity

Decks of cards

If we count:

- **Aces** as 1
- **Jacks** as 11
- **Queens** as 12
- **Kings** as 13

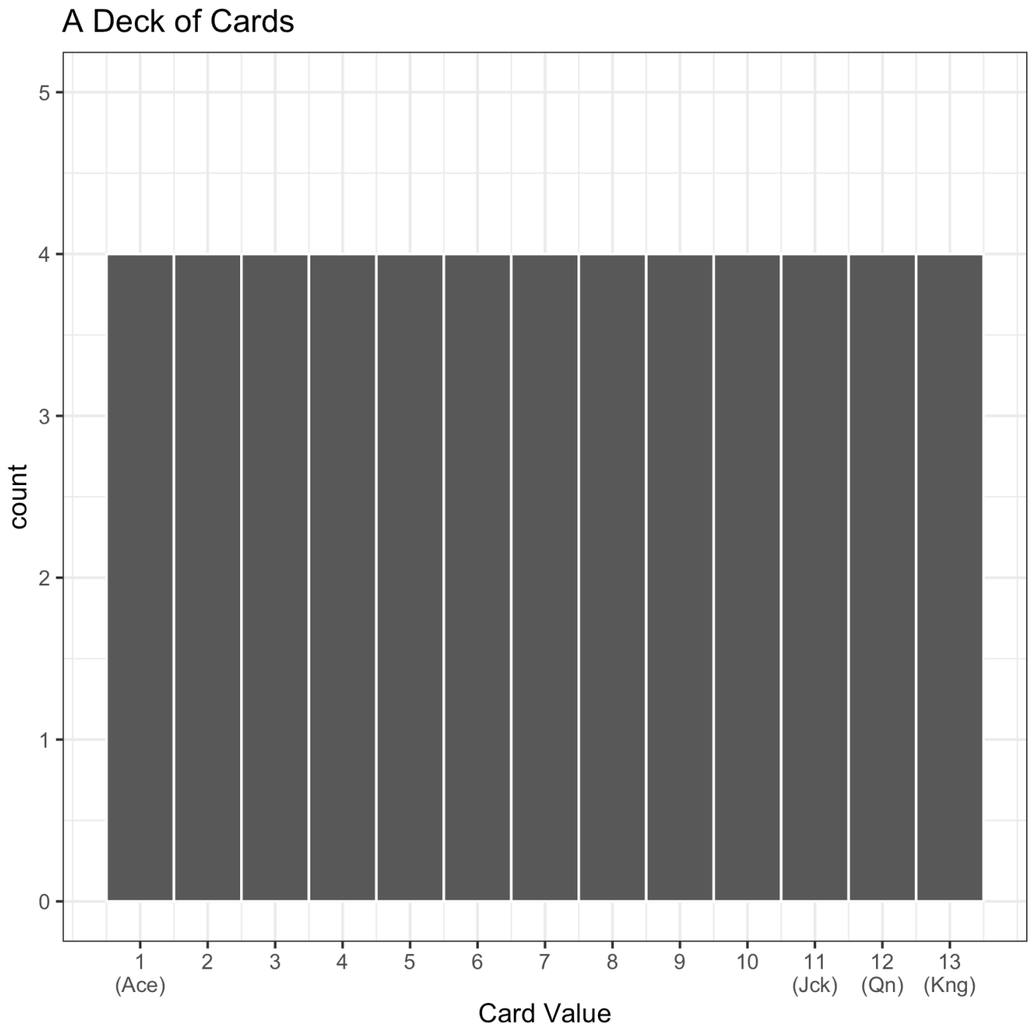
The distribution of numbers in a deck of cards looks like:



Drawing cards: Population, Sample, Parameter, and Statistic

- Today, we're going to be:
  - Randomly drawing 10 cards from a deck of cards
  - Calculating the average value of our 10 cards

- **Q:** What is the **sample**? What is the **population**?
- **Q:** What is the **statistic**? What is the **parameter**?



Activity Instructions

1. Thoroughly shuffle your group's deck of cards.
2. Draw 10 cards from the deck to form a sample.
3. Compute the average/mean value of your 10 cards
4. Write the value of the average/mean on a sticky note and add to chalkboard.
5. Repeat steps 1 - 4 an additional four times.

Each group should have calculated 5 averages from 5 different samples of 10 cards!

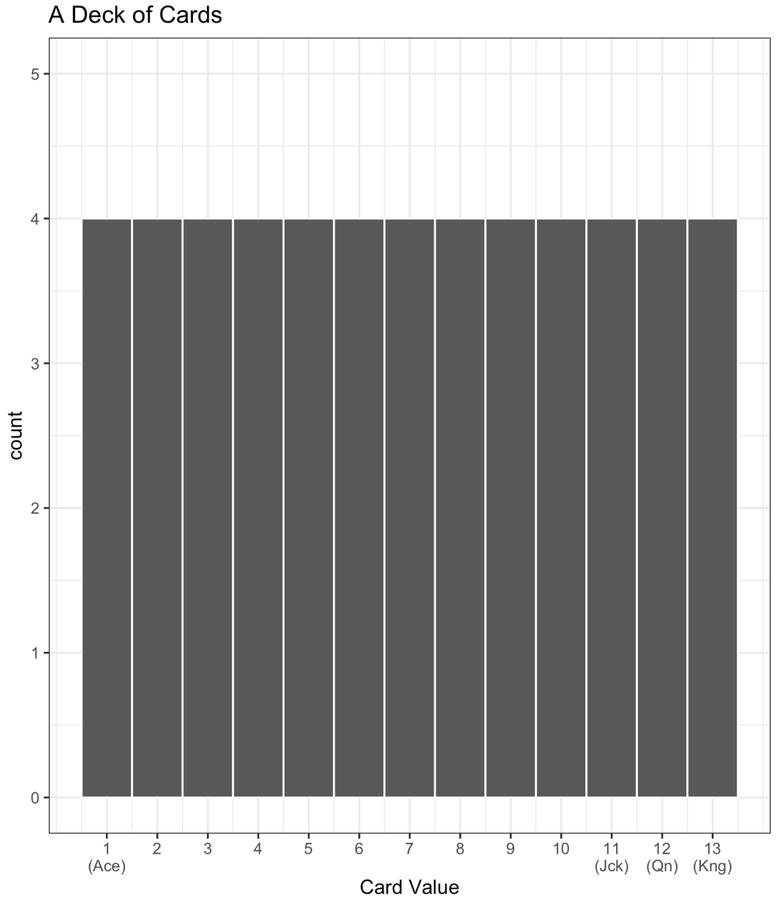
Small Group Discussion

Once you're done, discuss the following questions with your group:

1. What is the true average card value in a deck of cards?
2. How does the distribution of sample means compare to the distribution of card values in a deck of cards?
3. What is the relationship between the centers of the two distributions?
4. Which distribution appears to have more variability?
5. How do the shapes of the two distributions compare? Why do they differ?

Discussion

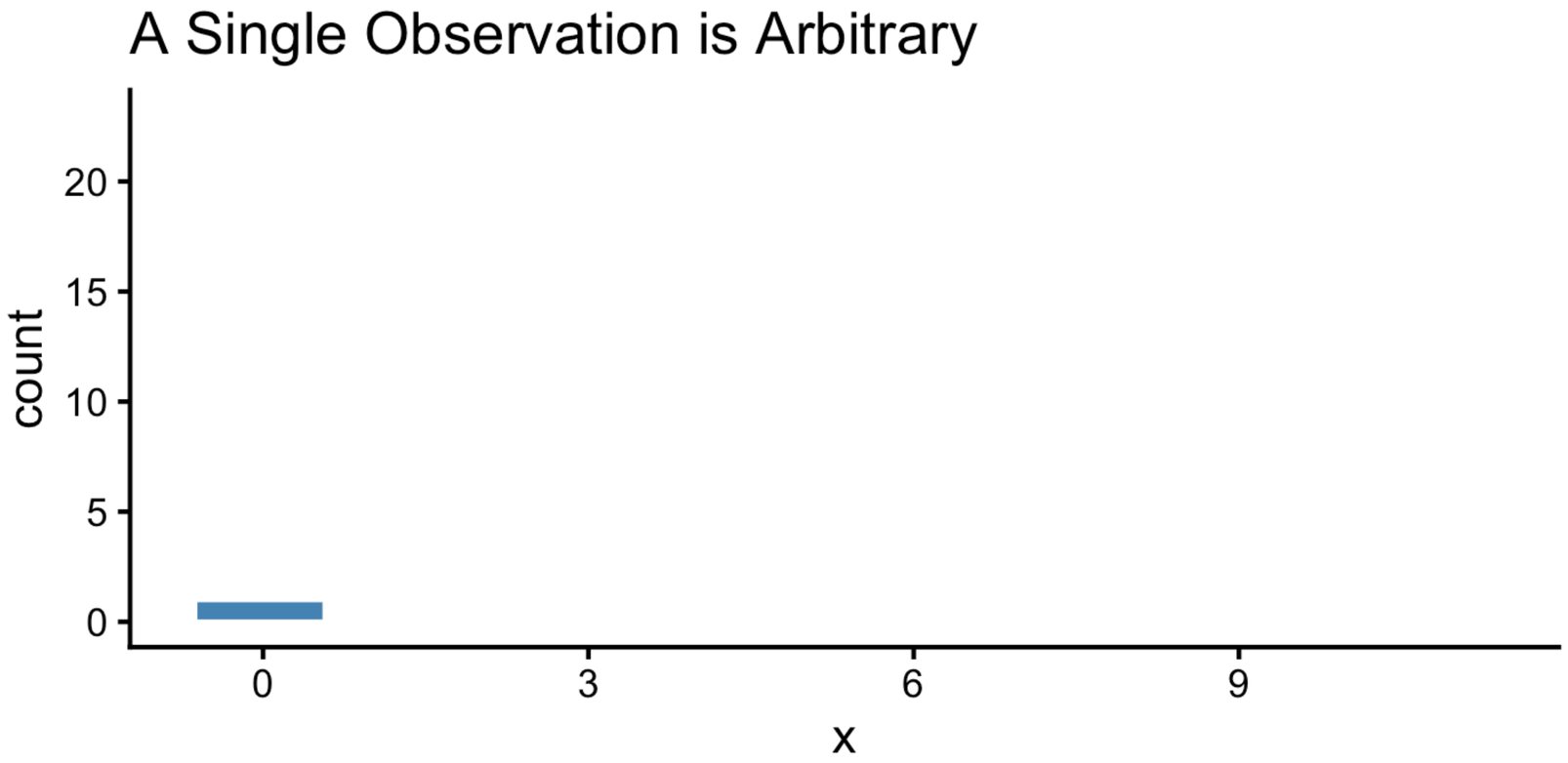
1. What is the true average card value in a deck of cards?
2. How does the distribution of sample means compare to the distribution of card values in a deck of cards?
3. What is the relationship between the centers of the two distributions?
4. Which distribution appears to have more variability?
5. How do the shapes of the two distributions compare? Why do they differ?



Sampling Overview

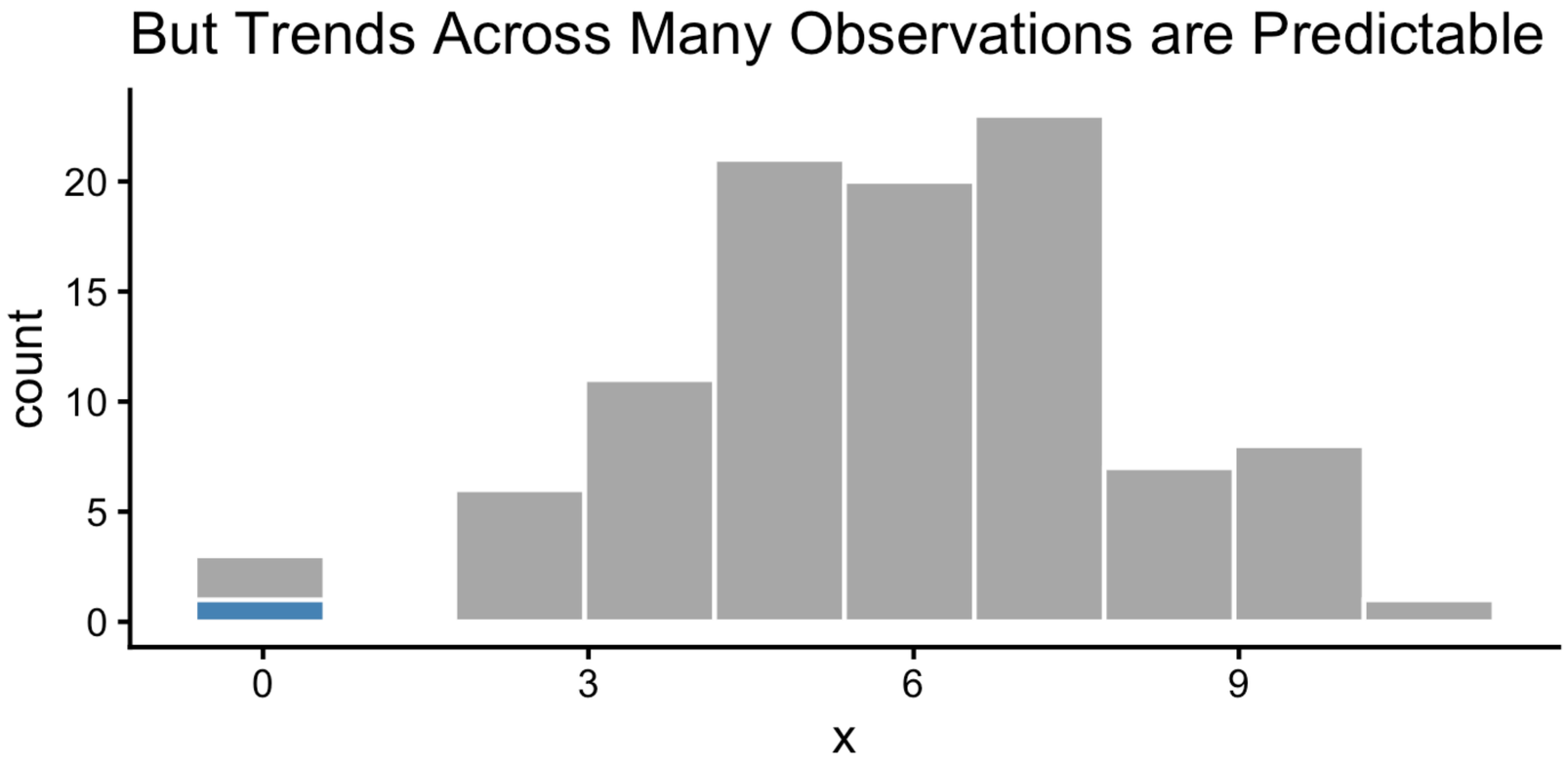
Sampling Overview

- The distribution of a data set allows us to quantify the shape, center, and spread of the data.
- While a single observation in a data set may appear arbitrary…



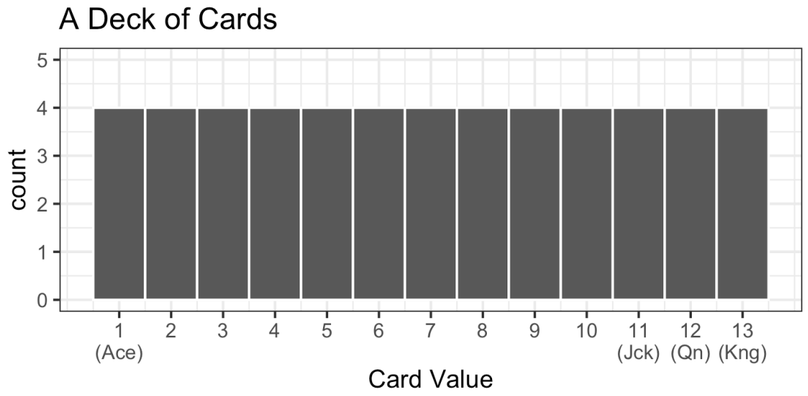
Sampling Overview

- The distribution of a data set allows us to quantify the shape, center, and spread of the data.
- While a single observation in a data set may appear arbitrary… repeated trials often show that outcomes follow certain patterns.



Sampling Overview

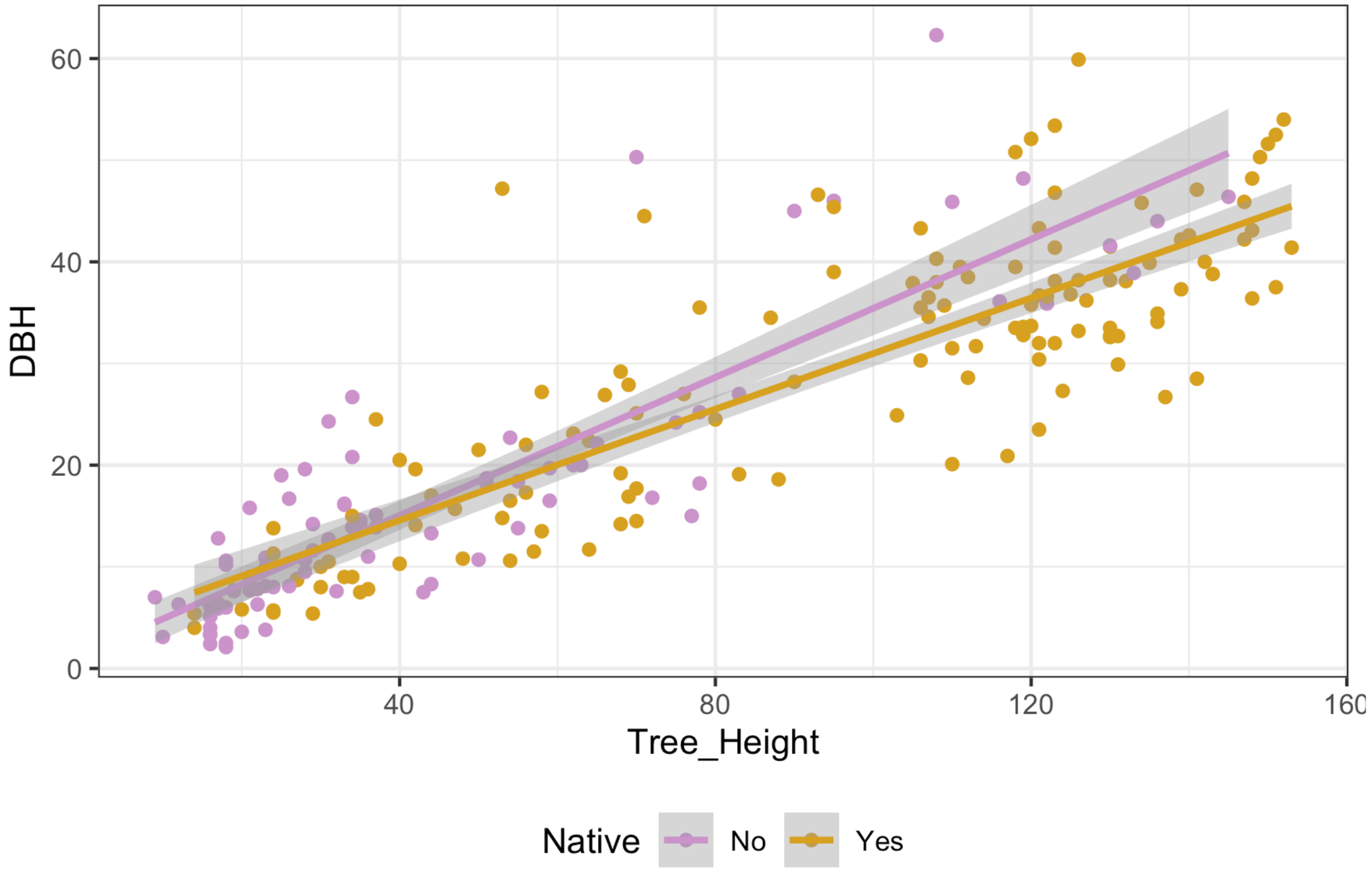
- We know that a variable from a population has a distribution
  - e.g., the distribution of card values in a deck of cards



- **Moral of Today:** Statistics (e.g., mean of variable in a sample) have distributions too!!
  - How? We only have ONE statistic – the statistic from the ONE sample we drew
  - Distribution is **over all the possible samples we could have drawn** (we just see 1)
- **Implication:** Statistics themselves have a *mean*, *standard deviation*, *5-number summary*
  - The mean tells us the statistic's typical value in a randomly chosen sample.
  - The standard deviation tells us how the statistic fluctuates from sample to sample.
- **Very Powerful:** e.g., Can use this distribution to give **plausible** ranges for the parameter and a sense of **uncertainty** in a given statistic.

Bigger Picture - Quantifying Our Uncertainty

R has been giving us uncertainty estimates (ex. `geom_smooth` when we don't set `se = FALSE`):


60
40
20
0
DBH
40
80
120
160
Tree_Height
Native
No
Yes

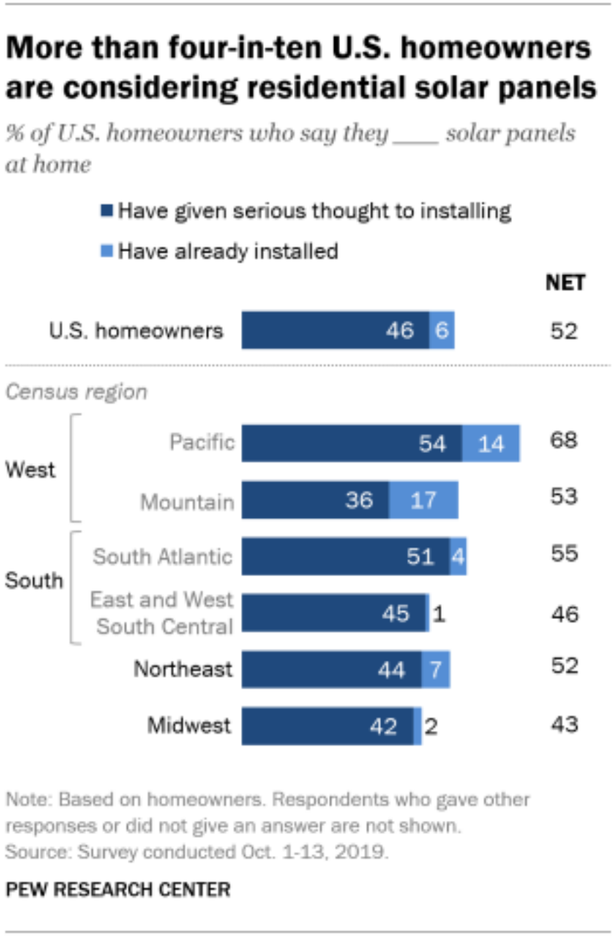
Bigger Picture - Quantifying Our Uncertainty

R has been giving us uncertainty estimates (ex. `std_error` in summaries from `lm()`):

```
# A tibble: 4 × 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept        7.39      3.65       2.03   0.044    0.201   14.6
2 DBH              2.25      0.17      13.3    0        1.92     2.59
3 Native: Yes     11.1       5.59       1.98   0.049    0.067   22.1
4 DBH:NativeYes    0.315     0.215      1.47   0.144   -0.108    0.739
```

Bigger Picture - Quantifying Our Uncertainty in Statistics

Uncertainty estimates are constantly reported in news and journal articles:



Note: The findings are based on a survey conducted Oct. 1-13, 2019, among 3,627 U.S. adults on Pew Research Center's American Trends Panel. The margin of sampling error for the full sample is plus or minus 2.1 percentage points. The margin of error for the 2,564 U.S. homeowners is plus or minus 2.5 percentage points. See full topline results.

Bigger Picture - Quantifying Our Uncertainty in Statistics

Uncertainty estimates are constantly reported in news and journal articles:

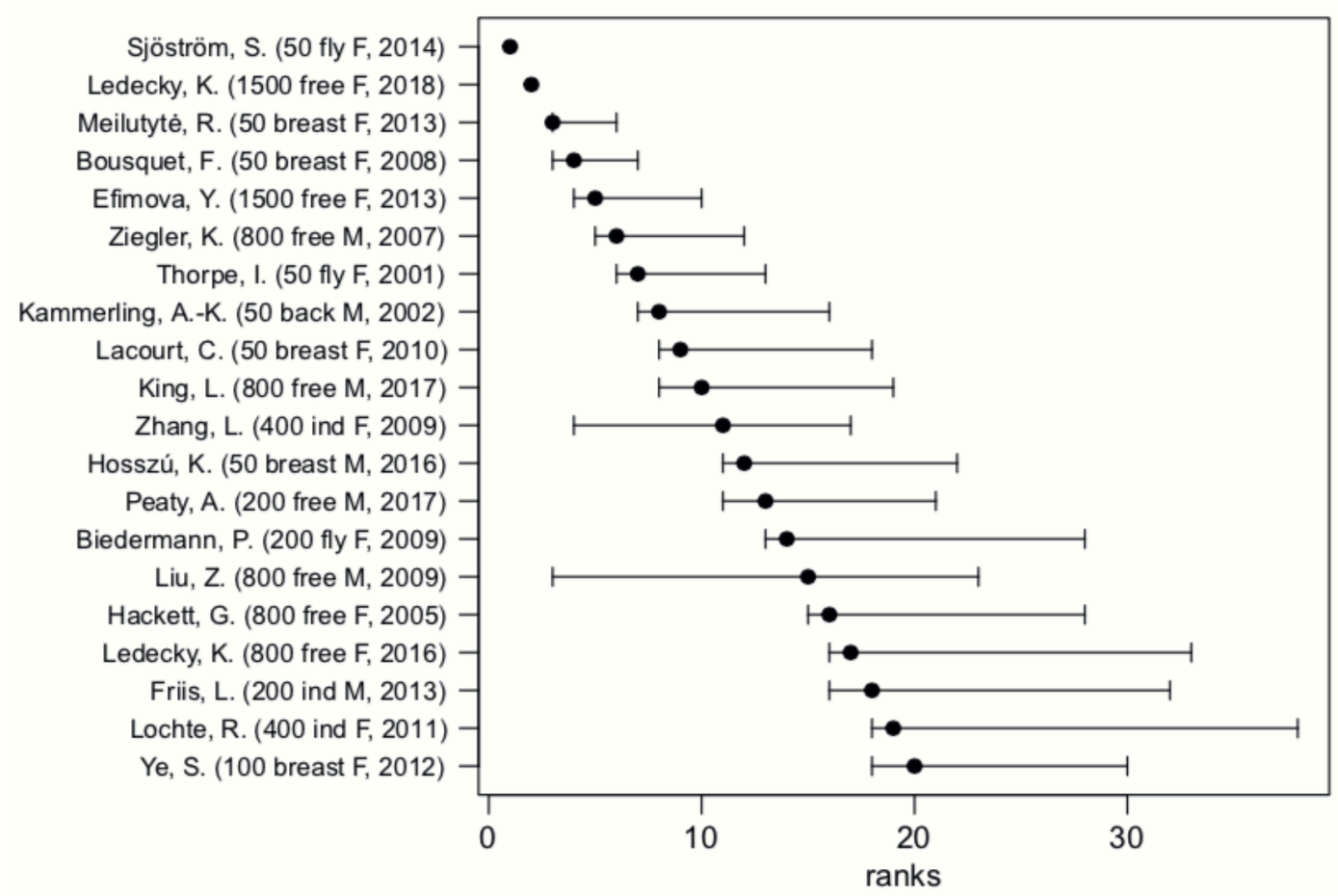
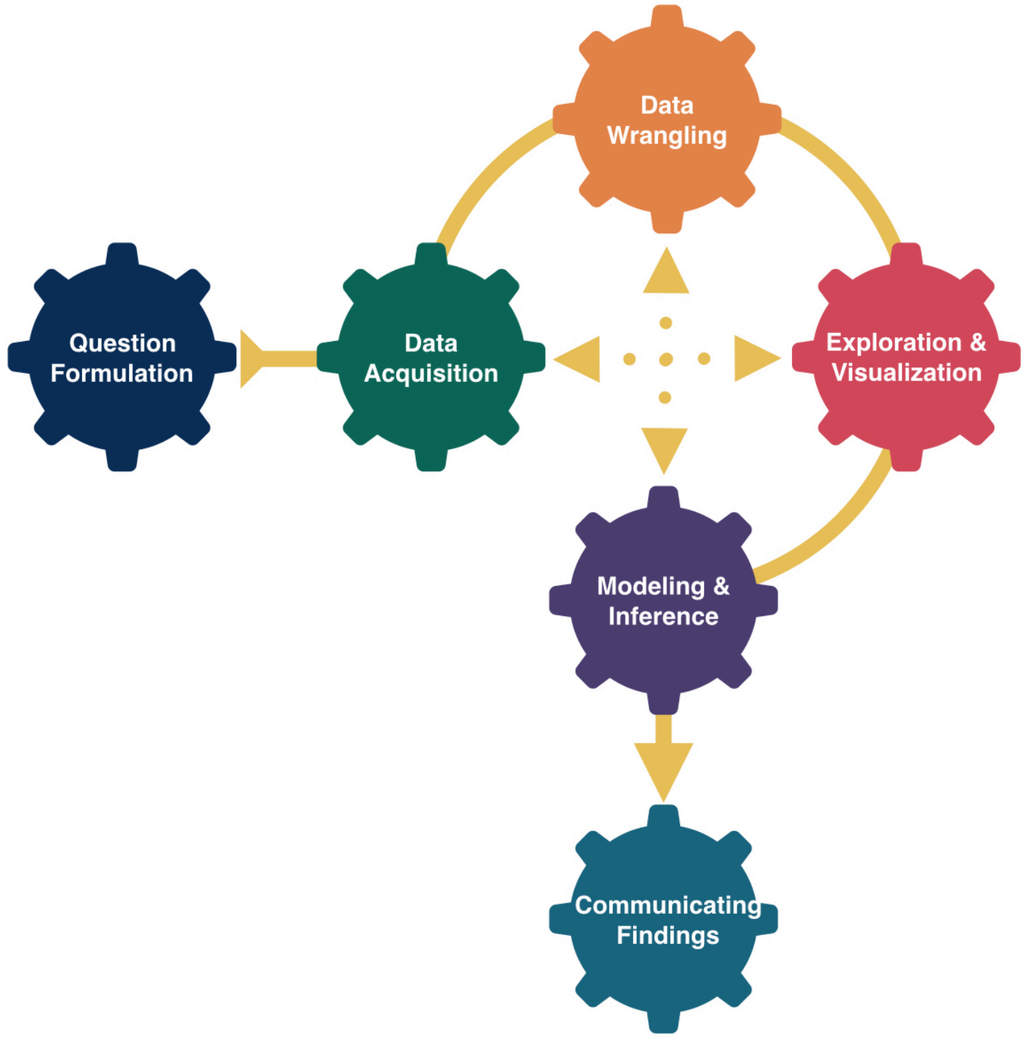


Figure 2: Ranking of the top 20 personal best swims (and their swimmers), 2001–2018, with 95% confidence intervals. Better-ranked swimmers are higher on the *y*-axis.

Statistical Inference

- **Goal**: Draw conclusions about the population based on the sample.
  - We've seen how to calculate a statistic from our sample
  - But different samples give different statistics: how do we know whether to trust the one we have?
  - Sampling distributions show how widely our statistic can range across samples
  - This helps us understand how much to trust any single statistic



Sampling Distributions II

Megan Ayers

Math 141 | Spring 2026

Monday, Week 6

Announcements

- Midterm studyguide posted on course website
- We'll work on the study guide in lab this week in groups - it would be helpful to take a look before lab but you don't need to prepare any work

Goals for Today

- Discuss the framework for random sampling
- Investigate properties of the “sampling distribution”
- Week 4 feedback (if time)

The ❤️ of statistical inference is quantifying uncertainty

Like with regression, need to distinguish between the **population** and the **sample**

- **Parameters**:
  - Based on the **population**
  - Unknown if we don't have data on the whole population
  - EX: β_0 and β_1
  - EX: μ (population mean) p (population proportion)
- **Statistics**:
  - Based on the **sample** data
  - Known
  - Usually estimate a population parameter
  - EX: $\hat{\beta}_0$ and $\hat{\beta}_1$
  - EX: $\bar{x}$ (sample mean), $\hat{p}$ (sample proportion)

Estimation and Uncertainty

- We are interested in the value of a **parameter** in a *population*, and use a **statistic** from a *sample* to estimate the parameter.
  - e.g., want to know the proportion (p) of Reed students with March birthdays
  - Estimate p by using the proportion in a sample, ($\hat{p}$), of 100 individuals.
- **Key question:** How accurate is $\hat{p}$ as an estimate of p?
- **Sub-question:** If we take many samples, how much will $\hat{p}$ vary?
- The distribution of all possible values of $\hat{p}$ (for a fixed sample size) is called the **Sampling Distribution**

Sampling Distribution of a Statistic

Steps to Construct a Sampling Distribution:

1. Decide on a sample size, n.
2. Determine all* possible samples of size n from the population.
3. Compute the sample statistic in each sample.

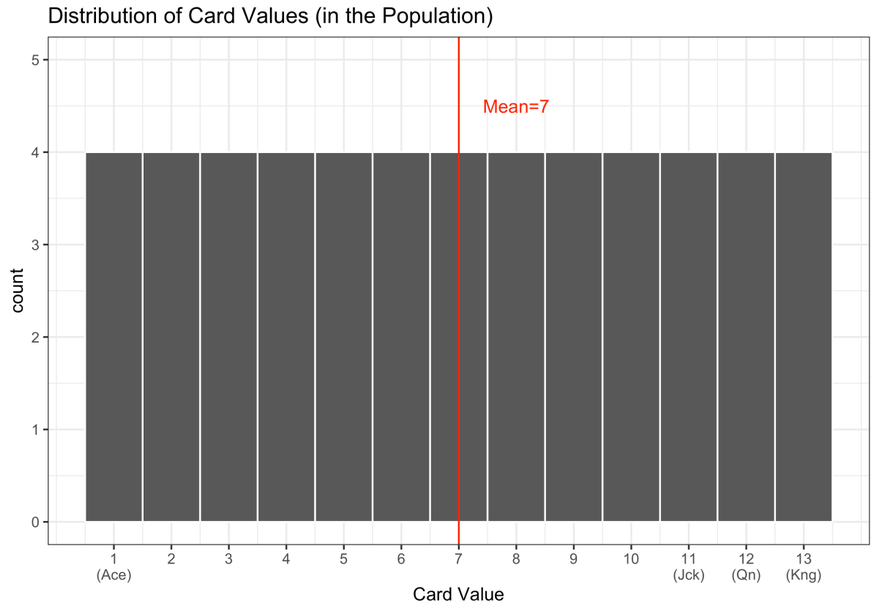
Approximate Sampling Distribution of a Statistic

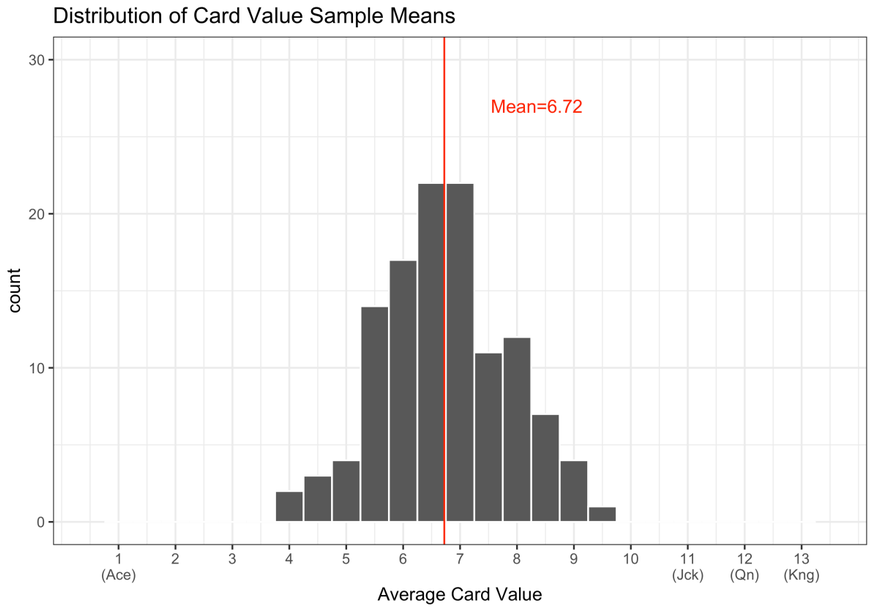
Steps to Construct an **Approximate** Sampling Distribution:

1. Decide on a sample size, n.
2. **Randomly select a sample of size n from the population.**
3. Compute the sample statistic in that sample.
4. **Put the sample back in.**
5. Repeat Steps 2 - 4 many (1000+) times.

Population and Sampling Distributions in Our Activity

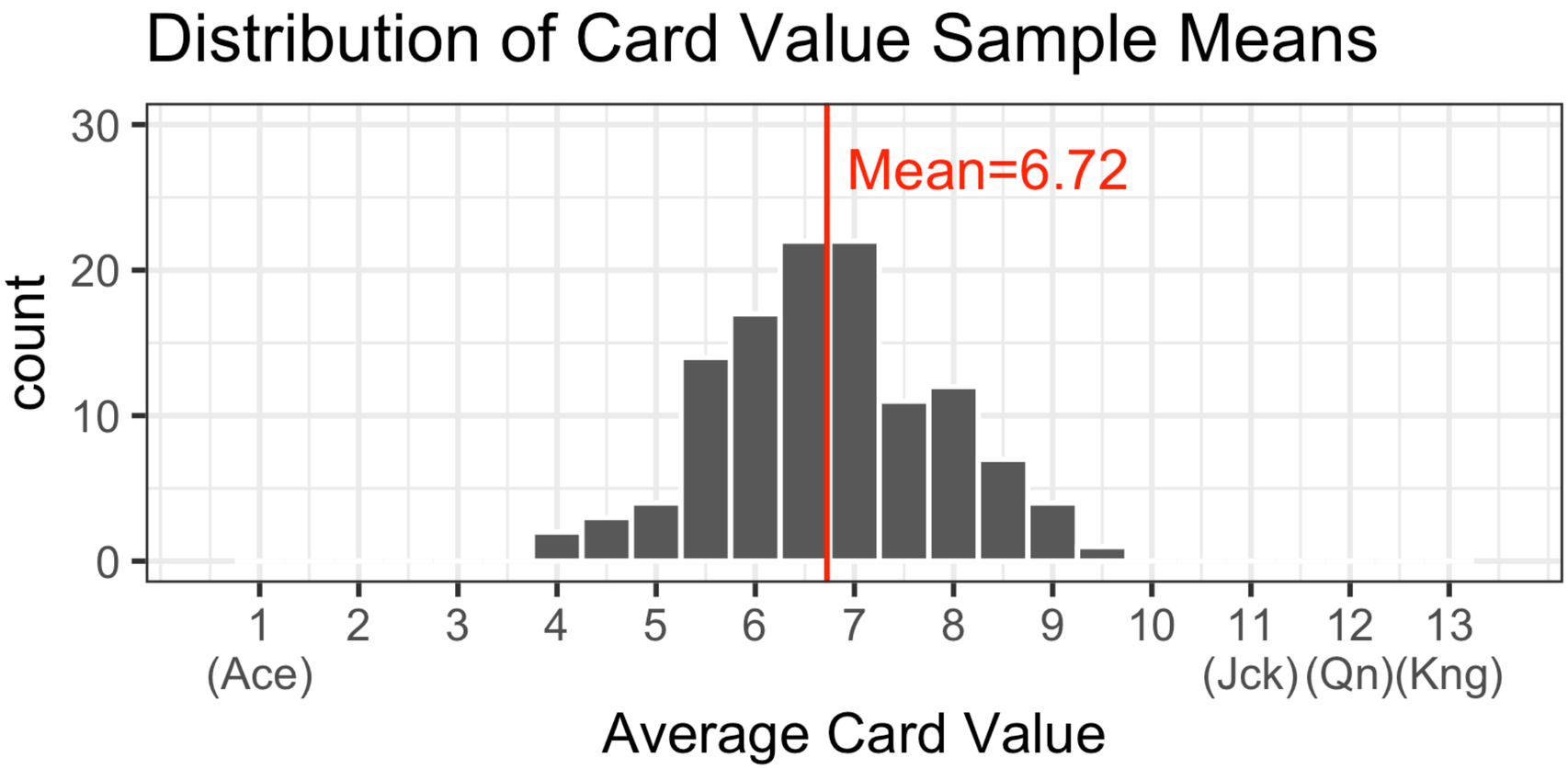
Last Class: got a "snapshot" of the sampling distribution for the **mean card value from a sample of 10 cards**.





Sampling Distribution in Our Activity

"Snapshot" of sampling distribution for **mean card value from a sample of 10 cards**:



Why just a "snapshot"?

- **Q:** How many elements are there in the sampling distribution?

$$\binom{52}{10} = 15,820,024,220 \text{ Samples}$$

- We only took 119 samples (combining both sections)…

Sampling Distribution: a smaller scale

Consider a **population of 4 students**, where the (true) population proportion of March birthdays is $p = \frac{1}{4} = 0.25$

name	bday_month
Joel	march
Maria	january
Arthur	april
Klaus	september

- **Practice:** With a neighbor, write out the sampling distribution for $\hat{p}$ (the sample proportion of those with March birthdays) with…

  1. All possible samples of size 2
  2. All possible samples of size 3

- And, what are the means of these two sampling distributions?

Sampling Distribution: a smaller scale

1. Sampling distribution for the sample proportion $\hat{p}$ from a sample of 2:

name	bday_month
Joel	march
Maria	january
Arthur	april
Klaus	september

sample	phat
Joel + Maria	0.5
Joel + Arthur	0.5
Joel + Klaus	0.5
Maria + Arthur	0.0
Maria + Klaus	0.0
Arthur + Klaus	0.0

- There are $\binom{4}{2} = 6$ elements in the sampling distribution
- **Mean=0.25**, which is p!

Sampling Distribution: a smaller scale

2. Sampling distribution for the sample proportion $\hat{p}$ from a sample of 3:

name	bday_month
Joel	march
Maria	january
Arthur	april
Klaus	september

sample	phat
Joel + Maria + Arthur	1/3
Joel + Maria + Klaus	1/3
Joel + Arthur + Klaus	1/3
Maria + Arthur + Klaus	0

- There are $\binom{4}{3} = 4$ elements in the sampling distribution
- **Mean=0.25**, which is p! (again)
- **Note**: The means of the sampling distributions were all $p = \frac{1}{4}$ *even though none of the sample proportions* $\hat{p}$ *were equal to* $\frac{1}{4}$!

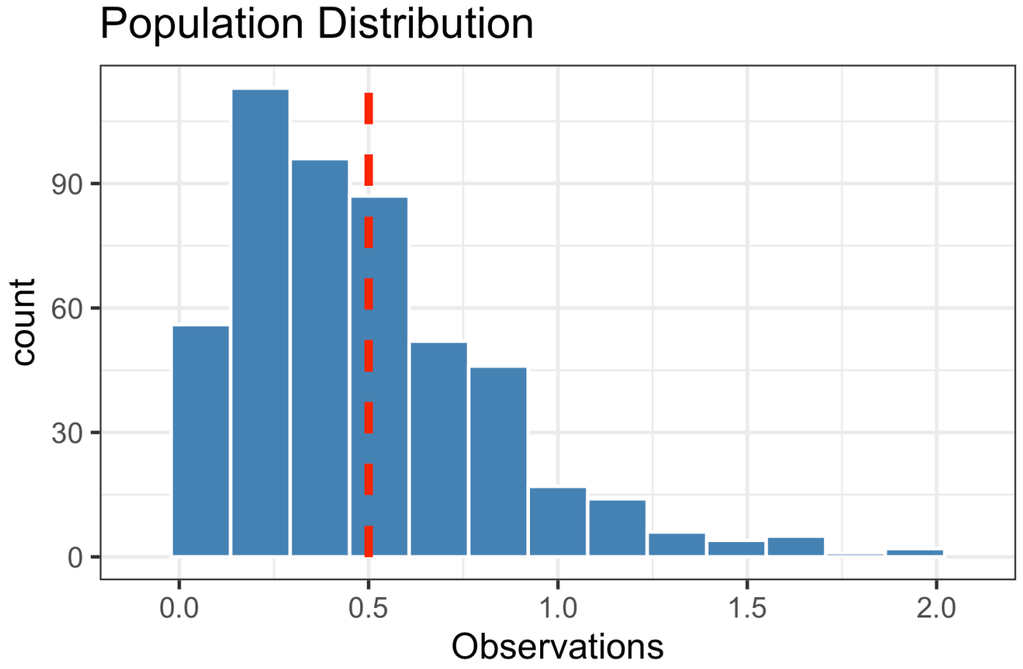
Sampling Distribution

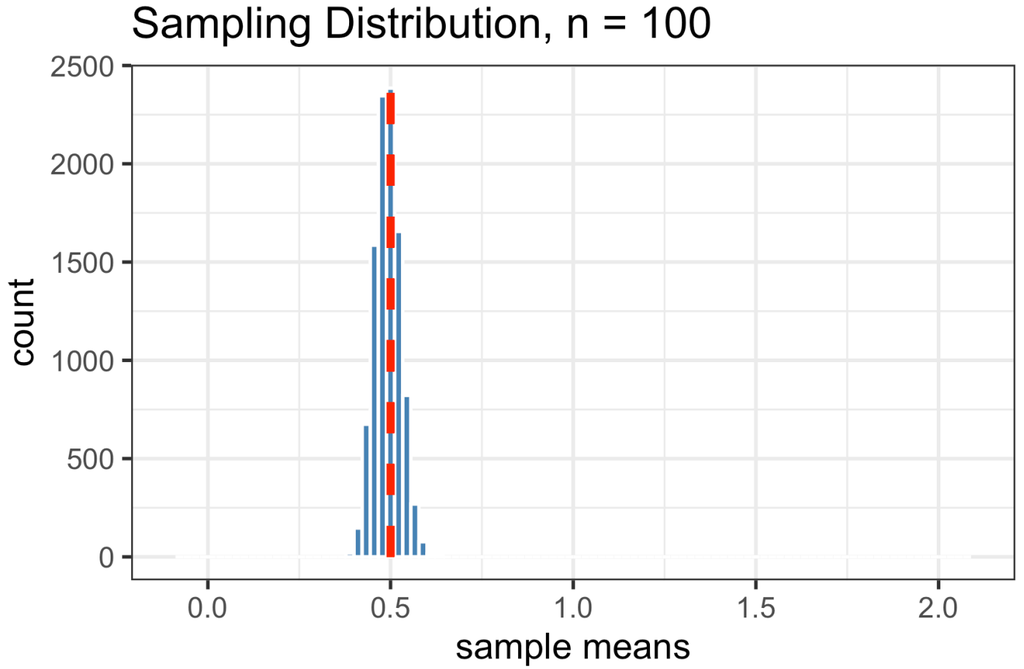
- Just like any distribution, the sampling distribution of a statistic has a mean, standard deviation, median, IQR, etc.
  - **Standard Error**: standard deviation in a sampling distribution
- Ex. When estimating a population proportion (p) with the sample proportion ($\hat{p}$) with sample size n, theory shows that:
  - The sampling distribution for $\hat{p}$ has mean p
  - and has standard deviation (aka standard error) $SE = \sqrt{\frac{p(1-p)}{n}}$
- This means if the population proportion $p = 0.20$ and $n = 100$, then in sampling distribution:
  - mean($\hat{p}$) = p = 0.20
  - $SE = \sqrt{\frac{0.20(1-0.20)}{100}} = 0.04$

Sampling Distribution vs. Population Distribution

For most sample statistics and sufficiently large sample sizes ($n \geq 30$ is a rule of thumb), **the sampling distribution will be approximately bell-shaped**, even if the population is not

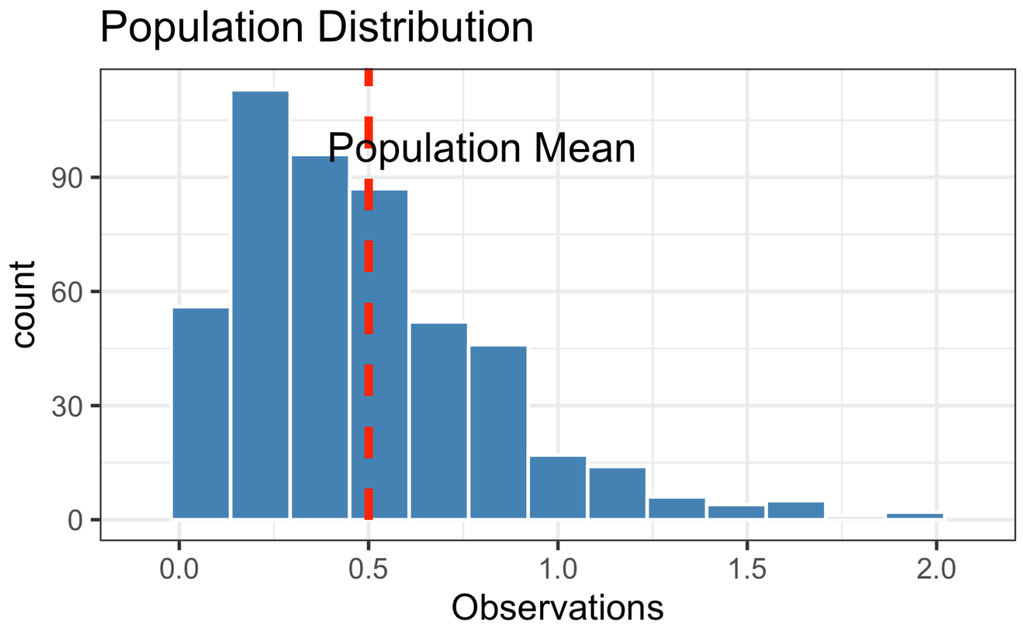
- Both distributions will have the same center.
- But, the sampling distribution will have lower variability than the population distribution.



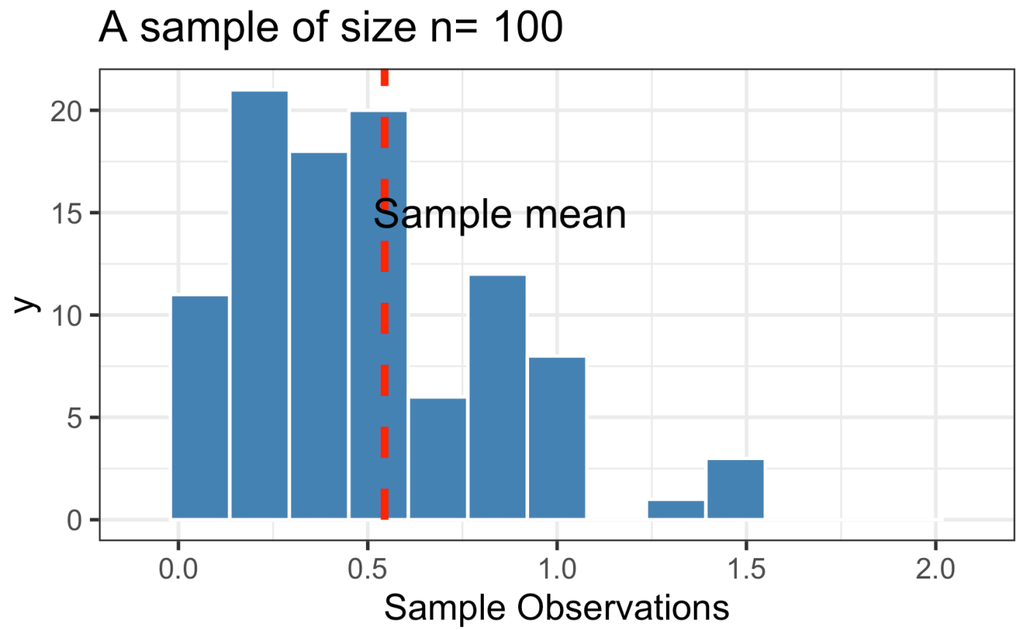


Why are Sampling Distributions Useful?

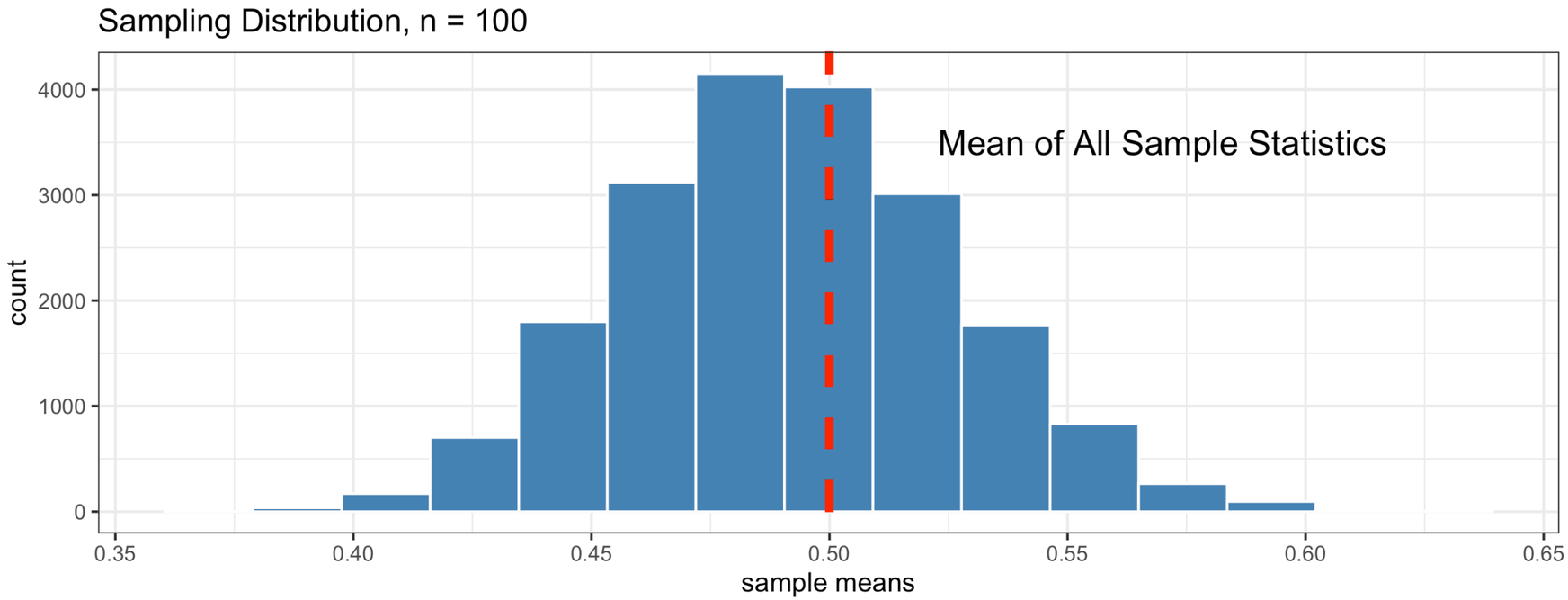
What we want to know:



What we have:



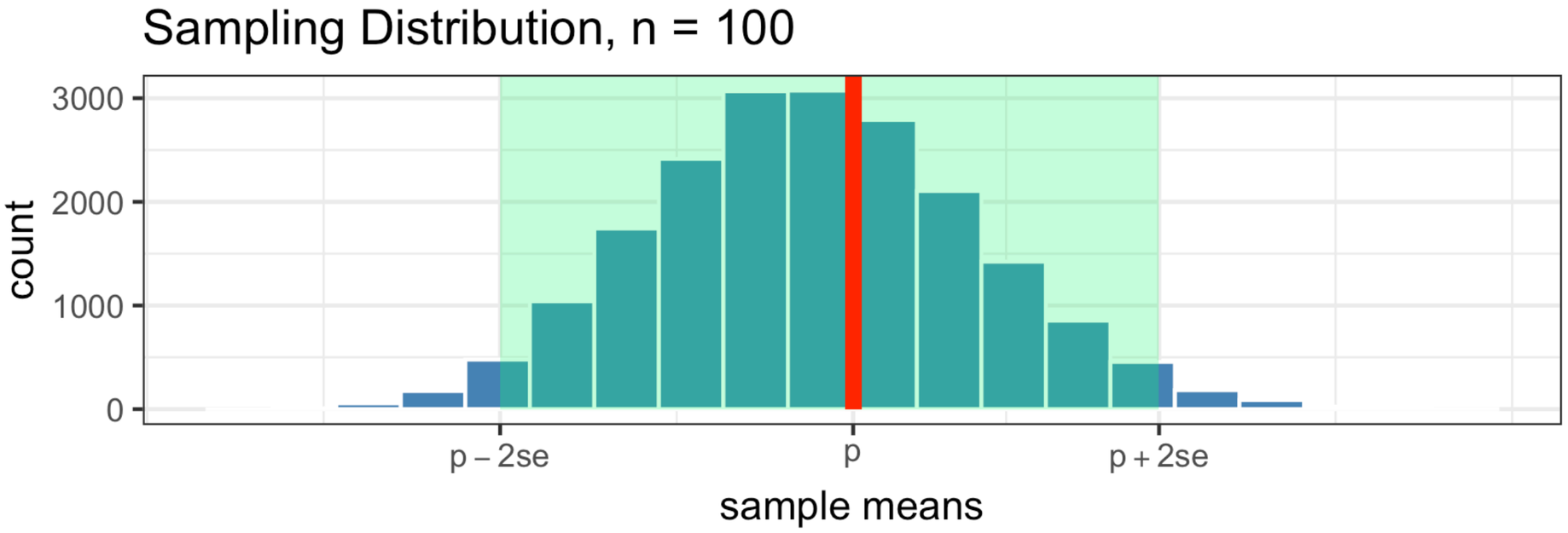
What sampling distributions tell us about what we have:


Sampling Distribution, n = 100
Mean of All Sample Statistics
count
4000
3000
2000
1000
0
0.35
0.40
0.45
0.50
0.55
0.60
0.65
sample means

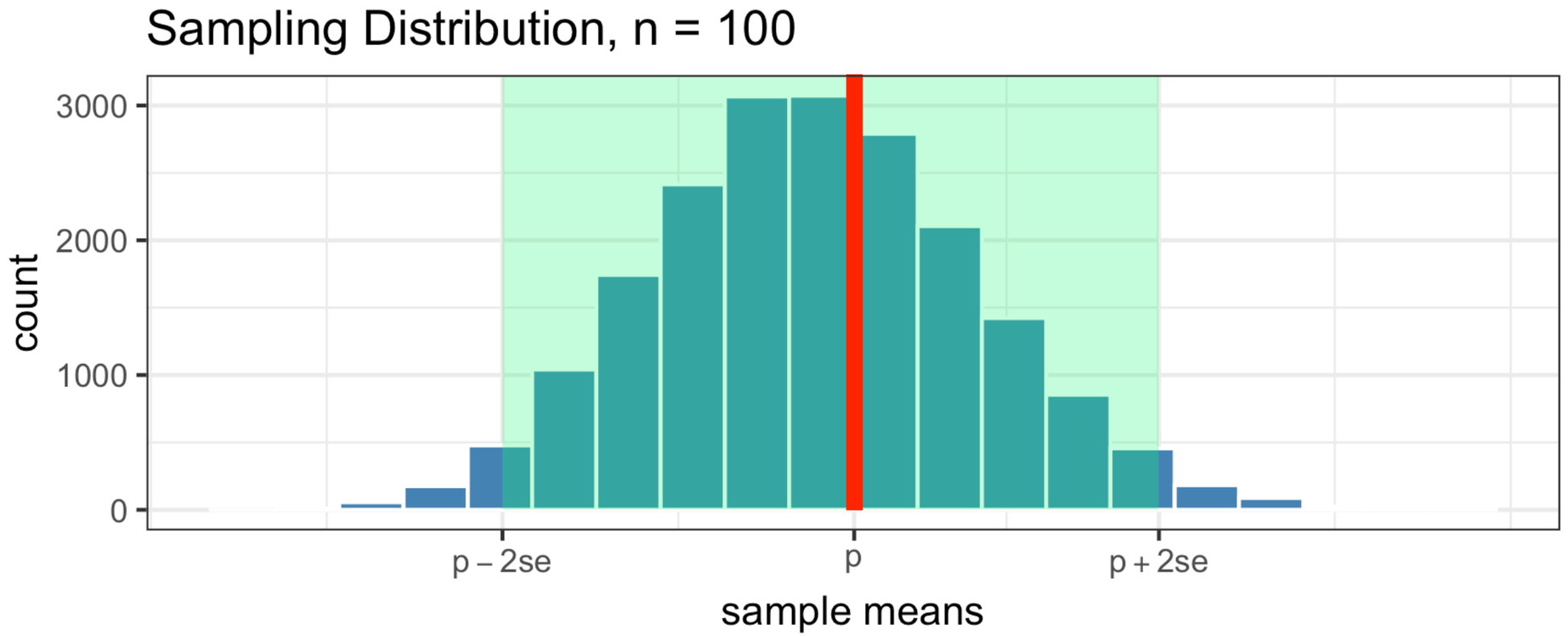
Variability in Samples

The **standard error** (se) of a sample statistic measures variability between different samples.

- For ≈ bell-shaped distributions, about 95% of observations fall within two standard errors of the population's mean μ.
- **Very useful implication!**
  - The sampling distribution for the sample average, $\bar{x}$ is ≈ bell-shaped
  - … and is centered at the population mean, μ.
- So **95% of all sample statistics, $\hat{p}$, fall within 2 standard errors of p!**



This is very powerful!

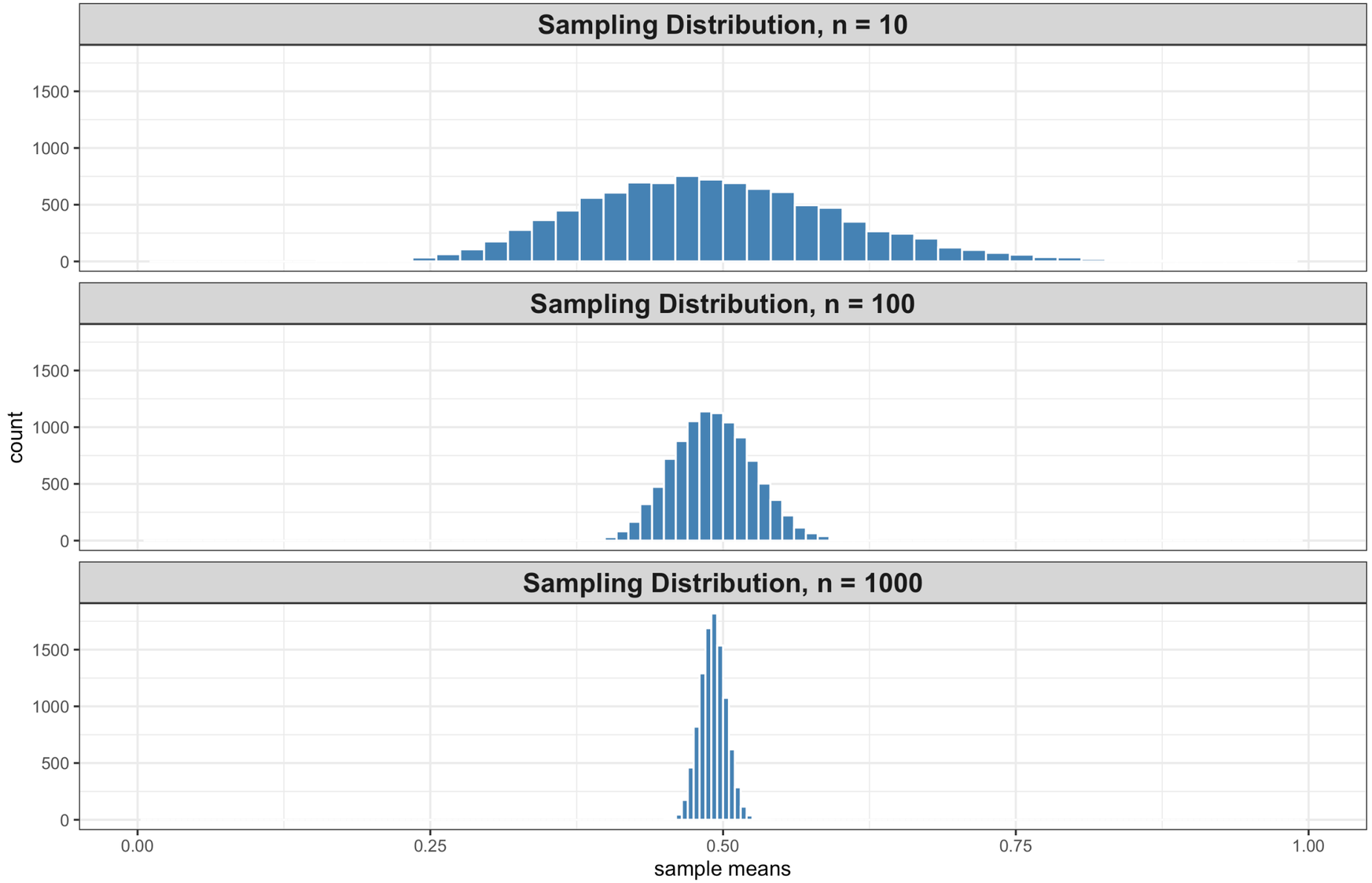


Even though we just have one sample, if we know se, **we know how far away from the parameter our statistic is likely to fall**!

- This allows us to give quantitative plausible ranges for the parameter

Variability and Sample Size

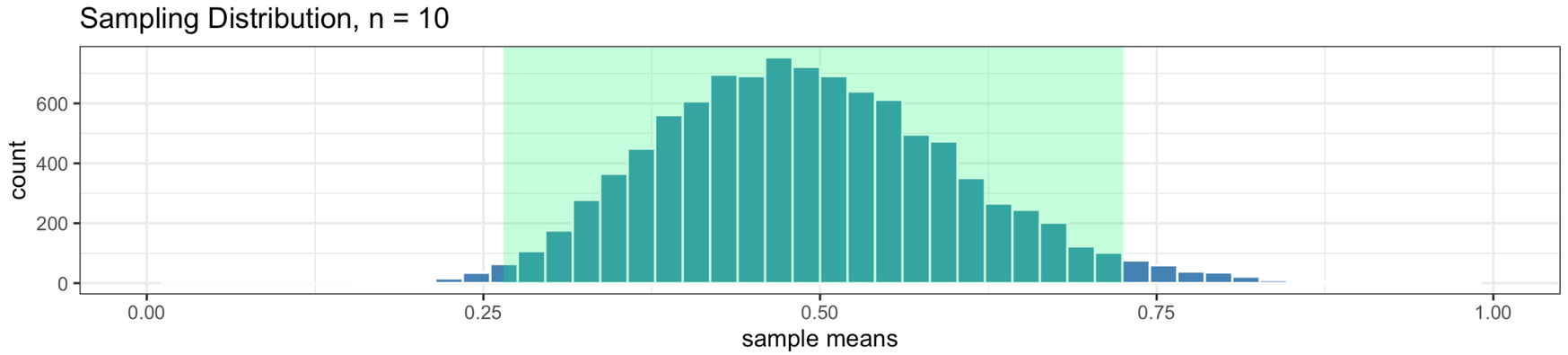
Variability of the sampling distribution generally decreases as sample size increases

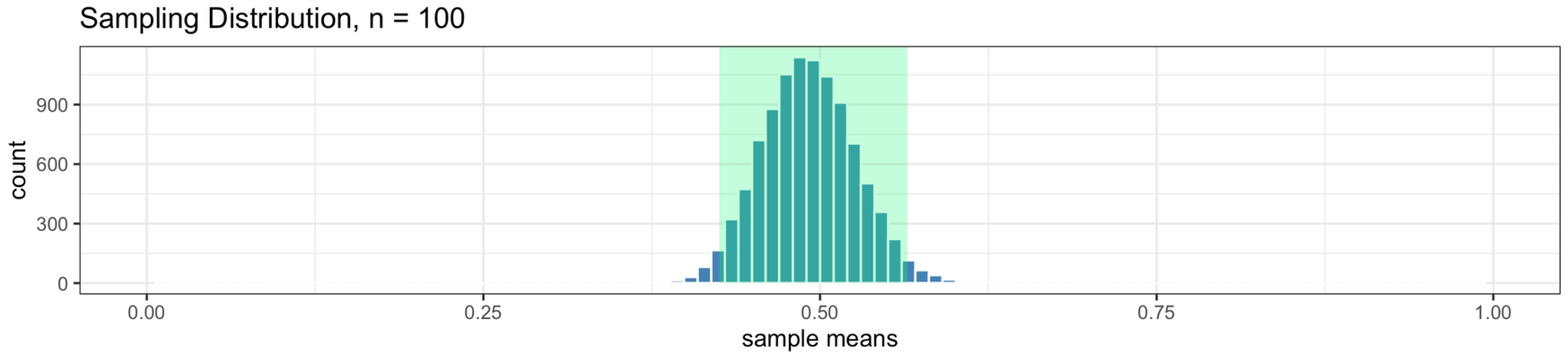


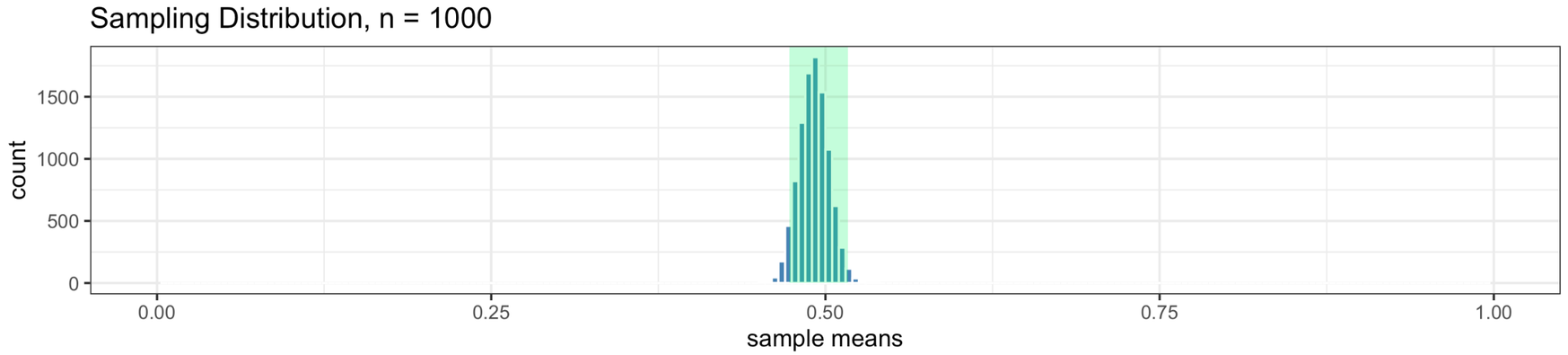
- **Implication:** Our statistic $\bar{x}$ is likely to be closer to parameter μ as sample size n increases.

Variability and Sample Size

- The sampling distributions for $n = 10, 100,$ and 1000 are all approximately bell-shaped, and so 95% of sample means are within 2 standard errors of the population mean.
- Highlighted in green are the intervals containing 95% of all sample means:


Sampling Distribution, n = 10
count
600
400
200
0
0.00
0.25
0.50
0.75
1.00
sample means

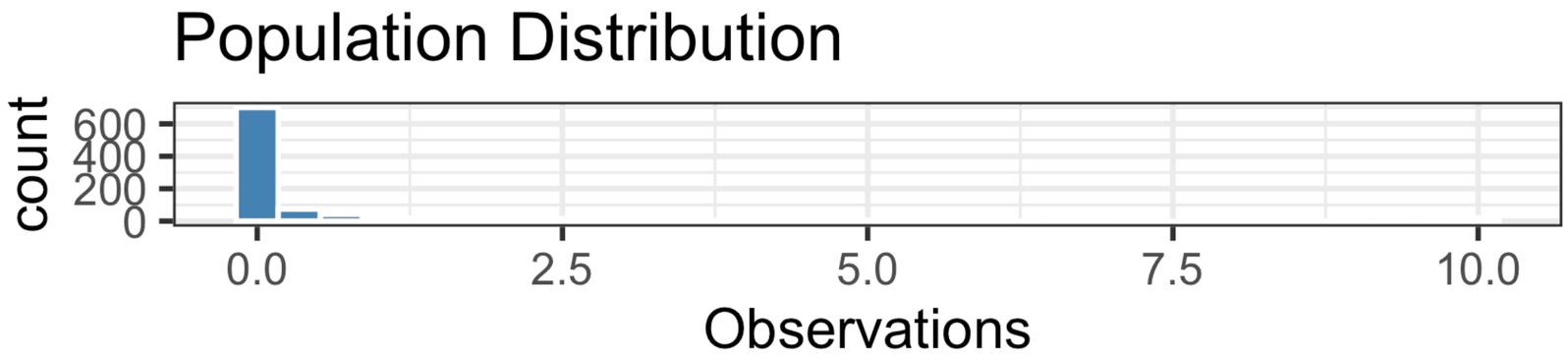

Sampling Distribution, n = 100
count
900
600
300
0
0.00
0.25
0.50
0.75
1.00
sample means

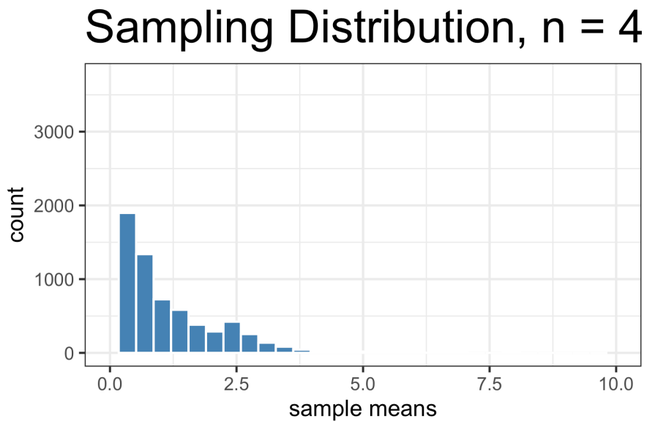

Sampling Distribution, n = 1000
count
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
sample means

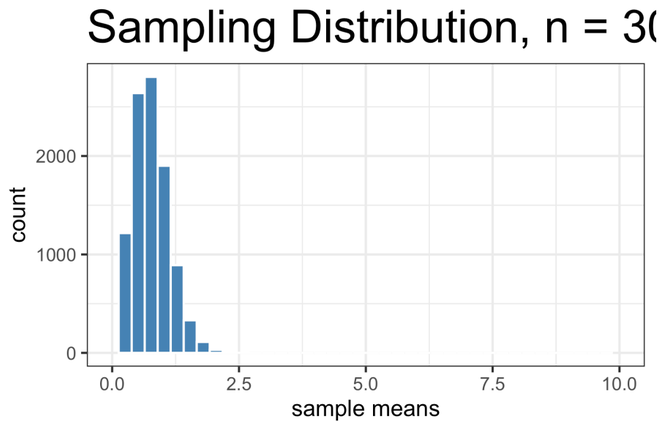
The Shape of the Sampling Distribution: word of warning

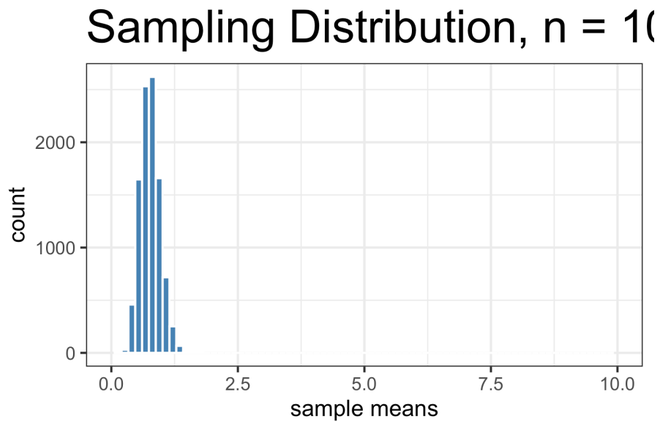
When $n \geq 30$, the sampling distribution is *usually* bell-shaped. But sometimes, you need a large sample size.

- Example: A funky population distribution.









Key Features of a Sampling Distribution

What did we learn about sampling distributions?

- Centered around the true population parameter.
- As the sample size increases, the **standard error** (SE) of the statistic decreases.
- As the sample size increases, the shape of the sampling distribution becomes more bell-shaped and symmetric.

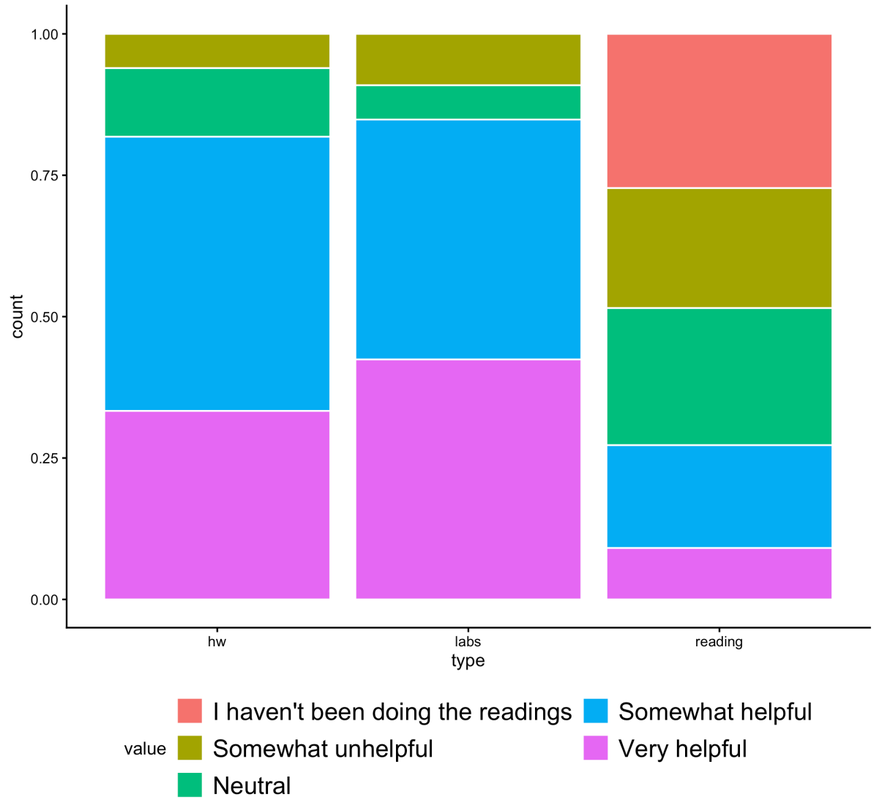
Question:

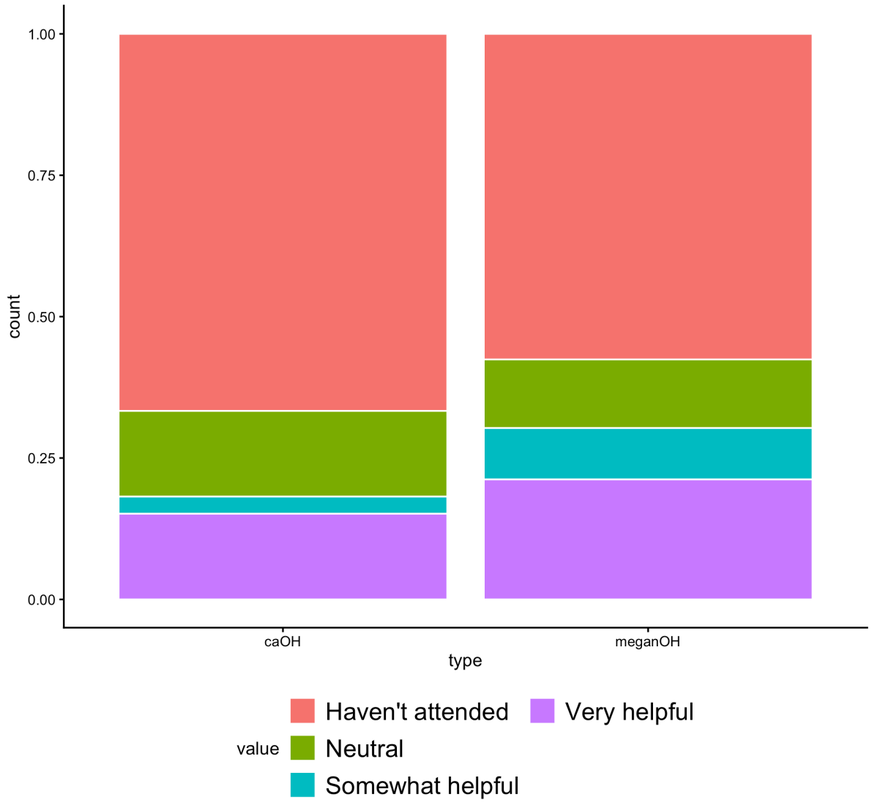
If I am estimating a parameter in a real example, why won't I be able to construct the sampling distribution?

Cliffhanger:

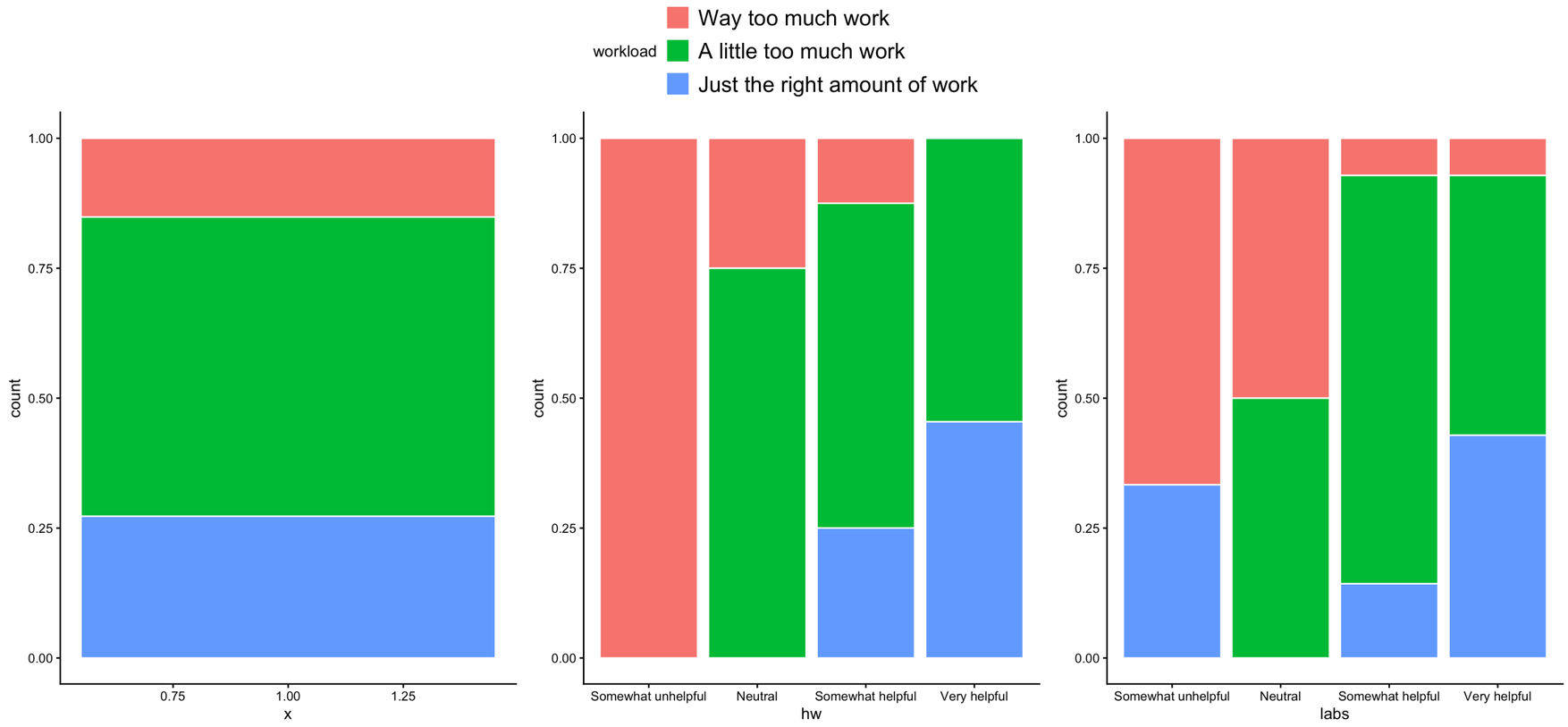
How can we learn from the sampling distribution if we only have one sample?

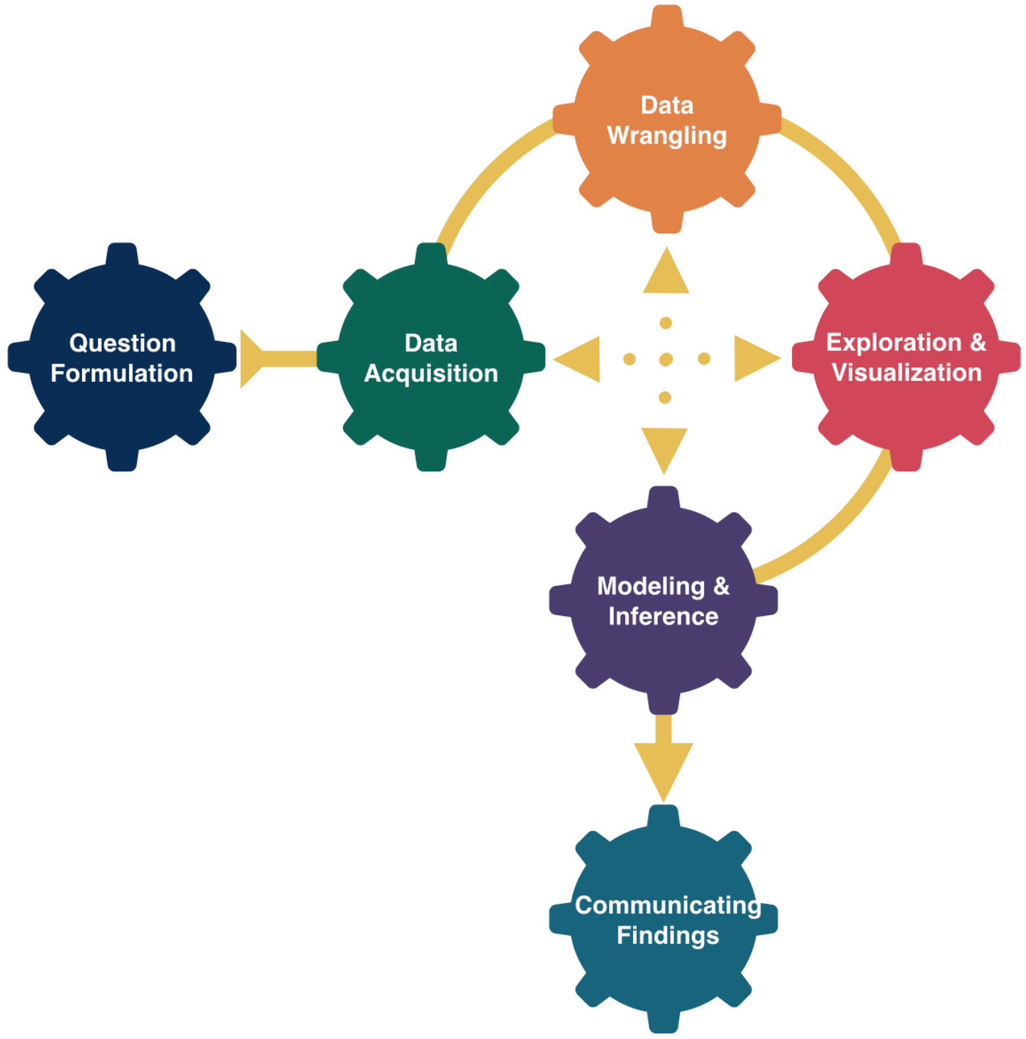
Week 4 Feedback (n = 33)





Week 4 Feedback (n = 33)





Bootstrapping

Megan Ayers

Math 141 | Spring 2026

Wednesday, Week 6

Midterm Check In

- If you have exam accommodations and have **not** discussed them with me via email, please let me know asap!

Goals for Today

- Review how sampling distributions can be used to assess sampling variability
- Discuss bootstrapping as a way of approximating the sampling distribution

Sampling Distribution: Polling Example

An October 2020 poll by Marist College surveyed by phone, asking

> If November's election were held today, whom would you support?

- 50% of respondents supported Biden/Harris, 46% supported Trump/Pence, 1% supported another candidate, and 3% were undecided
- In the Nov. 3 2020 election, Biden/Harris had 50.01% of the vote, while Trump/Pence had 48.84% of the vote.
  - **Population**: All registered voters in Pennsylvania ($N \approx 9$ million)
  - **Population Parameter**: The proportion p of registered voters who plan to vote for Trump/Pence. Based on election results, $p = 0.4884$.
  - **Sampling Method**: SRS(?) of size $n = 1020$ obtained using phone-numbers
  - **Sample Statistic**: The sample proportion $\hat{p}$ of Americans who plan to vote for Trump/Pence. In this case, $\hat{p} = 0.46$.

Sampling Variability

How confident should we be in the accuracy of our estimate of $\hat{p} = 0.46$?

- There are about 9 million registered voters in Pennsylvania. Marist College surveyed only 1020 of them (0.01% of the population)
- We should be skeptical that our estimate is *exactly* equal to true proportion.
- But we should feel confident that our estimate is *close* to the true proportion.

Why?

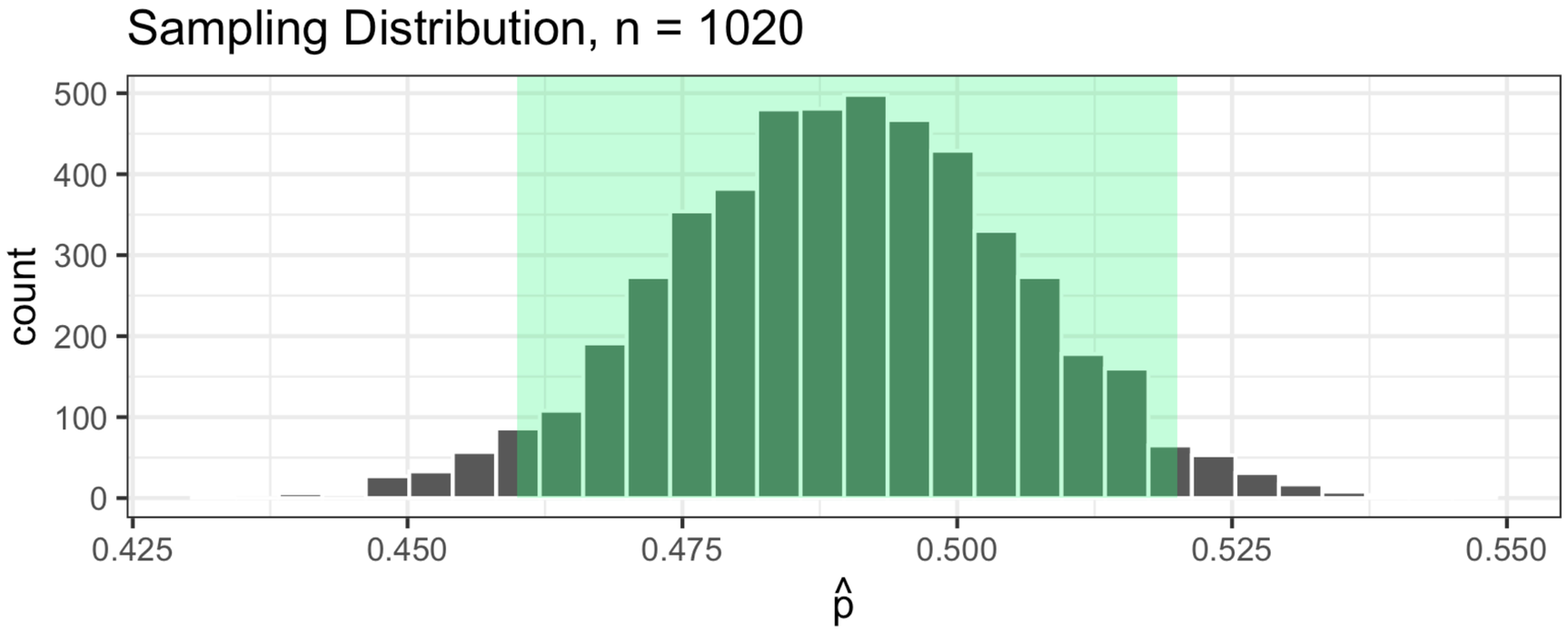
- The sampling distribution tells us how much variability to expect from sample to sample.
- Using probability theory, we know standard error for the sampling distribution for the sample proportion with sample size n is $SE = \sqrt{\frac{p(1-p)}{n}}$
- With $n = 1020$ and $p = 0.4884$, the standard error is $SE \approx 0.016$.

Sampling Variability

Suppose the true proportion of support for Trump/Pence was actually $p = 0.49$

- **Reminder**: In our sample, $\hat{p} = 0.46$ (0.03 away from $p = 0.49$).

Let's draw 5000 simulated samples of size 1020 to see how many have $\hat{p}$ far from $p = 0.49$.



- In 95% of samples, the sample proportion $\hat{p}$ is at most 0.03 away from the true proportion p!
  - Implication: If 49% of voters support Trump/Pence, then $\approx 95\%$ of samples of size 1020 will show Trump/Pence's support, $\hat{p}$, to be $46\% \leq \hat{p} \leq 52\%$
  - **Q**: How does this contextualize conclusions based on the poll's sample statistic?

The Problem

- For sampling distributions that are ≈ bell-shaped, 95% of sample statistics will be within 2 standard errors of the true parameter.
- This helps us assess **how close** a sample **statistic** tends to be to the population **parameter**.
- But in practice, we don't know the sampling distribution!
  - In the polls example, we **guessed** what the true p was to get a sampling distribution.
  - This is helpful as a thought experiment, but what if our guess was way off?
  - In order to form the actual sampling distribution, we need to collect many, many samples.
  - In the real world, we usually only have **one sample!**
- **The fix?**

Bootstrapping

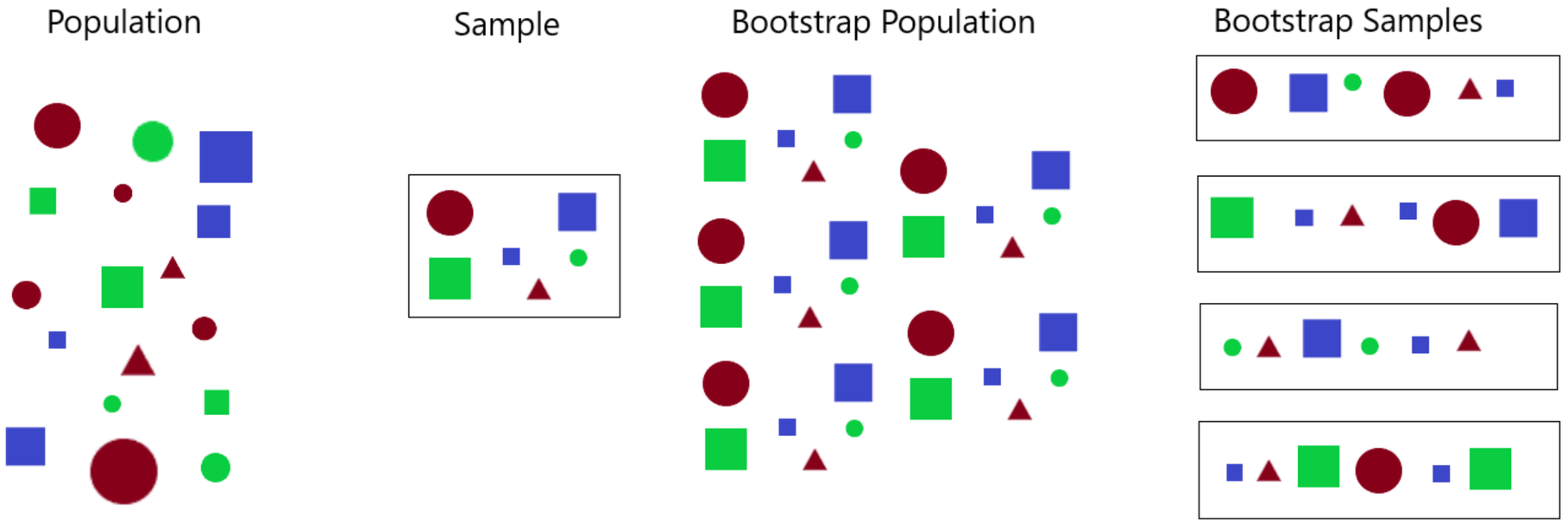
Bootstrapping

- The term *bootstrapping* refers to the phrase “to pull oneself up by one’s bootstraps”
- Originated in the 19th century as reference to a ludicrous or impossible feat
- By mid 20th century, meaning had changed to suggest a success by one’s own efforts, without outside help (the “American Dream” myth)
- Its use in statistics (dating from 1979) alludes to both interpretations.

The Bootstrap Trick

- The Impossible Task:
  - How can we learn about the sampling distribution, if we only have 1 sample?
- The "Ludicrous" Solution obtained without outside help:
  - Draw repeated samples from the **original sample**
  - Compute the statistic of interest in each new sample, and plot the resulting distribution
- The Main Idea:
  - The original sample approximates the population
  - Resampling from the sample approximates sampling many times from the population
  - The distribution of statistics from the resamples approximates the sampling distribution

Theory


Population
Sample
Bootstrap Population
Bootstrap Samples

The Bootstrap Procedure

To generate a **Bootstrap Distribution** given a sample of size n from the population,

1. Generate a **bootstrap sample** of size n by resampling *with* replacement from the original sample
2. Repeat (1) a large number of times (with technology, at least 1000 times)
3. For each bootstrap sample, calculate the appropriate statistic (called the **bootstrap statistic**)
4. The collection of the bootstrap statistics form the **Bootstrap distribution**.

Q: How does this process of generating a bootstrap distribution differ from the process of generating the *sampling* distribution?

- We sample from the original sample here; for sampling distributions, we sample from the population
- We sample with replacement here; for sampling distributions, we sample without replacement

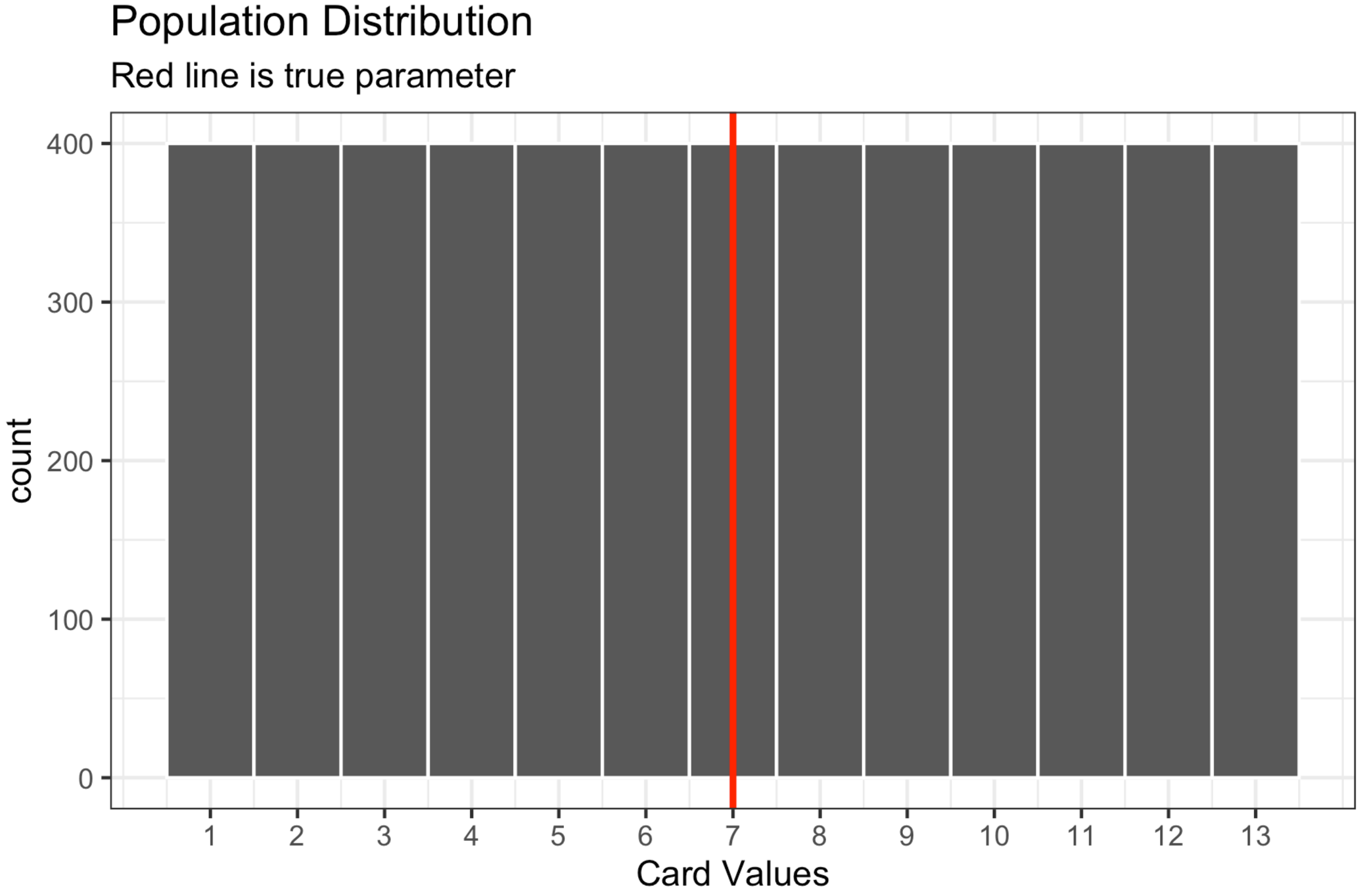
Proof of Concept

- **Population:** Consider a very large deck of cards (5200 cards) with 400 of each card value.
- **Question:** What's the mean value from the deck of cards?
- **How We'll Answer:**
  - Draw a sample hand of $n = 25$ cards, and calculate the mean value.
  - Estimate uncertainty using the bootstrap distribution.

Since we have the deck of cards, we can look at:

1. The population distribution
2. The sampling distribution for sample means
3. The single sample's distribution
4. The bootstrap distribution for sample means

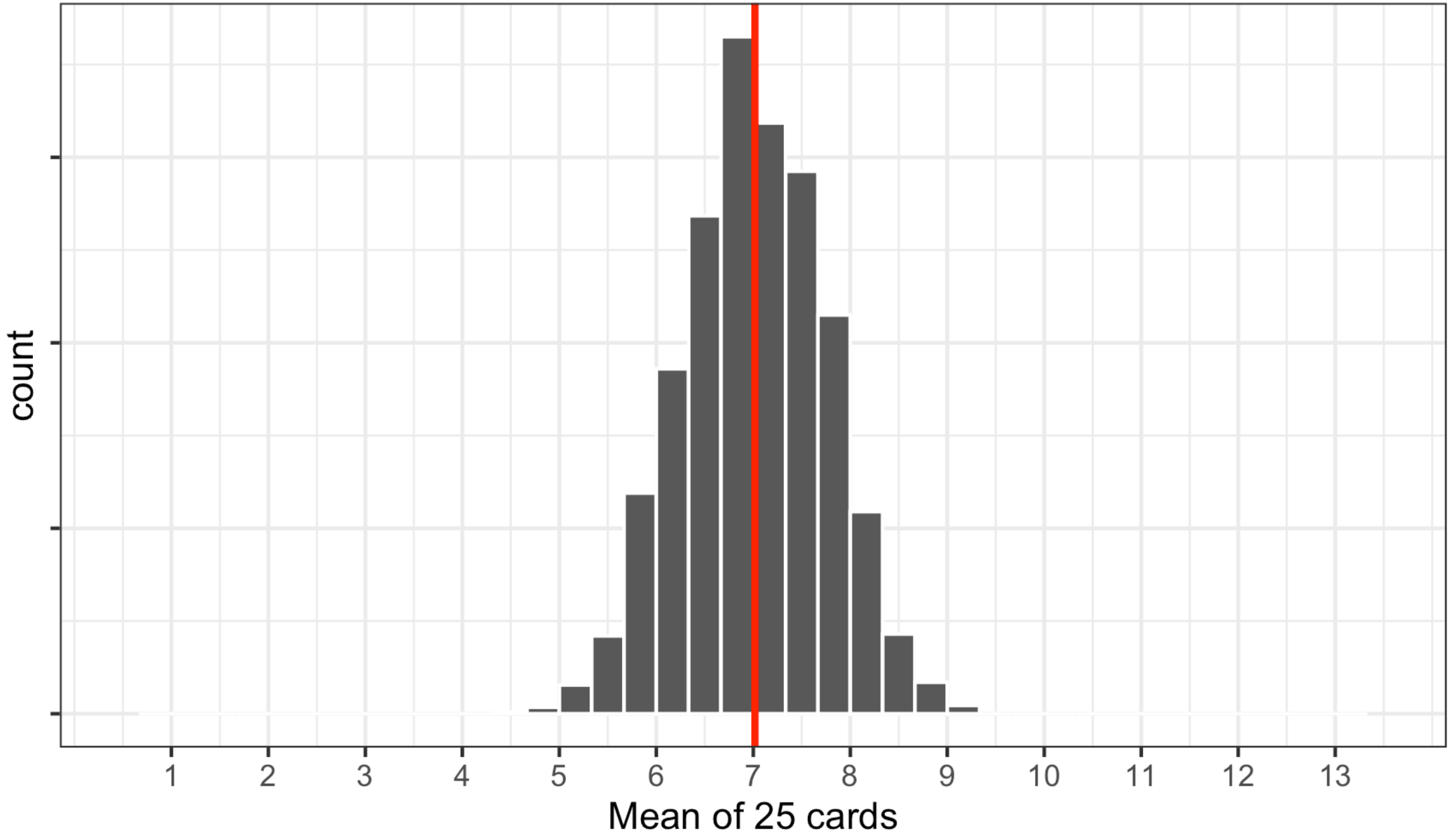
World's Largest Deck of Cards: Population


Population Distribution
Red line is true parameter
400
300
200
100
0
count
1
2
3
4
5
6
7
8
9
10
11
12
13
Card Values

World's Largest Deck of Cards: Sampling Distribution

Sampling Distribution

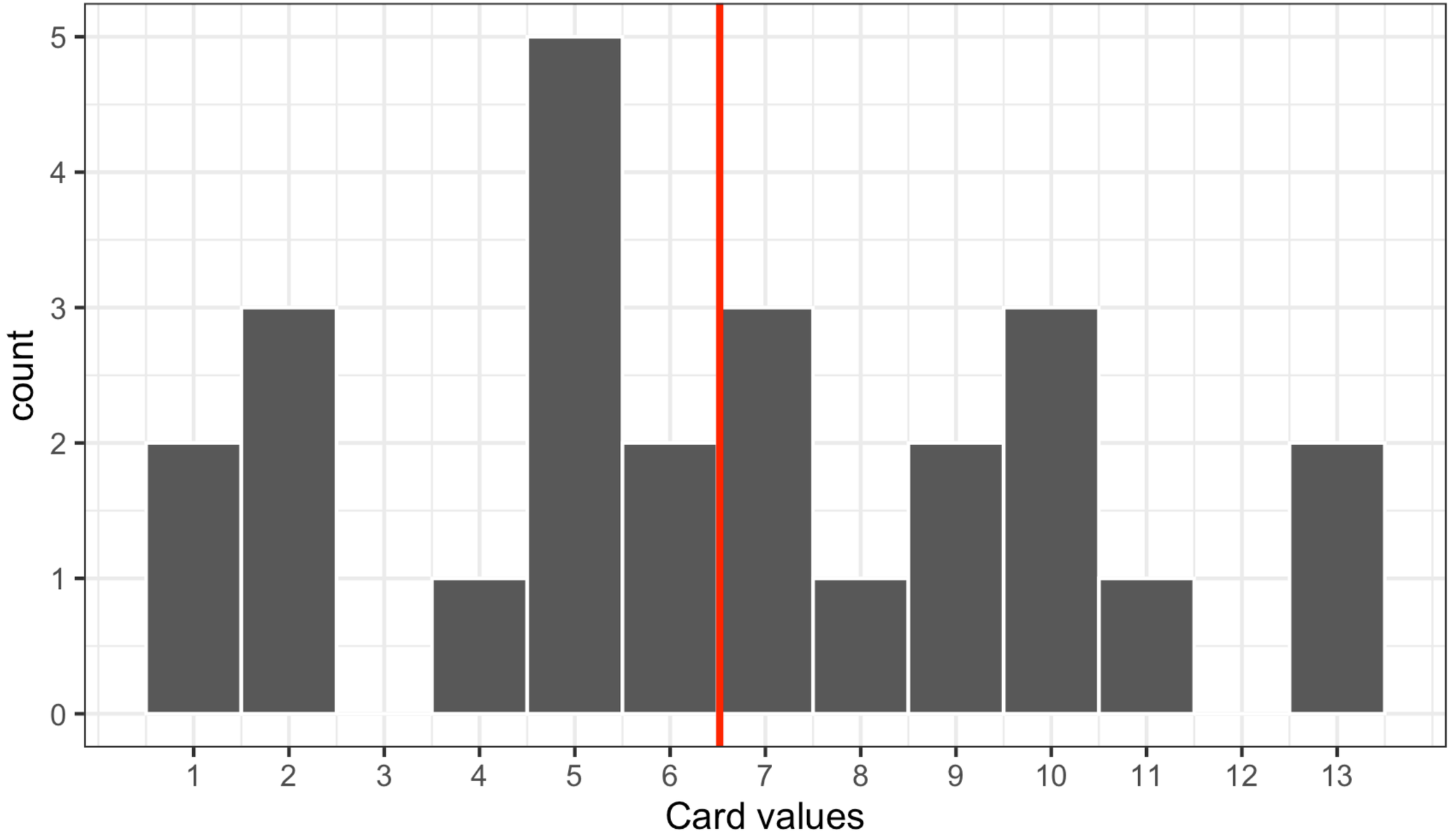
IF we could take lots of samples (we can't in the real world!)



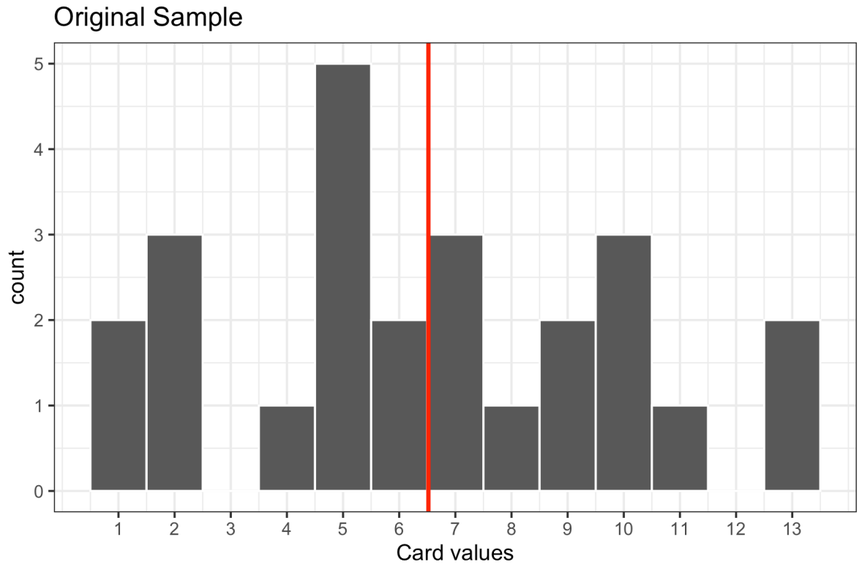
World's Largest Deck of Cards: Sample Distribution

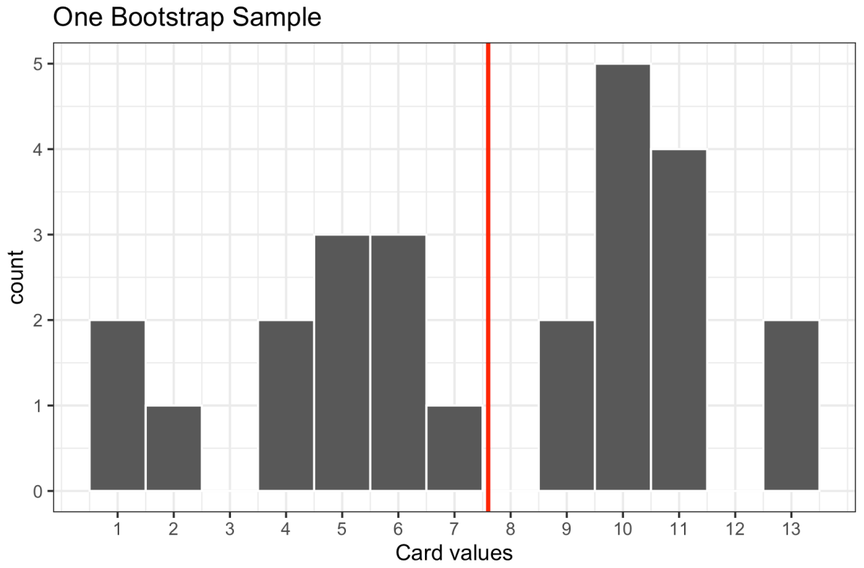
Sample's Distribution

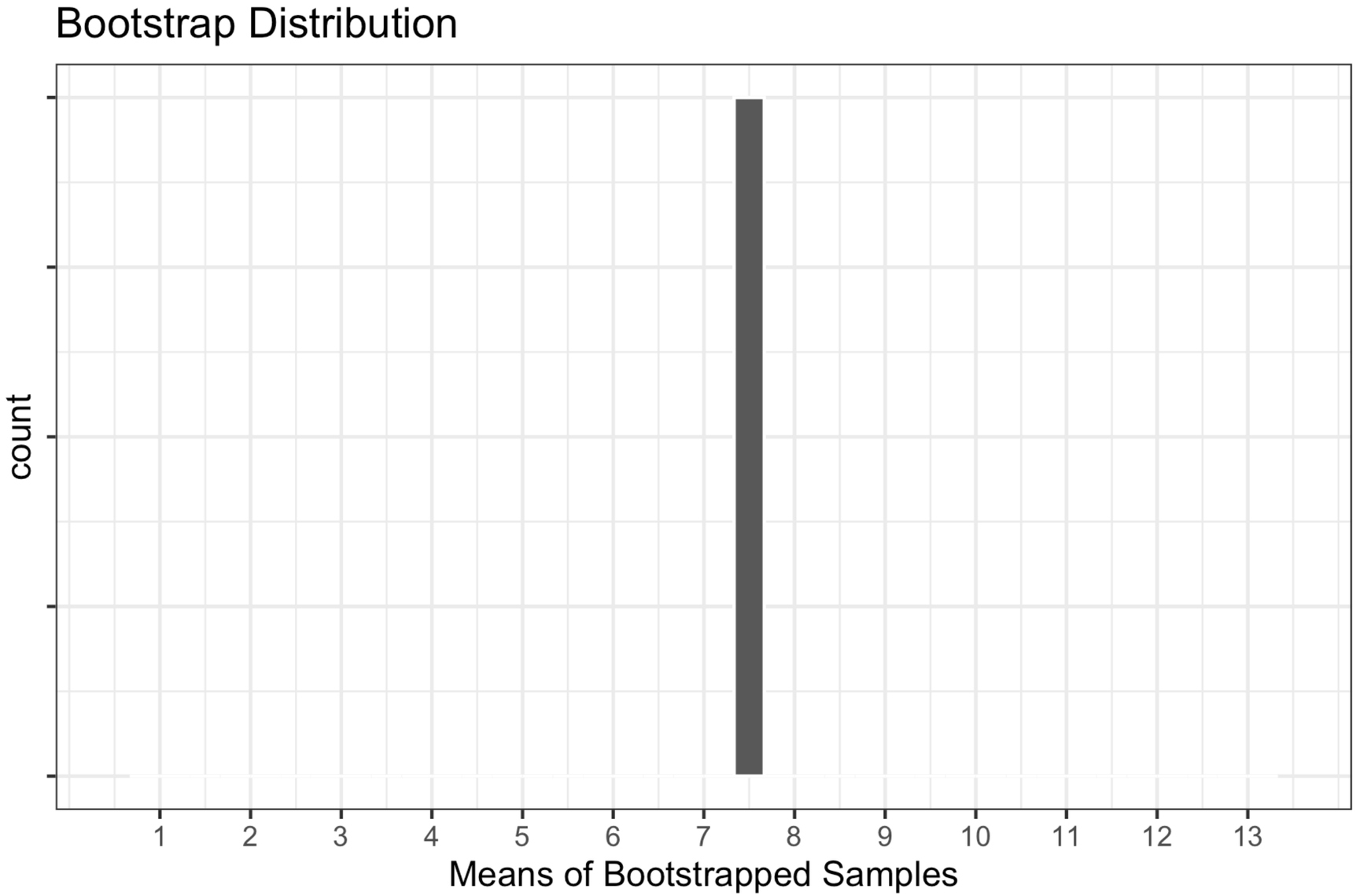
Values of cards in our sample of 25 cards



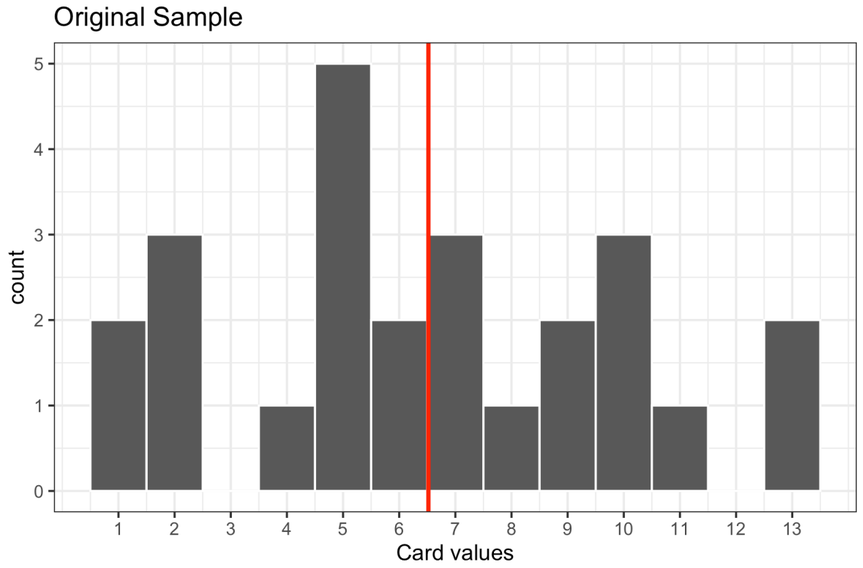
World's Largest Deck of Cards: Bootstrap Samples

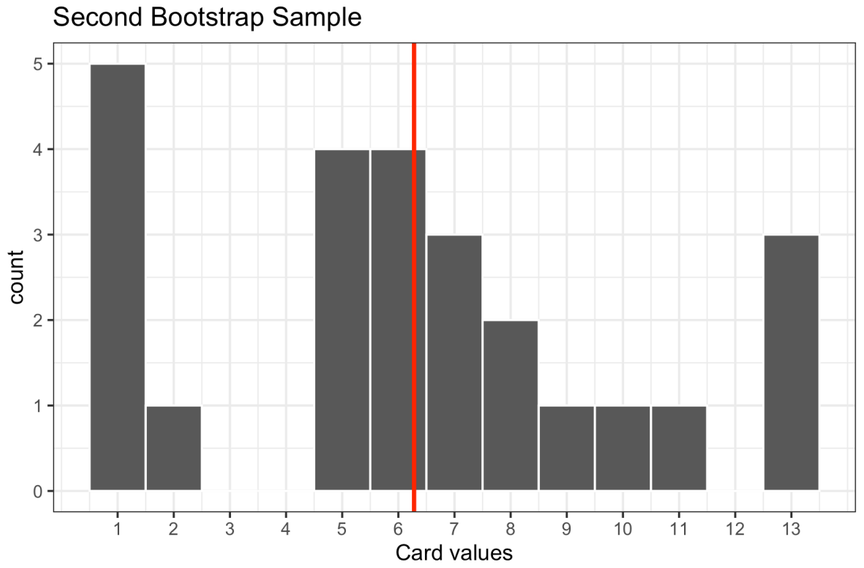


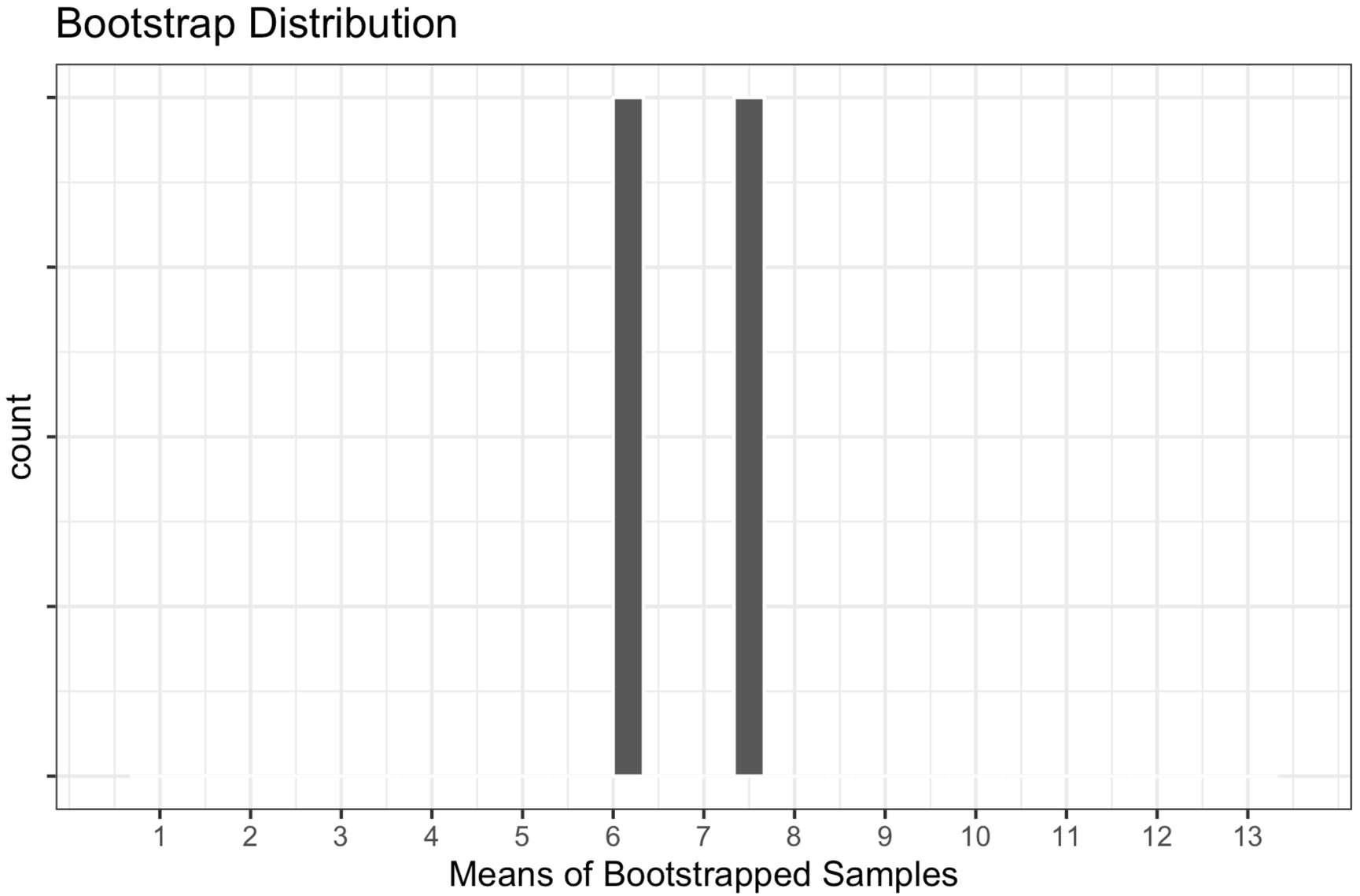



Bootstrap Distribution
count
1
2
3
4
5
6
7
8
9
10
11
12
13
Means of Bootstrapped Samples

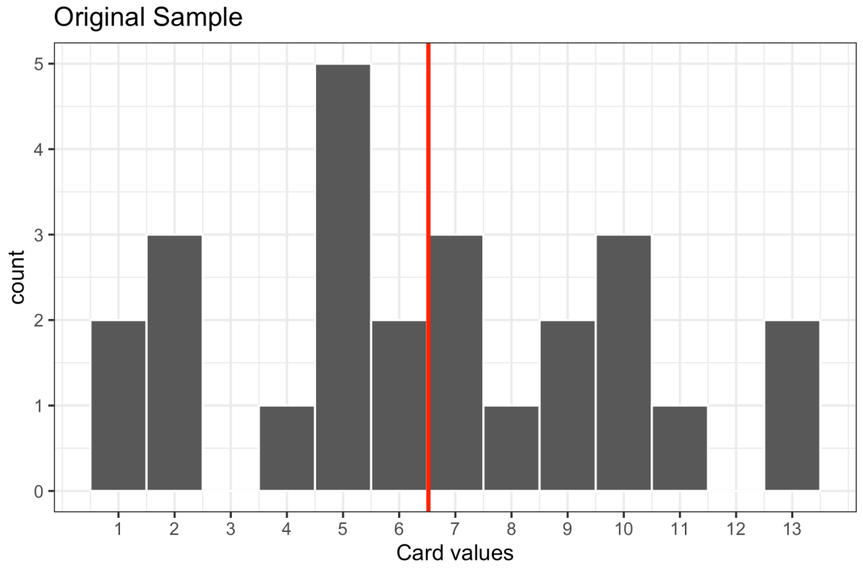
World's Largest Deck of Cards: Bootstrap Samples

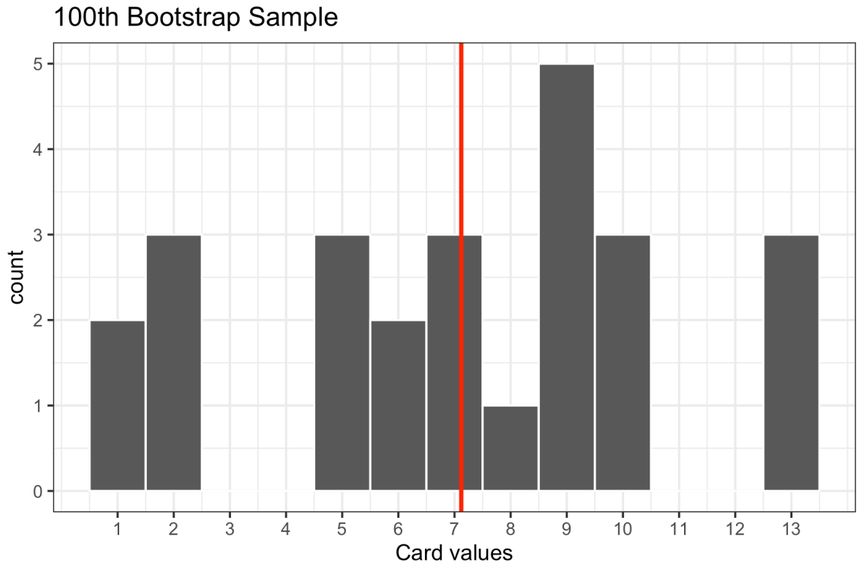


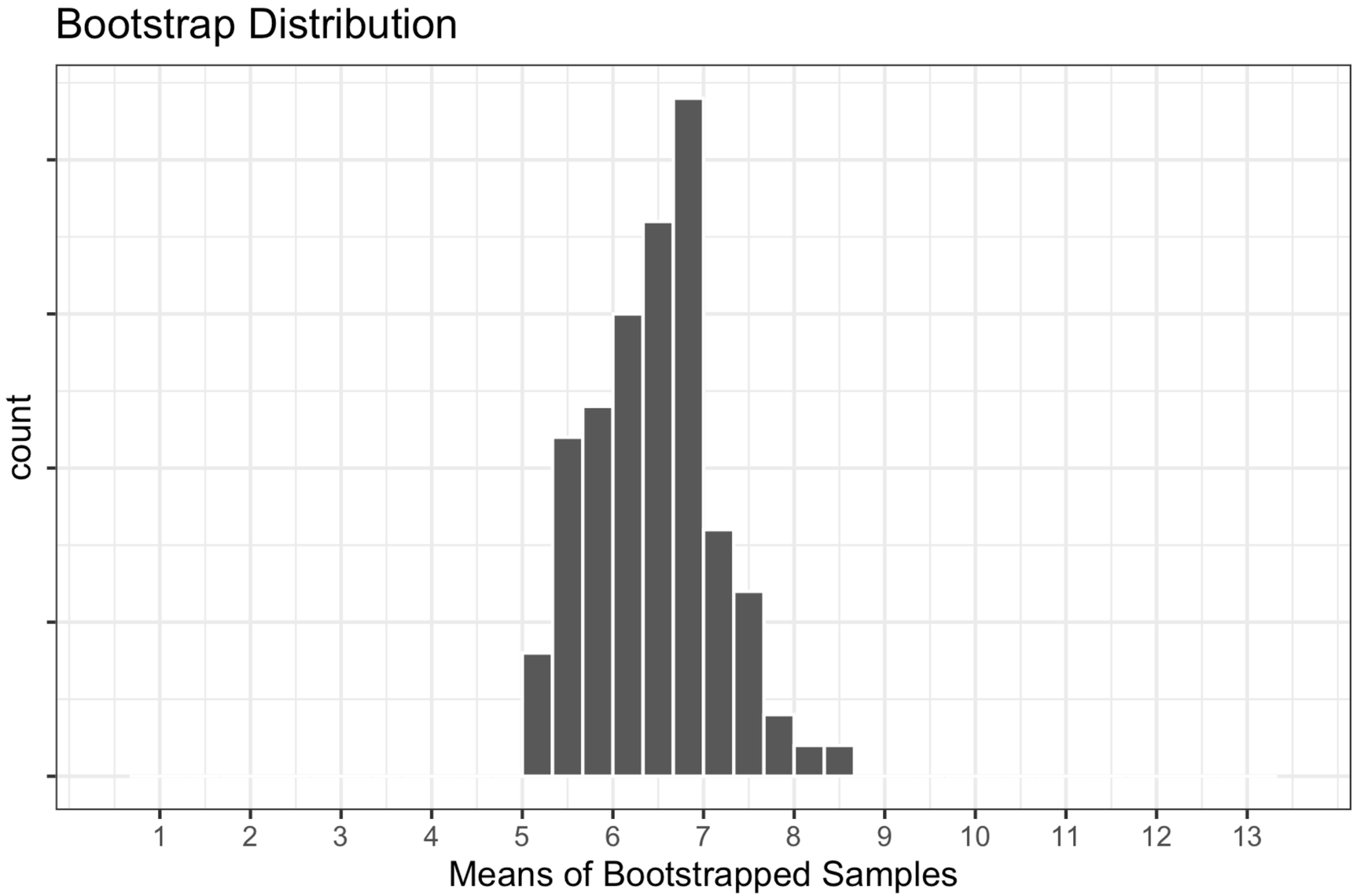



Bootstrap Distribution
count
1
2
3
4
5
6
7
8
9
10
11
12
13
Means of Bootstrapped Samples

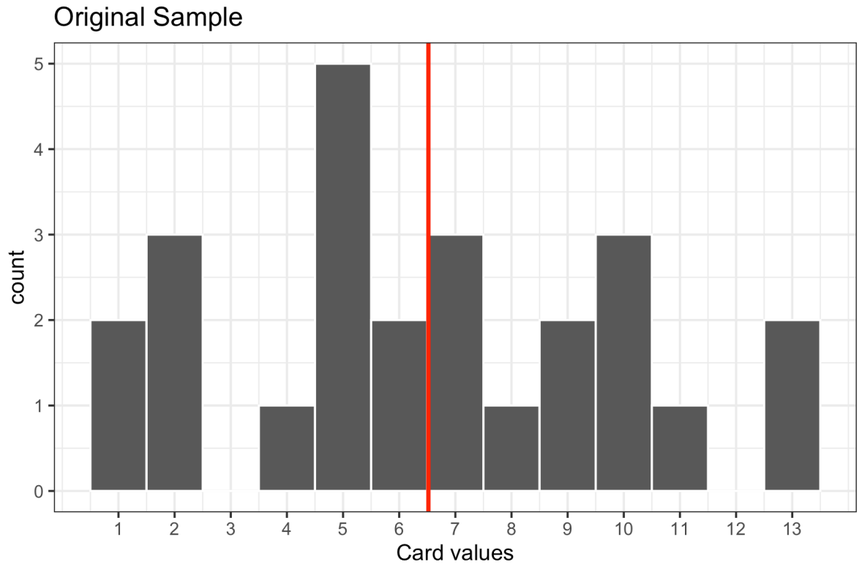
World's Largest Deck of Cards: Bootstrap Samples

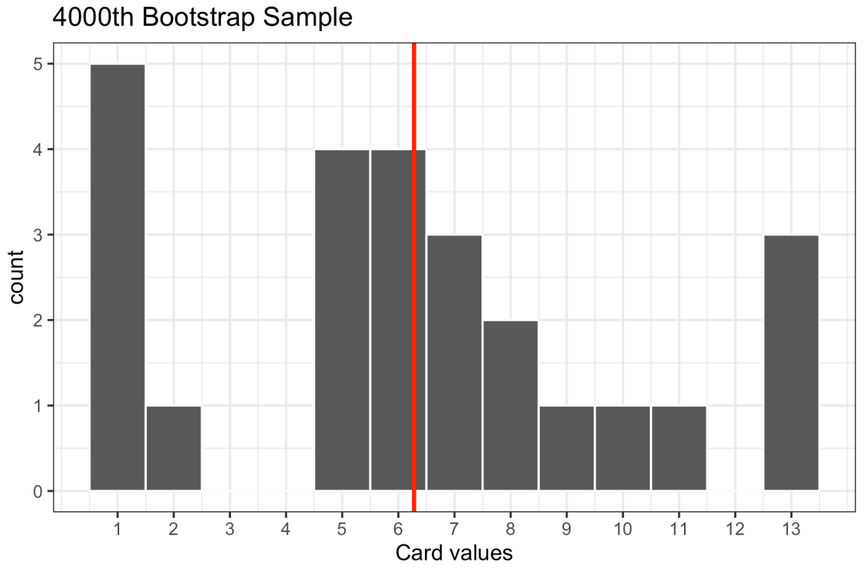


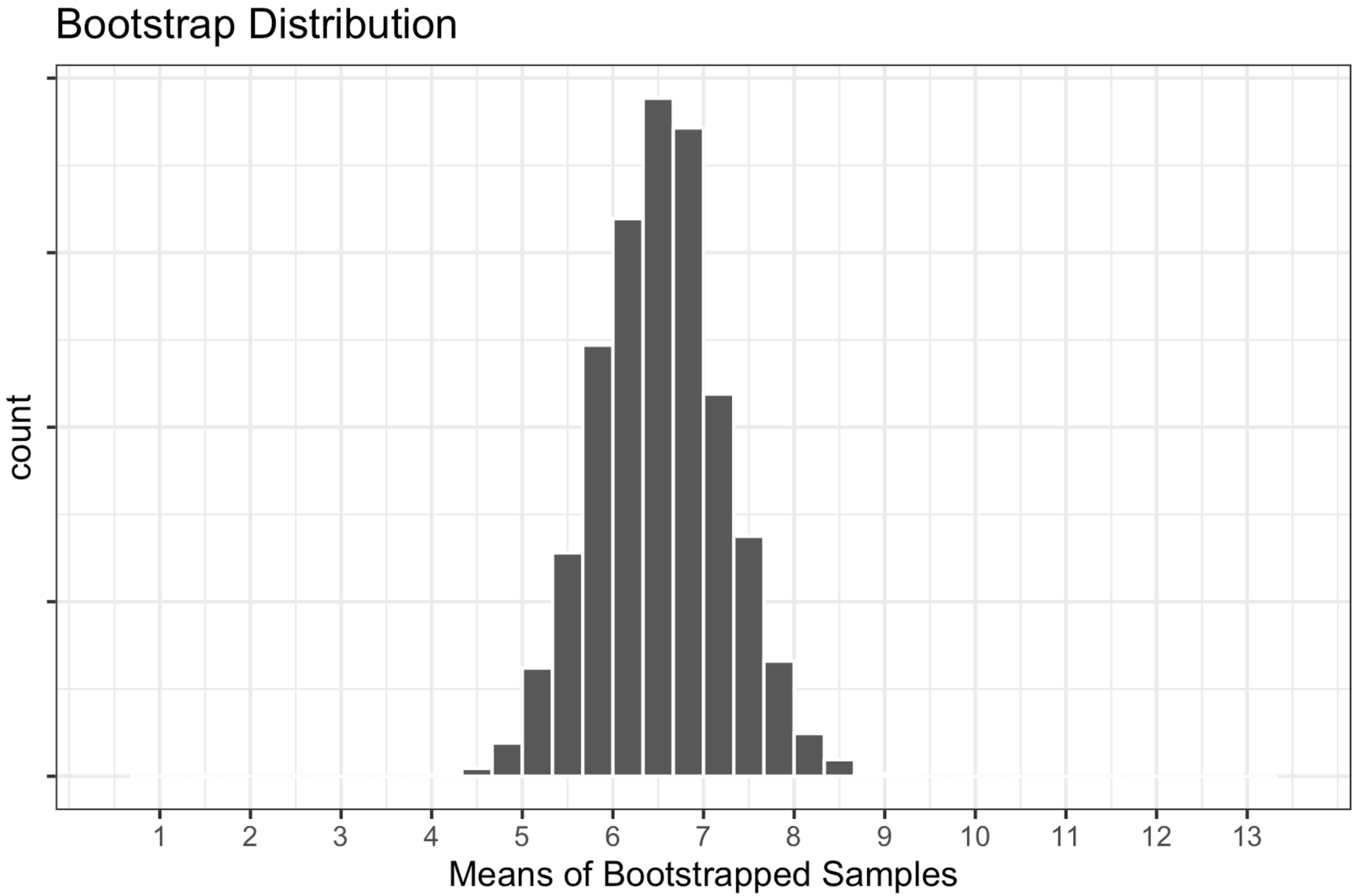



Bootstrap Distribution
count
1
2
3
4
5
6
7
8
9
10
11
12
13
Means of Bootstrapped Samples

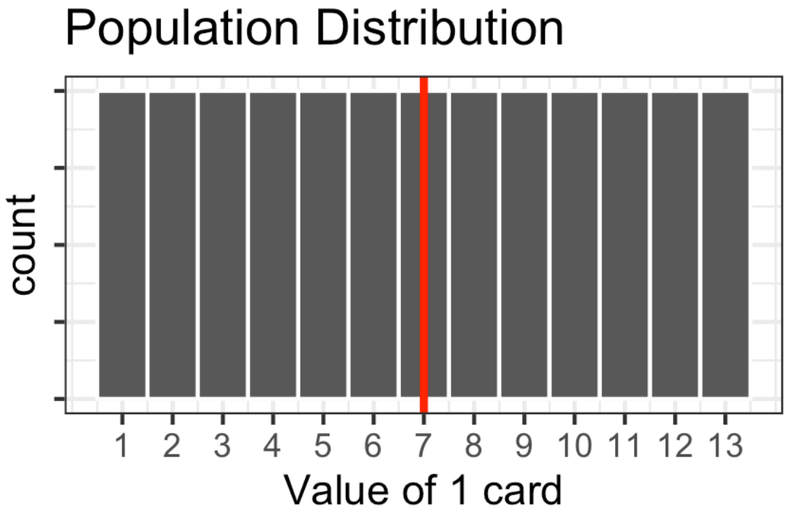
World's Largest Deck of Cards: Bootstrap Samples

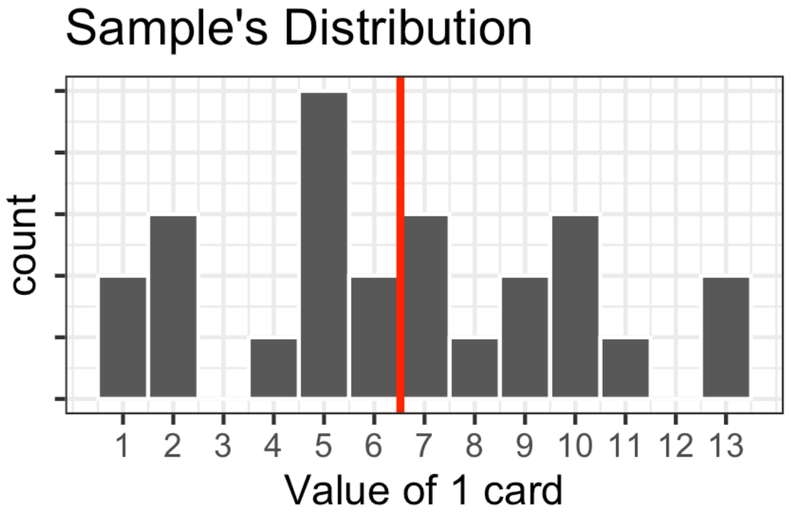


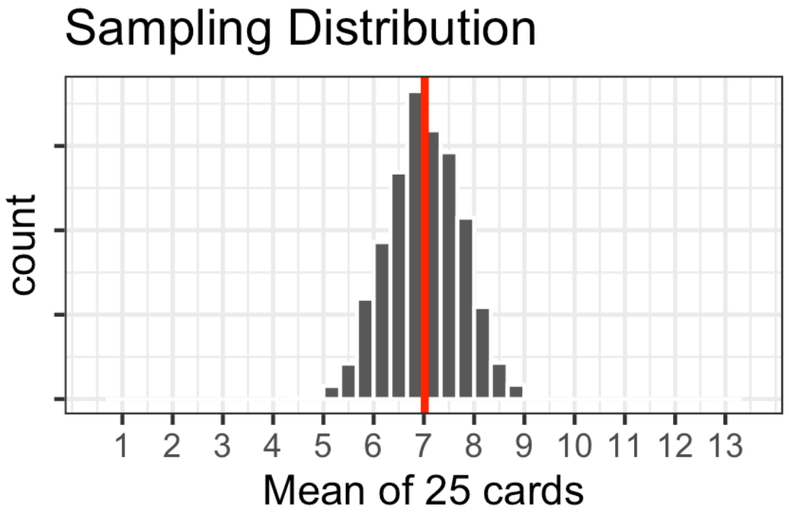


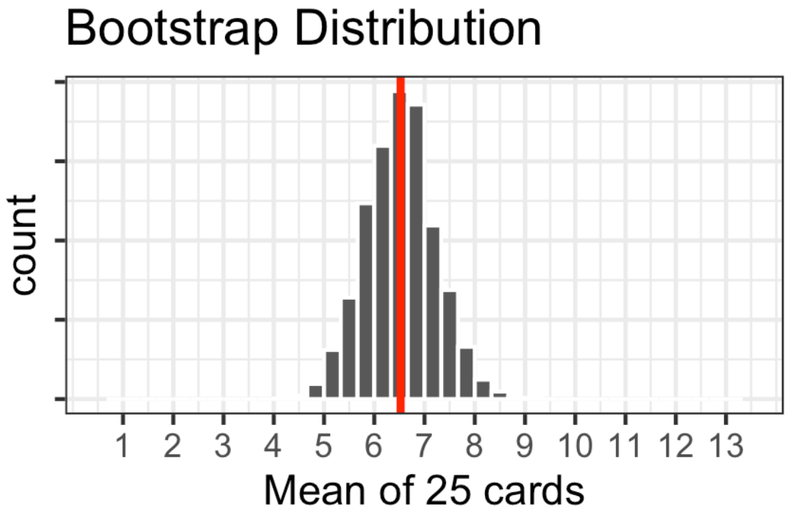

Bootstrap Distribution
count
1
2
3
4
5
6
7
8
9
10
11
12
13
Means of Bootstrapped Samples

World's Largest Deck of Cards: Recap









Q: Compare the Sampling Distribution and the Bootstrap Distribution. How are they similar? How do they differ?

World's Largest Deck of Cards

We can compute some relevant statistics:

Population:

mean_value	sd_value
7	3.742017

Sample:

mean_value	sd_value
6.52	3.513308

Sampling Distribution:

mean_xbar	sd_xbar
7.01703	0.7331422

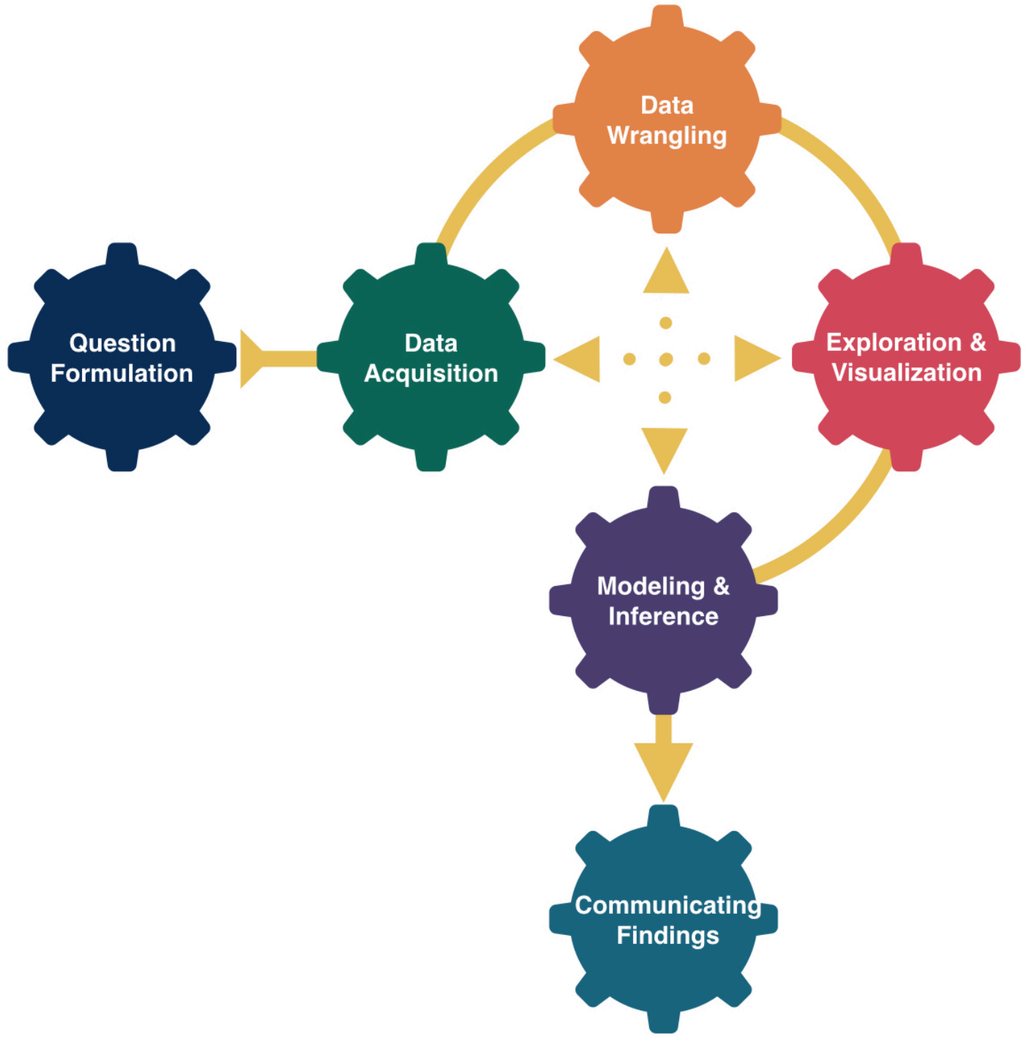
Bootstrap Distribution:

mean_xbar	sd_xbar
6.52426	0.6902801

- Sampling and Bootstrap have slightly different means, but have similar standard deviations!
- Mean of Sampling is the *true* mean
- Mean of Bootstrap is the *sample* mean

Recap

- Sampling distributions are useful for understanding how close a statistic (ex. $\hat{p}$) might be to a parameter (ex. p)
  - But we rarely know the sampling distribution. We usually only have one sample!
- Bootstrap samples allow us to **approximate the sampling distribution.**
  - Almost magically, we can use a single sample to generate many "bootstrap samples"
  - Bootstrap samples behave very much like regular samples
- Bootstrap distributions look very similar to sampling distributions,
  - …but without the work of taking many samples!
  - The center is shifted (around the statistic instead of the parameter), but, very importantly, similar shape and **spread**!
  - The **standard deviation of the bootstrap distribution** can help us estimate the **standard error of the sampling distribution** (next week)



Sampling and Bootstrap Distributions

Megan Ayers

Math 141 | Spring 2026

Friday, Week 6

ASA DataFest 2026

- Are you interested in data science and looking for a challenge? DataFest is an exciting opportunity to work with real-world data, collaborate with peers, and gain valuable experience!
- Friday April 17 - Sunday April 19 at Willamette
- To learn more and sign up, visit https://my.willamette.edu/site/computer-science/data-fest

Goals for Today

- Complete a worksheet comparing sampling and bootstrap distributions
- Review our answers together

Activity

Instructions

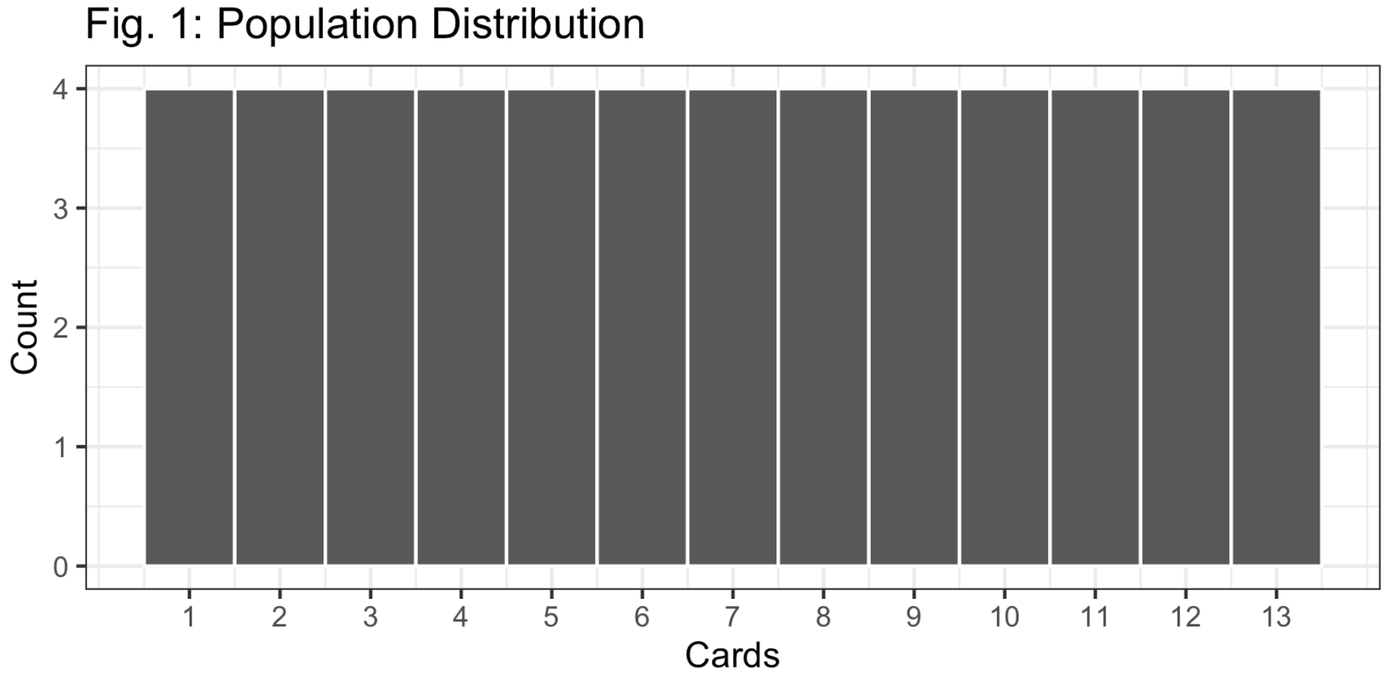
- In small groups, carefully complete the worksheet

- When you are finished, **submit on Gradescope for a small completion grade**
  - You may work on HW 5 or midterm review if you have extra time!

- We'll come together to discuss with 10-15 minutes remaining

Question 1

```
1 deck <- data.frame(cards = rep(1:13, each = 4))
```



Q1: In Figure 1, why is the height of each bar 4? Describe this distribution.

- There are four cards of each value in a deck.
- The distribution has a “flat” shape, centered at 7 (mean = 7 and median = 7).
- The distribution has a “large” spread. Standard Deviation: 3.78

Question 2

```
1 set.seed(1) # ensures we all get the same "random" sample
2 single_sample <- deck %>% rep_sample_n(size = 10, replace = FALSE, reps = 1)
3 single_sample$cards
```

```
[1]  1 10  1  9  6 11  4  5  9  6
```

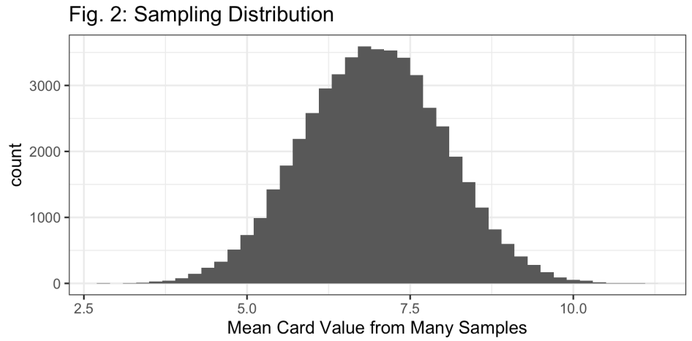
Q2: Based on the code, which cards are in our sample? Use the cards to calculate our sample statistic (the sample mean) based on this sample.

- We have two Ace's (1's), a 4, 5, two 6's, two 9's, a 10, and a Jack (11).
- The sample mean is 6.2.

Question 3

```
deck %>%
  rep_sample_n(size = 10,
               replace = FALSE, reps = 50000) %>%
  group_by(replicate) %>%
  summarize(x_bar = mean(cards)) %>%
  ggplot(aes(x = x_bar)) +
  geom_histogram(binwidth = 0.2) + theme_bw() +
  labs(x = "Mean Card Value from Many Samples",
       title = "Fig. 2: Sampling Distribution")
```

Q3: Based on the code above, how many cards are in each sample? How many different samples did we take to create the sampling distribution in Figure 2?

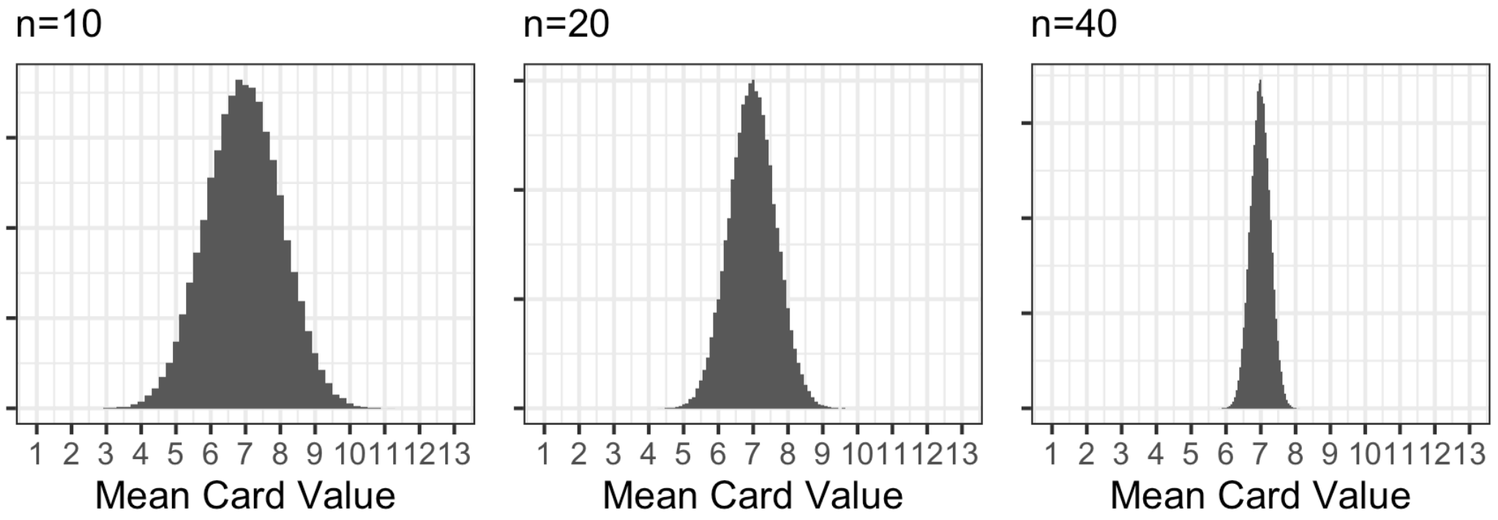


- There are 10 cards in each sample, since `size = 10`.
- We took 50,000 different samples, since `reps = 50000`.

Question 4

Q4: Figure 3 (below) displays sampling distributions for samples of size n=10, n=20, and n=40. How are they similar and how are they different? Why do larger samples have sampling distributions with less variability?

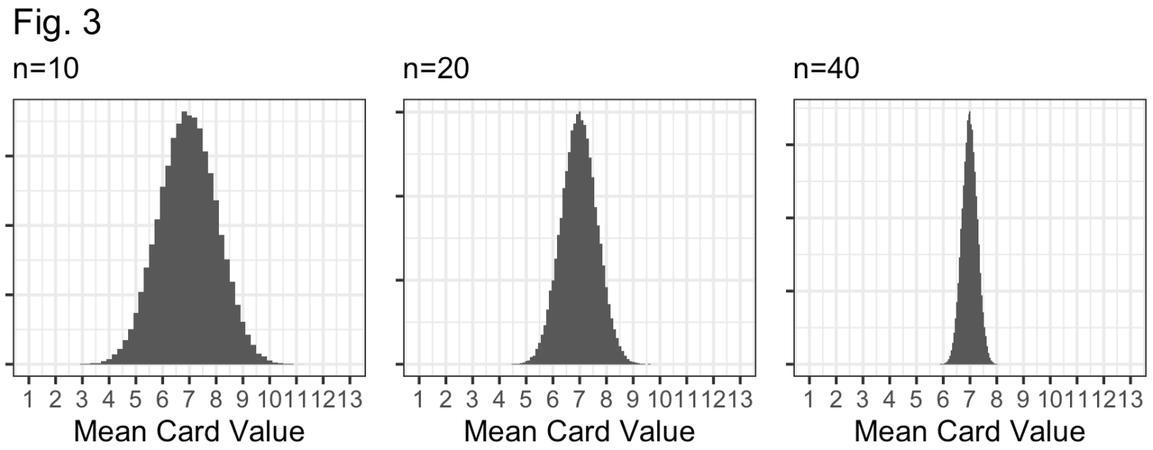
Fig. 3



- Bell-shaped and centered at the true mean, 7.
- They differ in their spread: $n = 10$ distribution has the largest standard error; the $n = 40$ distribution has the smallest standard error.

Question 4

Q4: Figure 3 (below) displays sampling distributions for samples of size n=10, n=20, and n=40. How are they similar and how are they different? Why do larger samples have sampling distributions with less variability?



- **Why?** If our sample size (n) is larger, we have more data and our sample should be "more representative" of the population.
- i.e., more data means a better glimpse at the true population, and better guesses about the population mean!

Questions 5 and 6

```
1 bootstrap_sample <- single_sample %>% rep_sample_n(size = 10, replace = TRUE, reps = 1)
2 bootstrap_sample$cards
```

```
[1] 11  6  9  1  5  6  1  1  9  9
```

Q5: In the code above, what are we sampling from (the population, or the single sample)? Are we sampling with replacement? What's our sample size?

- We're sampling from the single sample (`single_sample`).
- We're sampling with replacement (`replace = TRUE`).
- Our sample size is still 10 (`size = 10`)

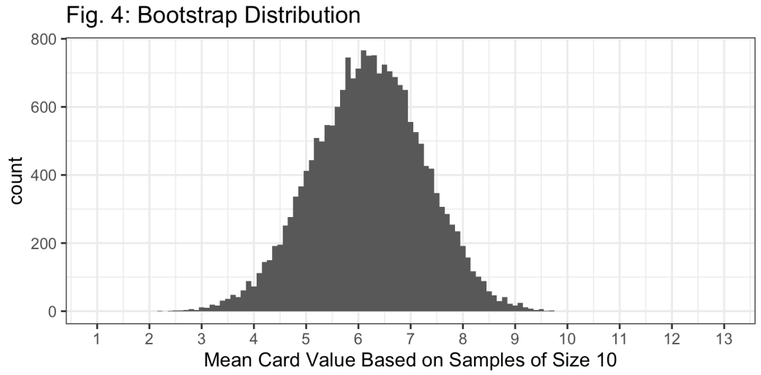
Q6: How would your answers to Q5 be different if we were talking about sampling for a sampling distribution?

- We sample from the **population**.
- We sample **without** replacement.
- Our sample size is **still 10**!

Question 7

```
single_sample %>% ungroup() %>% select(cards) %>%
  rep_sample_n(size = 10, replace = TRUE, reps = 20000) %>%
  group_by(replicate) %>%
  summarize(x_bar = mean(cards)) %>%
  ggplot(aes(x = x_bar)) + geom_histogram(binwidth = 0.1) +
  labs(x = "Mean Card Value Based on Samples of Size 10",
       title = "Fig. 4: Bootstrap Distribution") +
  theme_bw() +
  scale_x_continuous(breaks = 1:13, limits = c(1, 13))
```

Q7: Based on the code above, how many bootstrap samples are we taking to create the bootstrap distribution?



- 20,000 bootstrap samples because `reps=20000`.

Question 8

Fig. 3

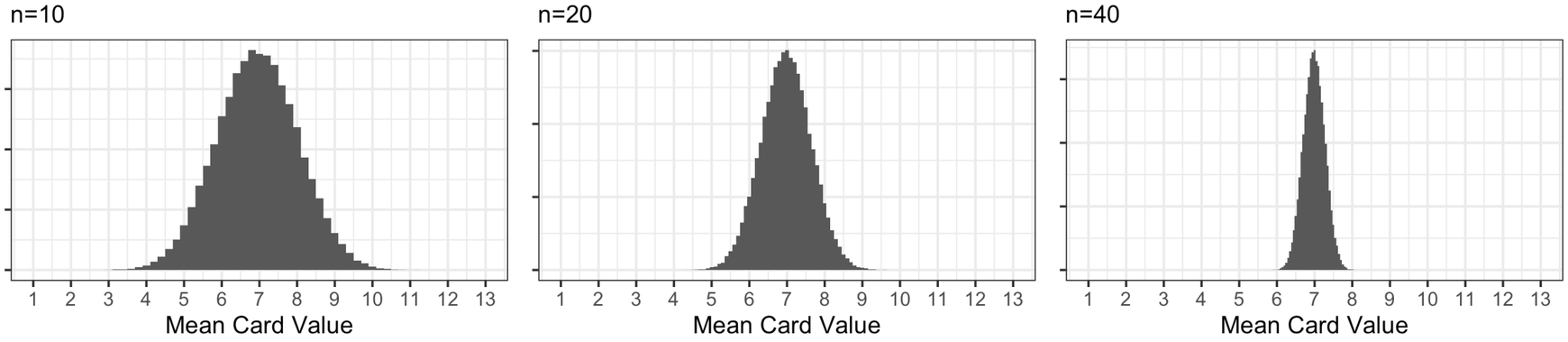
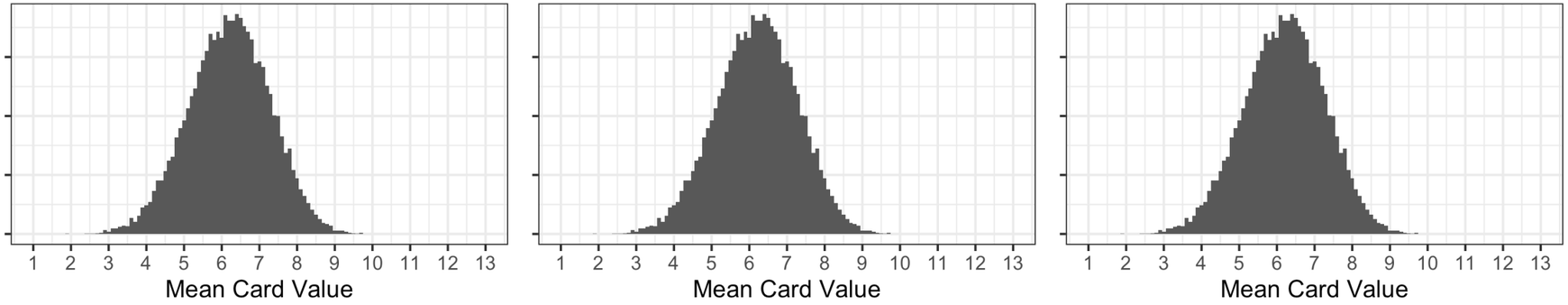


Fig. 4: Bootstrap



Q8: Which sampling distribution looks most like the bootstrap distribution in Figure 4? How specifically is it similar or different? Consider the shape, center, and spread of each distibution.

Question 8

Fig. 3

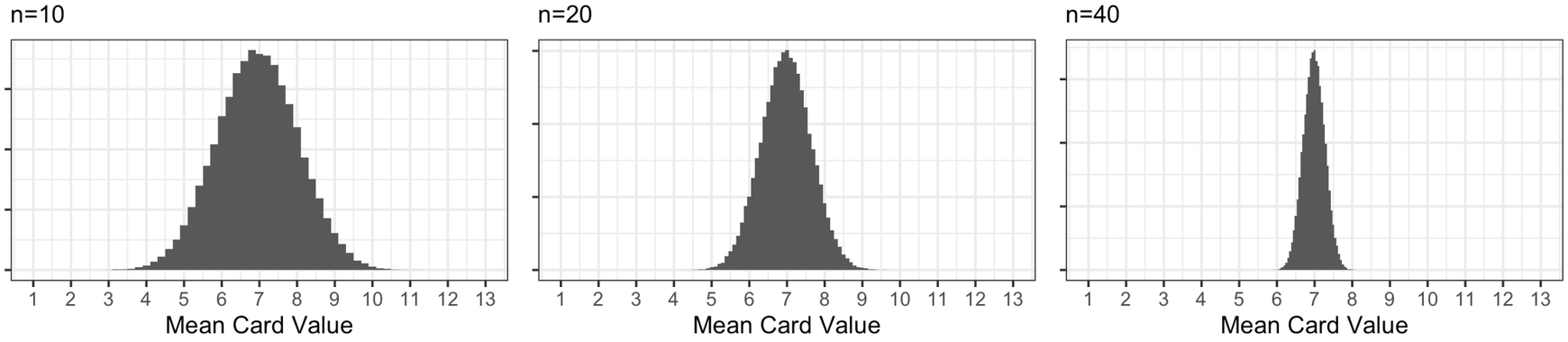
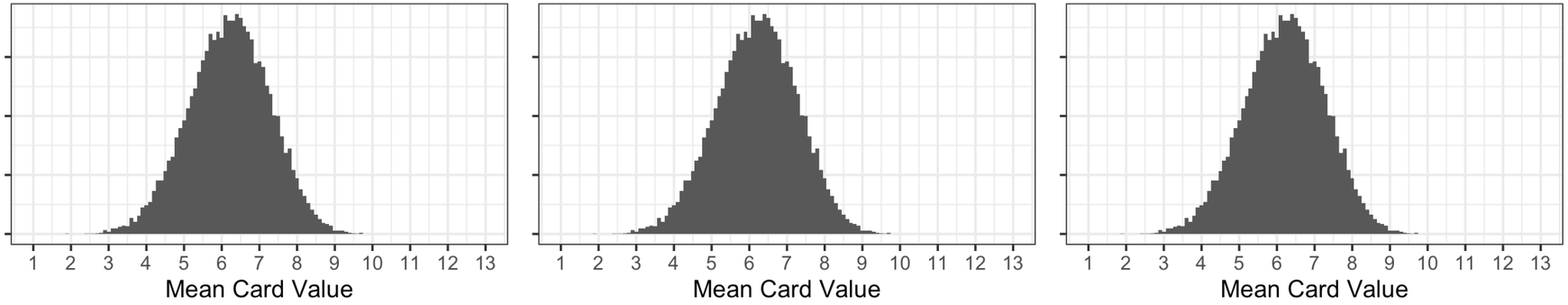


Fig. 4: Bootstrap



The $n = 10$ sampling distribution!

Question 8

Fig. 3

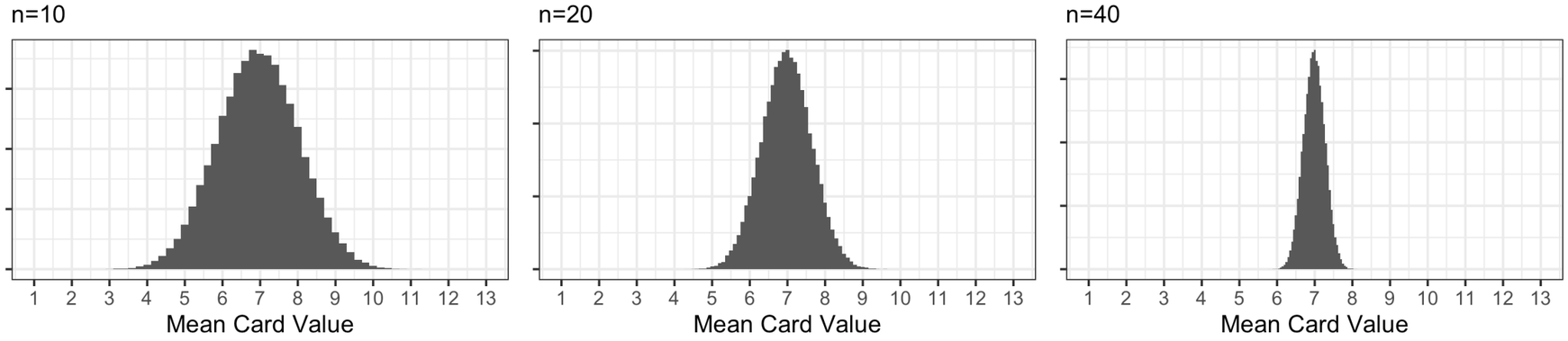
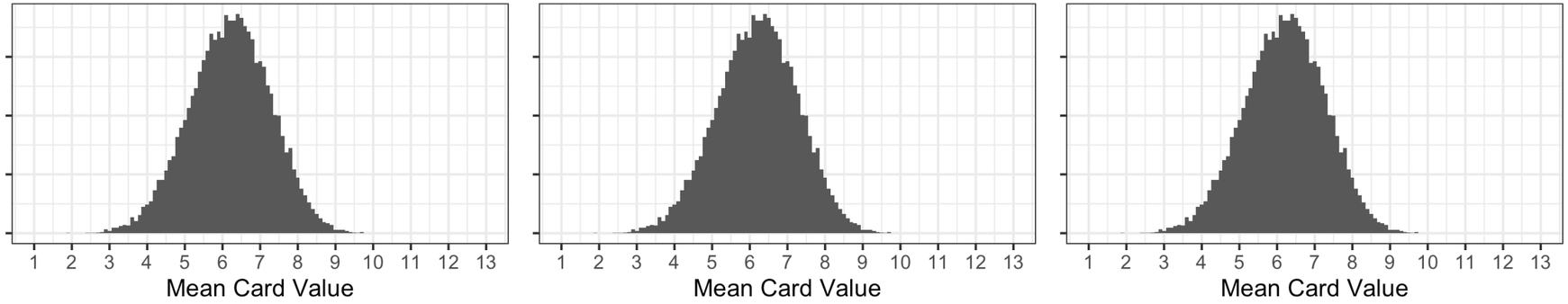


Fig. 4: Bootstrap



Shape: All distributions look bell-shaped, so this doesn't distinguish them at all.

Question 8

Fig. 3

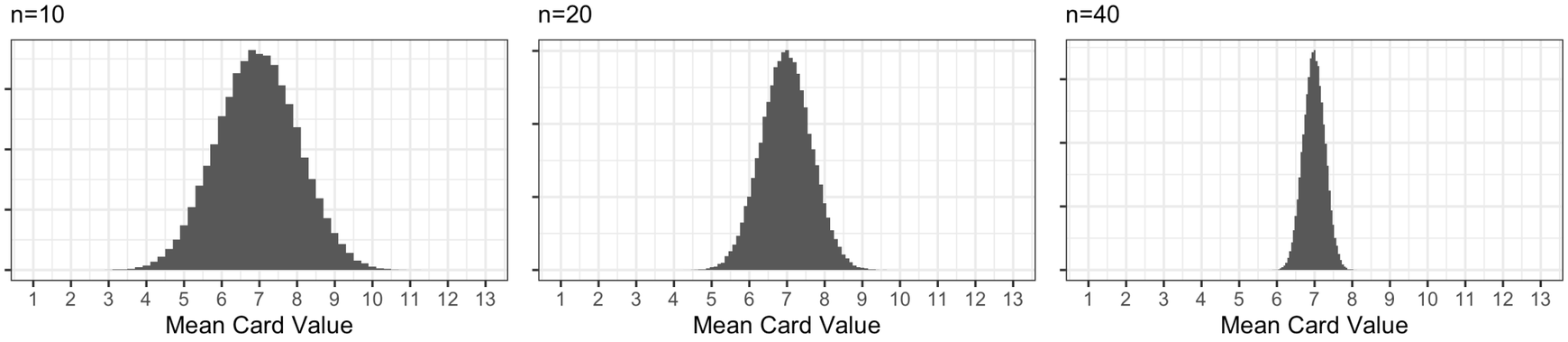
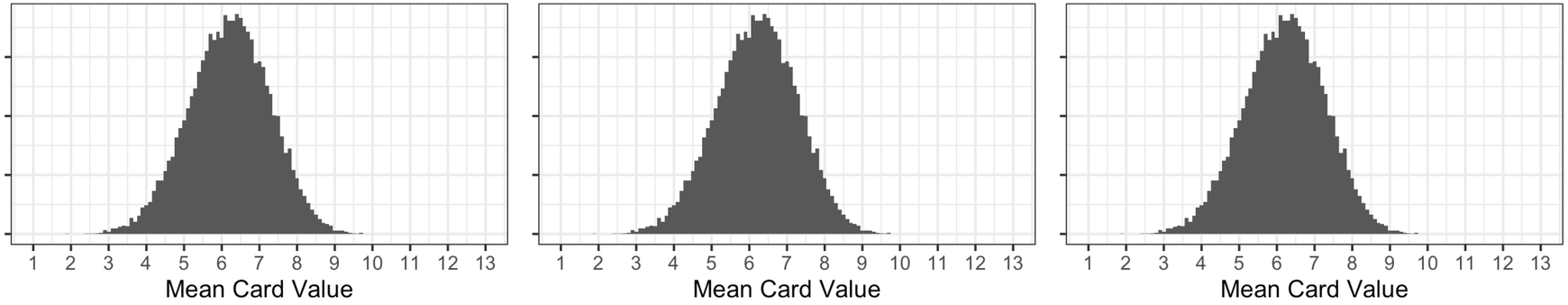


Fig. 4: Bootstrap



Center: Bootstrap is centered at the sample mean (6.2); all sampling distributions are centered at the population mean (7).

Question 8

Fig. 3

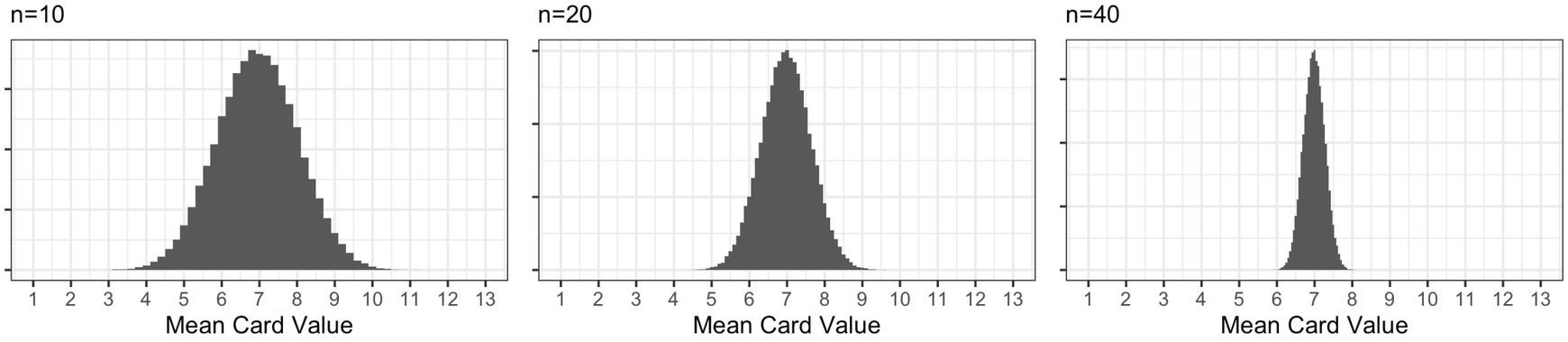
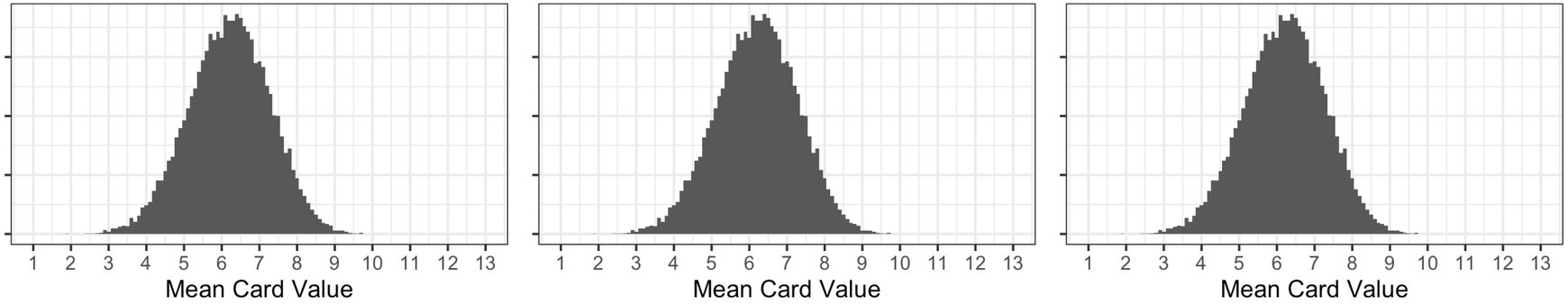


Fig. 4: Bootstrap

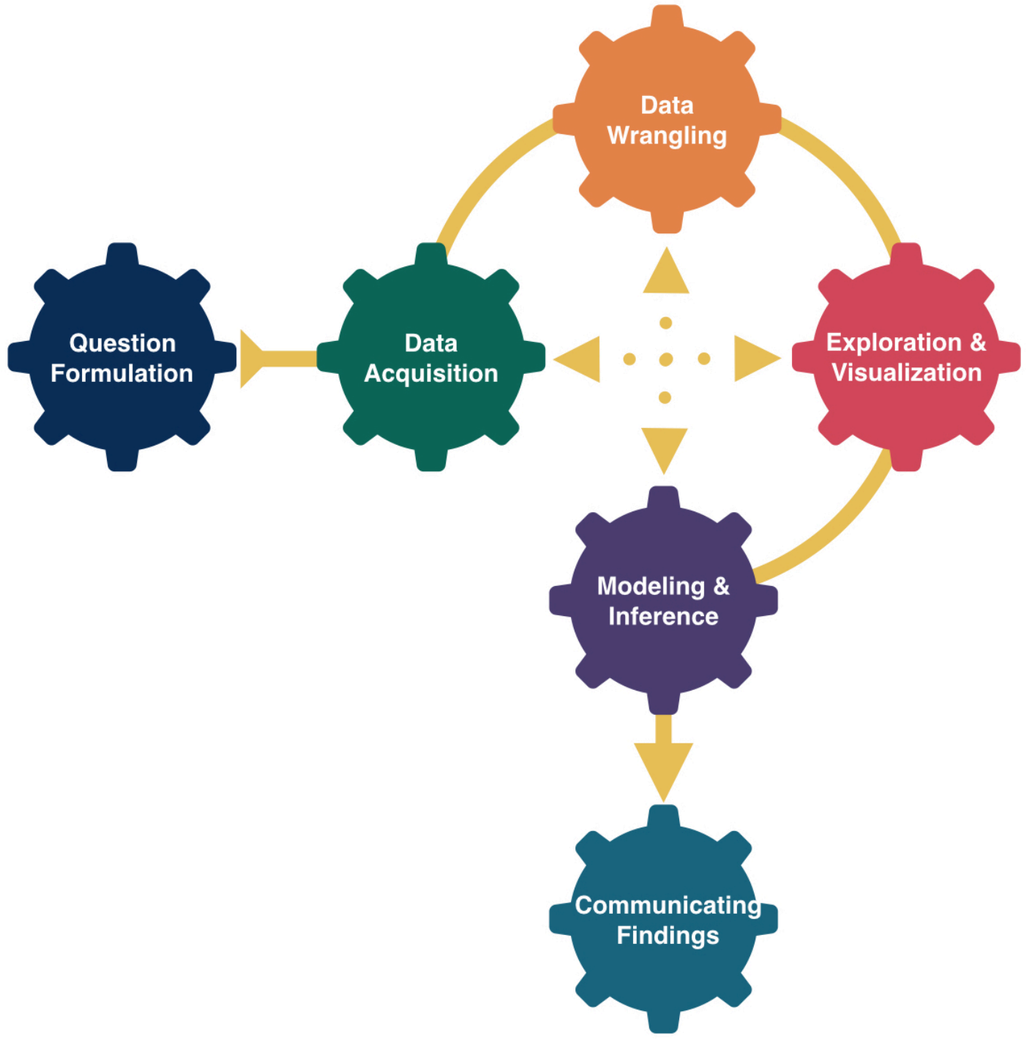


Spread: The spread of the bootstrap distribution looks similar to the spread of the sampling distribution with $n = 10$.

Question 9

Q9: Thinking more generally, compare and contrast sampling distributions and bootstrap distributions.

- Sampling distributions and bootstrap distributions both help us conceptualize the distribution of a statistic, for the purpose of understanding plausible values for a parameter.
- Sampling distributions and bootstrap distributions should have similar spread (e.g., a similar standard deviation).
- Their means should also be similar, although a sampling distribution is centered at the true parameter, while a bootstrap distribution is centered at the sample mean.
- Sampling distributions are usually not possible to obtain (we can only take one sample from the population). Bootstrap distributions approximate a sampling distribution, and are super easy to obtain (especially with code).



Confidence Intervals I

Megan Ayers

Math 141 | Spring 2026

Monday, Week 7

Announcements/reminders

- Homework 5 due tonight
- Homework 6 is posted, but not due until 3/20 (last day before spring break)
  - Today's content relevant to starting Ex. 4, 5.

Today's Goals

- Introduce confidence intervals as a method for estimating a parameter
- Use bootstrapping as a means of creating confidence intervals

Confidence Intervals

Point Estimates

- To estimate a population parameter, we can use a sample statistic.
  - **Ex:** You're hosting a pizza party for 200 people, and need to know what proportion p of vegetarian pizza to order.
  - **Idea:** Ask a (random) sample of attendees about their pizza preference and use the proportion $\hat{p}$ in the sample as an estimate for the total proportion p.
- Your sample statistic, $\hat{p}$, is a good guess for the population parameter.
  - **Terminology:** We sometimes call our sample statistic a point estimate.

Point Estimates vs. Interval Estimates

- A sample statistic is a good guess for the population parameter, but not the whole story
  - **Q:** After polling pizza party attendees, you find $\hat{p} = 0.33$. What factor(s) determine your (un)certainty in this estimate?
- It may be preferable to estimate the proportion using **a range of values**, with smaller intervals corresponding to precision.
  - With just $n = 9$ people, you might give a range of 0.03 **to** 0.63 for p.
  - But with $n = 48$, you might instead give the range 0.20 **to** 0.46.
- We call these ranges **interval estimates**.

Confidence Interval Estimates

A **confidence interval estimate** for a parameter *usually* takes the form

$$\text{Statistic} \pm \text{Margin of Error (ME)}$$

The confidence interval gives a range of plausible values for the parameter.

- e.g., when sampling pizza preferences with $n = 48$, we estimate p using the interval

$$0.20 \text{ to } 0.46 \qquad \text{or} \qquad \underbrace{0.33}_{\text{Statistic } (\hat{p})} \pm \underbrace{0.13}_{\text{ME}}$$

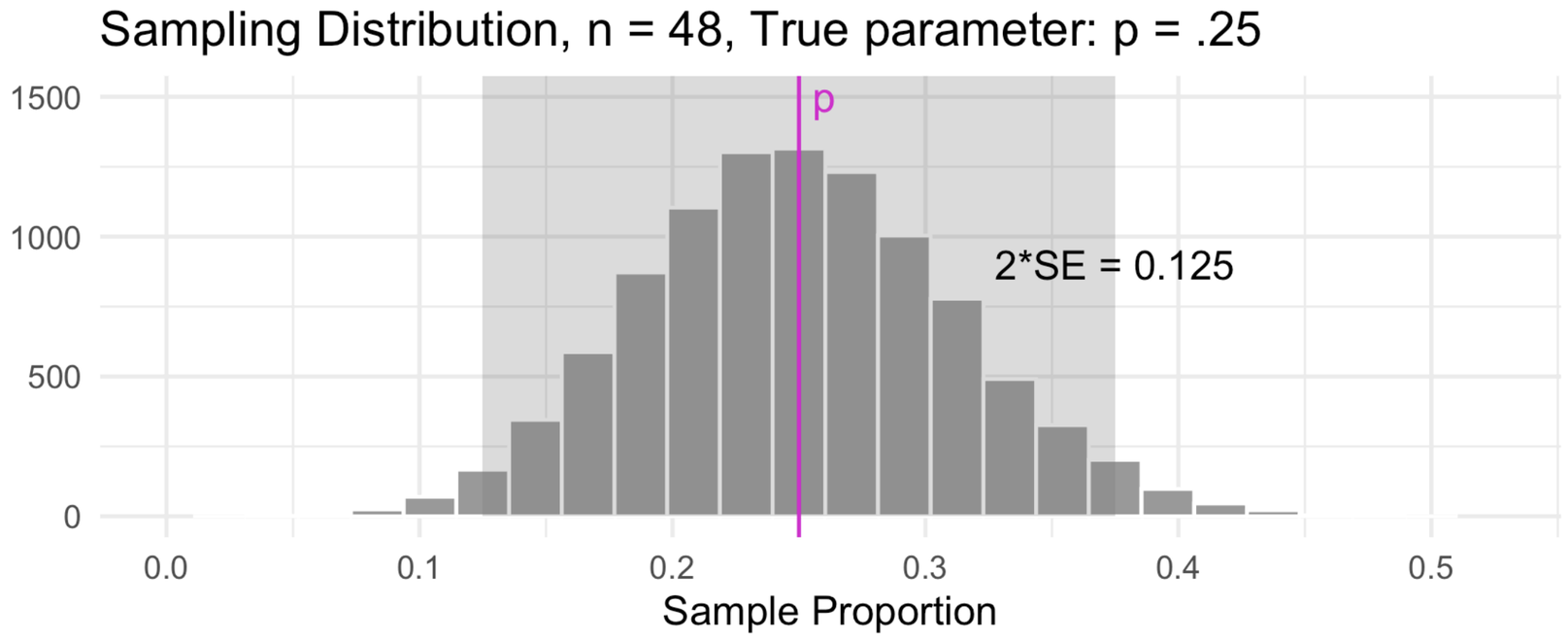
The **Margin of Error** determines the width of the interval ($2 * \text{Margin of Error}$)

- e.g., in our pizza interval, the width is:

$$0.46 - 0.20 = 0.26 = 2 * \underbrace{0.13}_{\text{ME}}$$

Confidence Intervals using the Sampling Distribution

- **Goal**: Figure out a reasonable **Margin of Error**
- In our example, suppose $p = 0.25$ (i.e., 25% of all party attendees prefer vegetarian)
- For approximately bell-shaped sampling distributions, **95% of all sample statistics are within 2 SE of the parameter.**



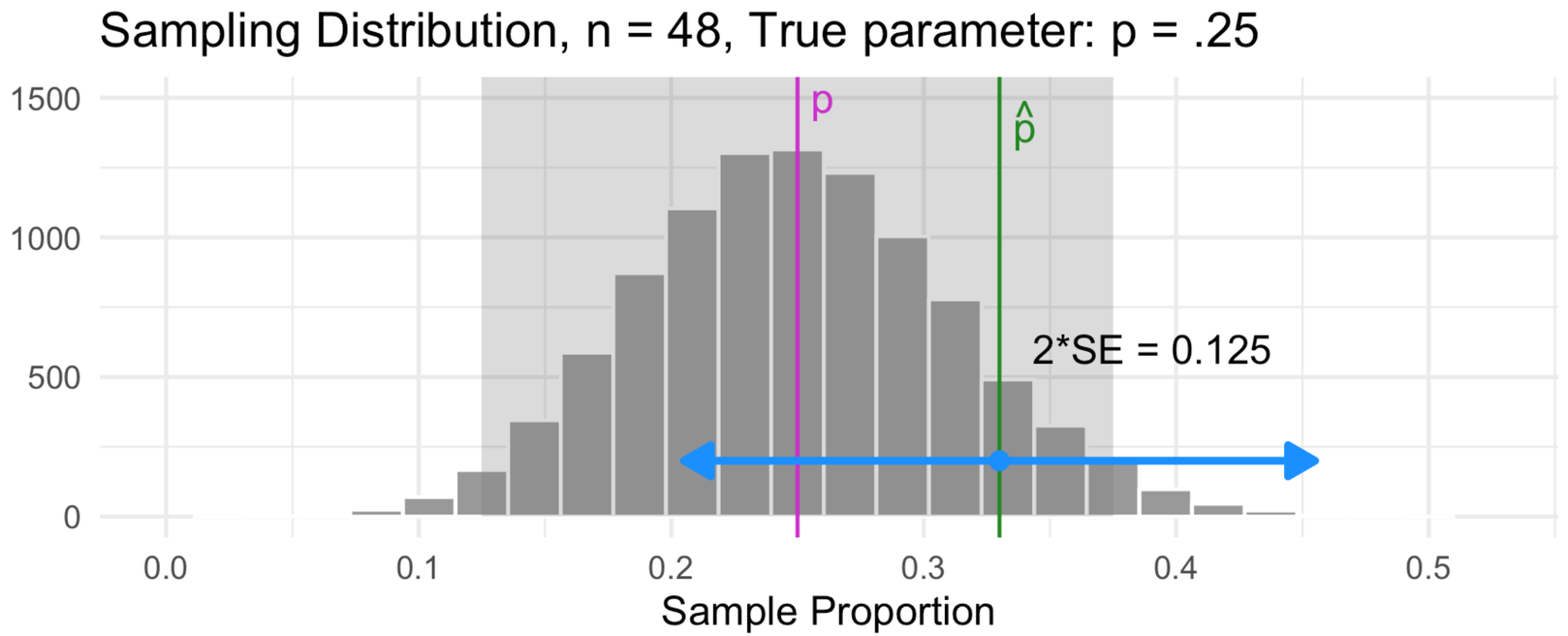
- This also means that **for 95% of all samples, the parameter will be within a distance of 2 SE of the sample statistic** (every sample in the gray region)

Confidence Intervals

Idea: Build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \text{SE}$$

- This interval will contain the parameter p in 95% of samples!



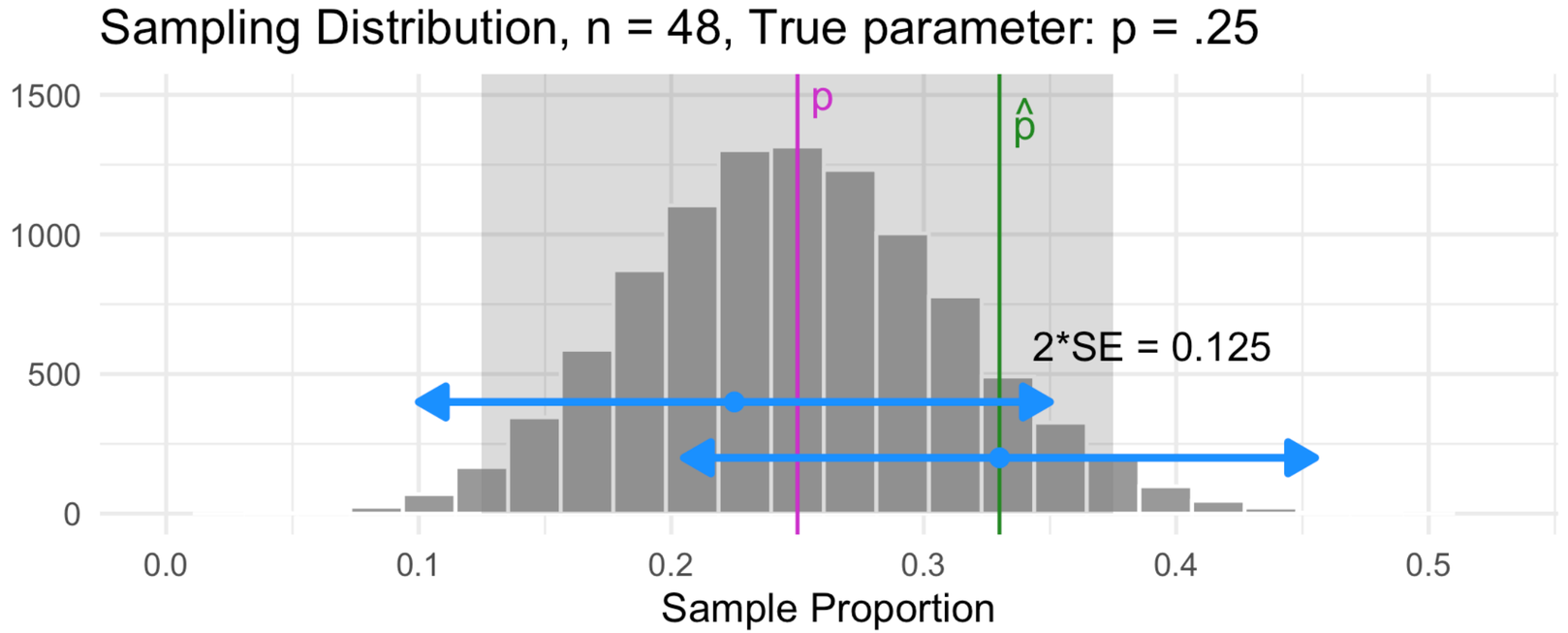
- The interval for our sample was 0.33 ± 0.125, which *does* contain the parameter p

Interval Estimates

Idea: Build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \text{SE}$$

- This interval will contain the parameter p in 95% of samples!



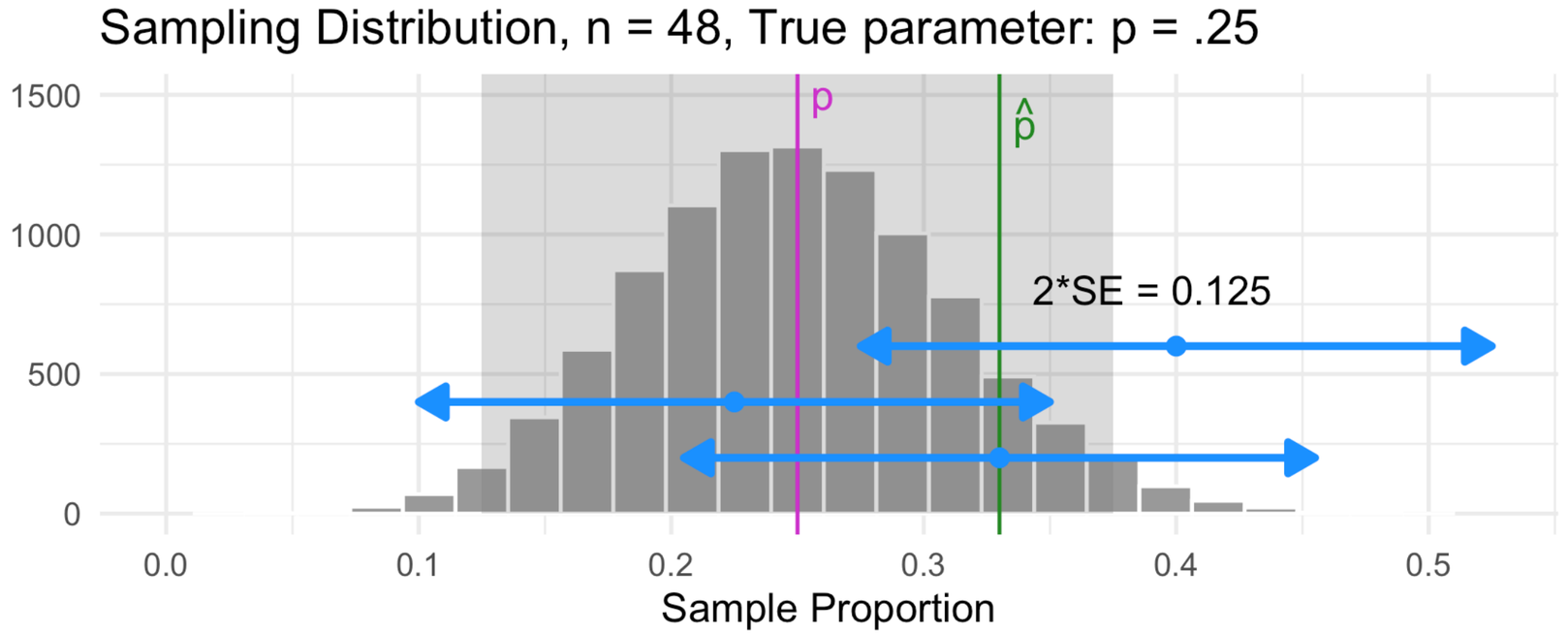
- Samples with $\hat{p}$ in the gray region have intervals that also contain the parameter p

Interval Estimates

Idea: Build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \text{SE}$$

- This interval will contain the parameter p in 95% of samples!



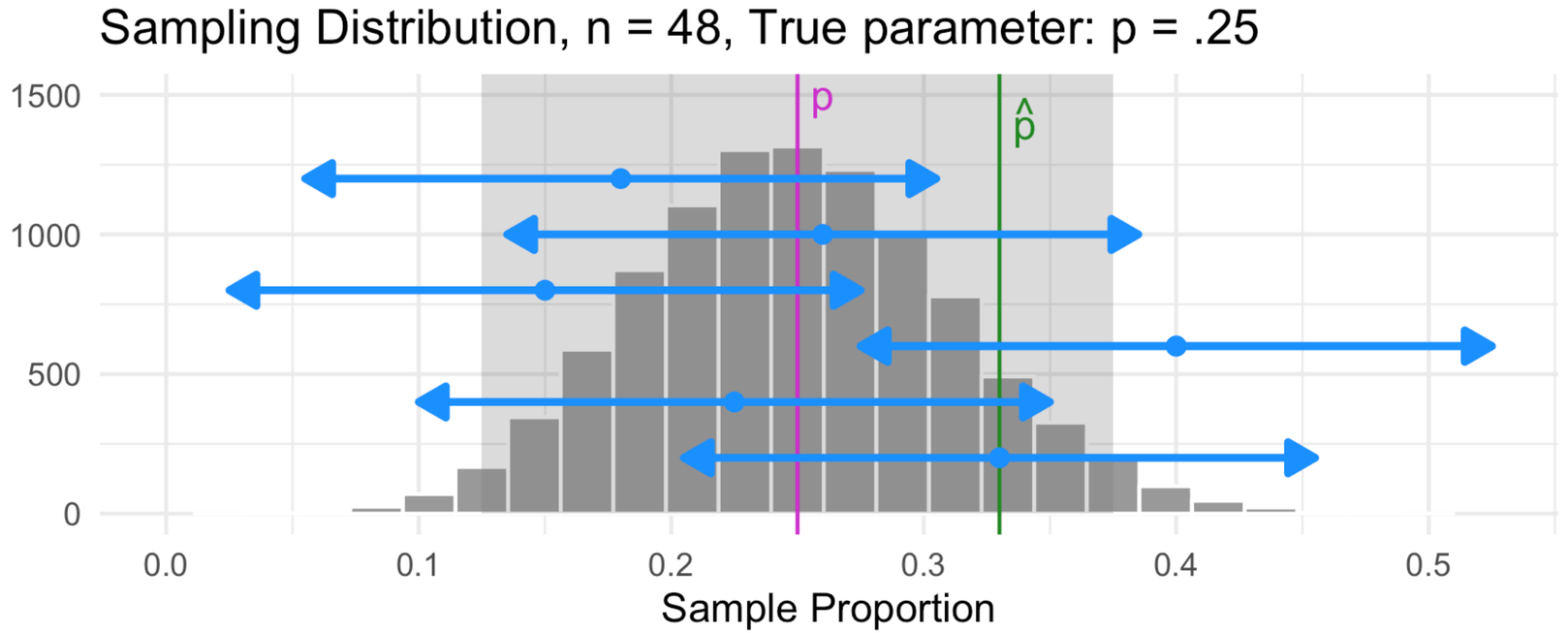
- Samples with $\hat{p}$ outside the gray region have intervals that don't contain p

Interval Estimates

Idea: Build an interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm 2 \cdot \text{SE}$$

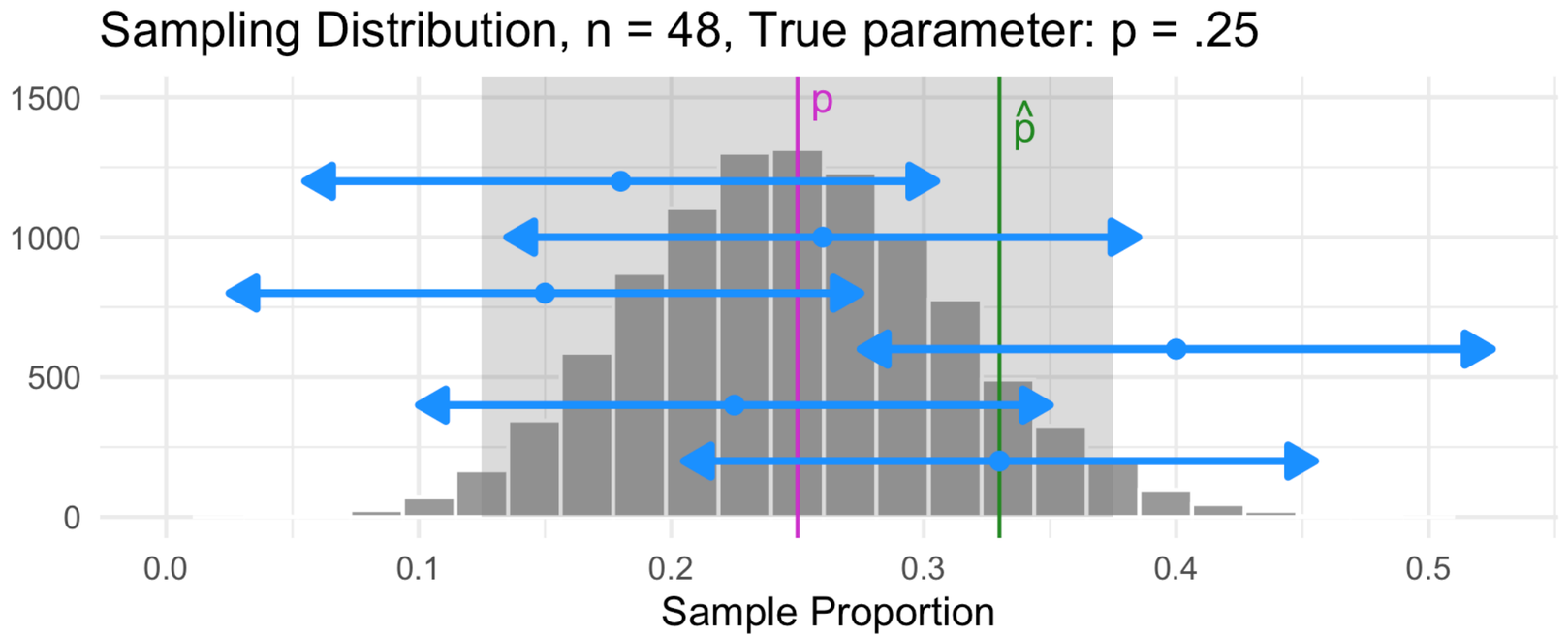
- This interval will contain the parameter p in 95% of samples!



- But 95% of all samples will have intervals that do contain p

Confidence levels

- On the previous slide, 95% of confidence intervals contained the true parameter, p.
- Thus, we call each interval estimate a **95% confidence interval**.
- Here, 95% is our **confidence level**: The success rate for our estimation technique.
  - e.g., The pizza interval 0.33 ± 0.13 has confidence level of 95%
- The confidence level corresponds to **the percentage of samples that would yield a corresponding confidence interval containing the true value of the parameter**.


Sampling Distribution, n = 48, True parameter: p = .25
1500
1000
500
0
p
p̂
0.0
0.1
0.2
0.3
0.4
0.5
Sample Proportion

Confidence levels

- **Confidence Level:** the percentage of samples that would yield a corresponding confidence interval containing the true value of the parameter.
- For example, were we to:
  - Repeatedly draw samples from the population, and
  - Create 95% confidence intervals in each sample…
  - 95% of those confidence intervals would contain the true parameter

Recap: Confidence Intervals

- Confidence intervals consist of 1. **an interval estimate** and 2. **a confidence level**.
- In our pizza party sample, we estimated the true proportion of vegetarian pizza-eaters was between 0.20 **and** 0.46, with 95% **confidence**.
- What does “95% confidence” mean?
  - It’s the success rate **of the process**
  - For 95% of all samples, the interval we construct will actually contain the parameter.
- Worth emphasizing: it’s the success rate **of The Process, NOT the specific interval you calculated in your sample.**
  - The parameter is either in your interval, or it’s not – there’s no success rate there!

Think-Pair-Share

Q: What's the difference between these two interpretations of a 95% confidence interval?

- For 95% of all samples, the interval we construct will actually contain the parameter.
- For a given interval, there is a 95% chance that the parameter will fall in the interval.

Think-Pair-Share (Answer)

Q: What's the difference between these two interpretations of a 95% confidence interval?

- For 95% of all samples, the interval we construct will actually contain the parameter.
- For a given interval, there is a 95% chance that the parameter will fall in the interval.

A:

- The first one is accurate! 95% confidence refers to the accuracy of the *process*
- **The second one is wrong, an easy mistake to make!!** It confuses probability of the process with a deterministic outcome. Our sample statistic "moves" with sampling variability - the parameter does not.

Interval Width and Sample Size (n)

Idea: Build a 95% confidence interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm \underbrace{2 \cdot \text{SE}}_{\text{Margin of Error}}$$

- This interval's **width is determined by the Standard Error (SE).**
- **Reminder:** The SE is the standard deviation of the sampling distribution.
- **Q:** What do we know about the SE as the sample size (n) increases?
- **Q:** What does this imply about the interval's width as the sample size (n) increases?
  - The SE gets smaller as n increases!
  - Our interval becomes narrower as n increases!

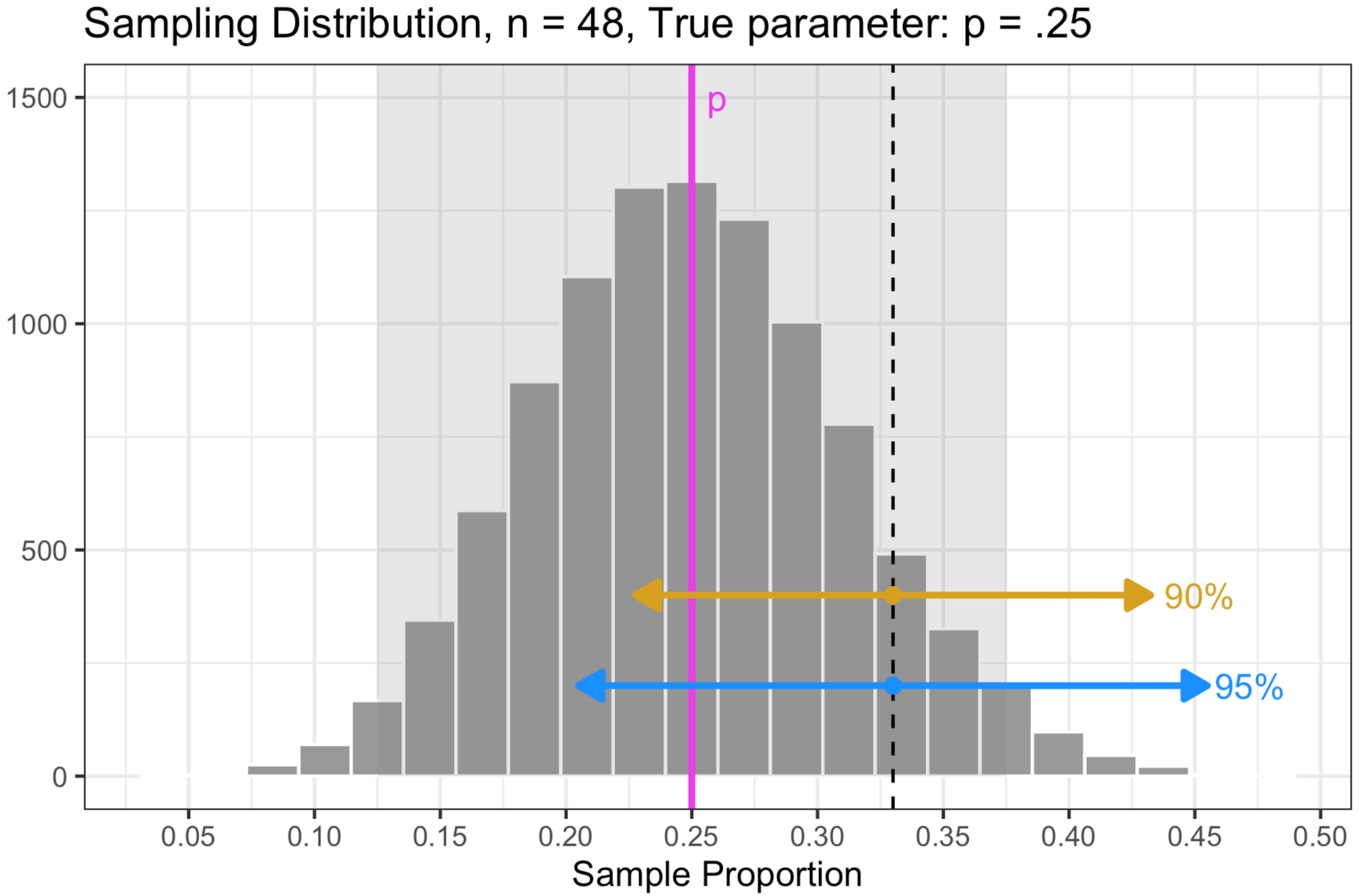
Interval Width and Confidence Level

Idea: Build an 95% confidence interval centered at the sample statistic, with a margin of error of $2 \cdot SE$:

$$\hat{p} \pm \underbrace{2 \cdot \text{SE}}_{\text{Margin of Error}}$$

Interval Width and Confidence Level

- **Takeaway: The confidence level also determines the width of the interval!**
  - higher confidence (e.g., 95%) means a *larger* margin of error (*wider* interval)
  - lower confidence (e.g., 90%) means a *smaller* margin of error (*narrower* interval)


Sampling Distribution, n = 48, True parameter: p = .25
1500
1000
500
0
p
90%
95%
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Sample Proportion

Problems(?) with Confidence Intervals

Let's say we're working with 95% confidence intervals

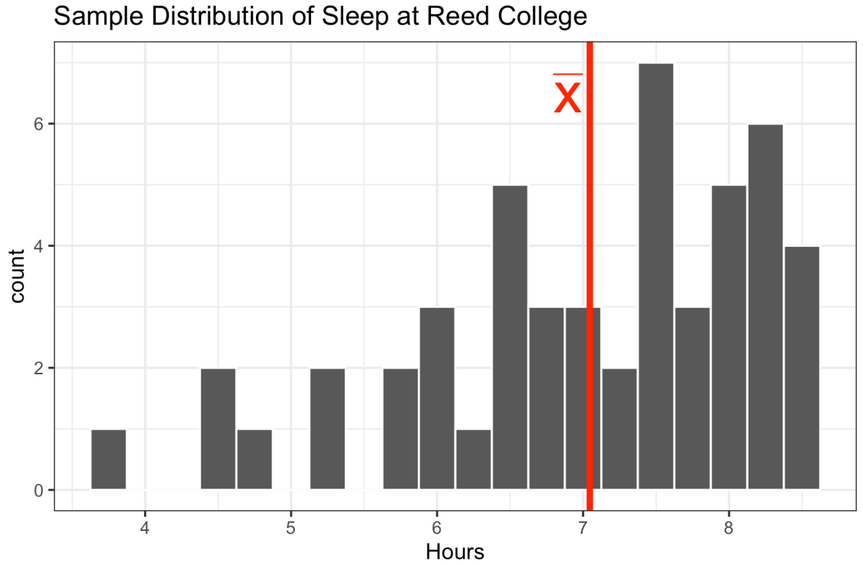
- **Problem 1:** We only have 1 sample, and we don't know if it belongs to the 95% of "good" samples, or the 5% of "bad" ones
  - **Consolation:** If I go through life constructing 95% confidence intervals, I will be right about 95% of the time
  - That's not bad!

- **Problem 2:** To make a confidence interval, we need the sampling distribution in order to compute the standard error. But in practice, we (often) don't have direct access to this.
  - **Solution:** approximate the sampling distribution via bootstrapping!

Bootstrap Confidence Intervals

Sleep for Reed Students

Suppose we wish to estimate **how many hours Reed students sleep, on average** with a sample of 50 students, who we surveys on hours of sleep.




```r
head(sleep, 4)
```

```
  Student    Hours
1       1 7.476953
2       2 7.717175
3       3 8.281295
4       4 7.117568
```

```r
sleep %>% summarize(MeanSleep = round(mean(Hours), 3))
```

```
  MeanSleep
1     7.046
```

- Is the true average hours of sleep at Reed is 7.046?
  - Surely not! This is just one sample of size 50
- Let's create a *confidence interval* for the true average hours
  - We can **use the bootstrap distribution to estimate the SE** needed for the interval

Bootstrap Average Sleep

Create the bootstrap samples:

- Q: What is each argument of `rep_sample_n()` doing?
- Each bootstrap sample consists of 50 observations sampled *with replacement* from the original sample (`size = 50`)
- We have a total of 10,000 bootstrap samples (`reps = 10000`)

```
1 bootstrap_samples <- sleep %>%
2   rep_sample_n(size = 50, replace = TRUE,
3                reps = 10000)
4 bootstrap_samples
```

```
# A tibble: 500,000 × 3
# Groups:   replicate [10,000]
   replicate Student Hours
       <int>   <int> <dbl>
 1         1      28  4.40
 2         1      12  6.45
 3         1       7  7.71
 4         1       4  7.12
 5         1       9  7.12
 6         1       1  7.48
 7         1      27  7.26
 8         1      39  6.38
 9         1      37  3.74
10         1      44  7.98
```

Bootstrap Average Sleep

Compute bootstrap statistics: (Mean of each *bootstrap* sample)

```
bootstrap_stats <- bootstrap_samples %>%
  group_by(replicate) %>%
  summarize(x_bar = mean(Hours))
```

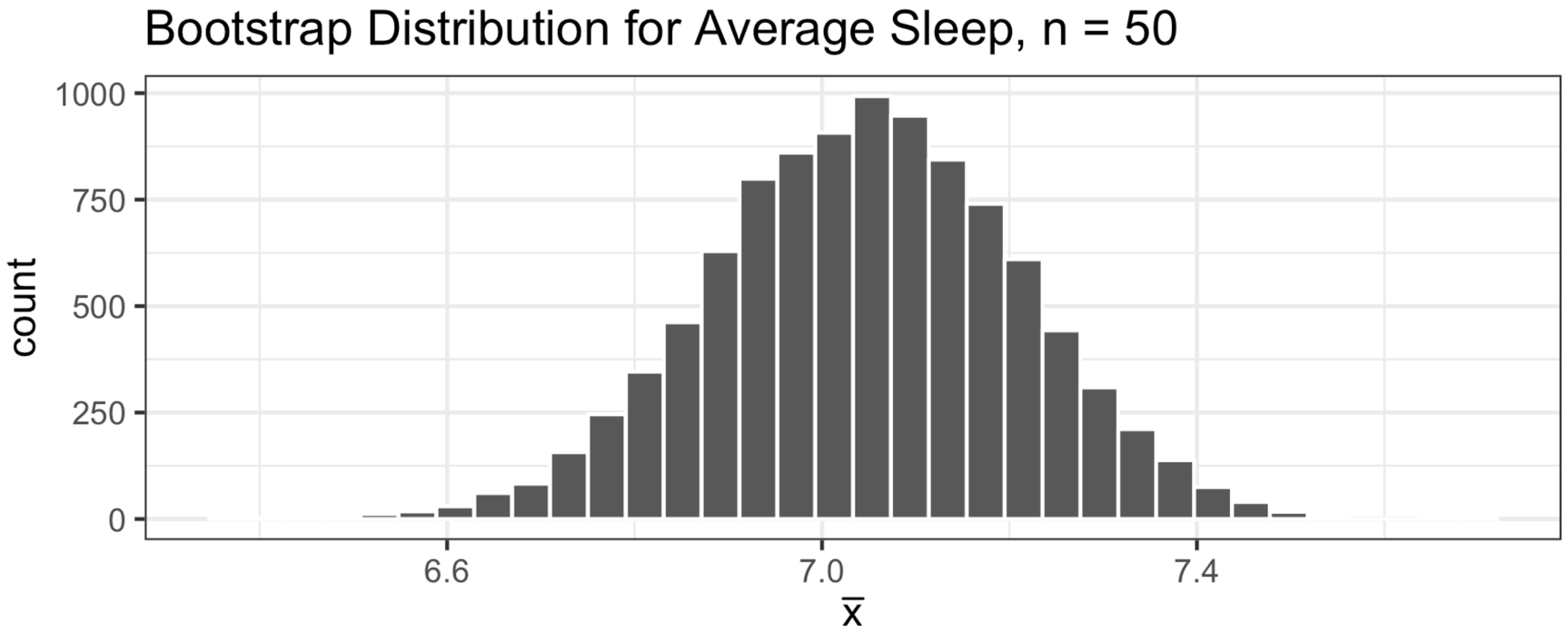
```
bootstrap_stats
```

```
# A tibble: 10,000 × 2
   replicate x_bar
       <int> <dbl>
 1         1  7.05
 2         2  6.97
 3         3  6.94
 4         4  7.31
 5         5  7.36
 6         6  7.11
 7         7  6.93
 8         8  7.10
 9         9  7.06
10        10  6.76
# ℹ 9,990 more rows
```

- We now have 10,000 sample means based on the bootstrap samples, and can assess their variability

Bootstrap Average Sleep

Graph the bootstrap distribution:



- Use the bootstrap distribution to **estimate** the standard error:

```
1 bootstrap_stats %>% summarize(SE = sd(x_bar))
```

```
# A tibble: 1 × 1
     SE
```

```
  <dbl>
1 0.165
```

Confidence Interval for Average Sleep

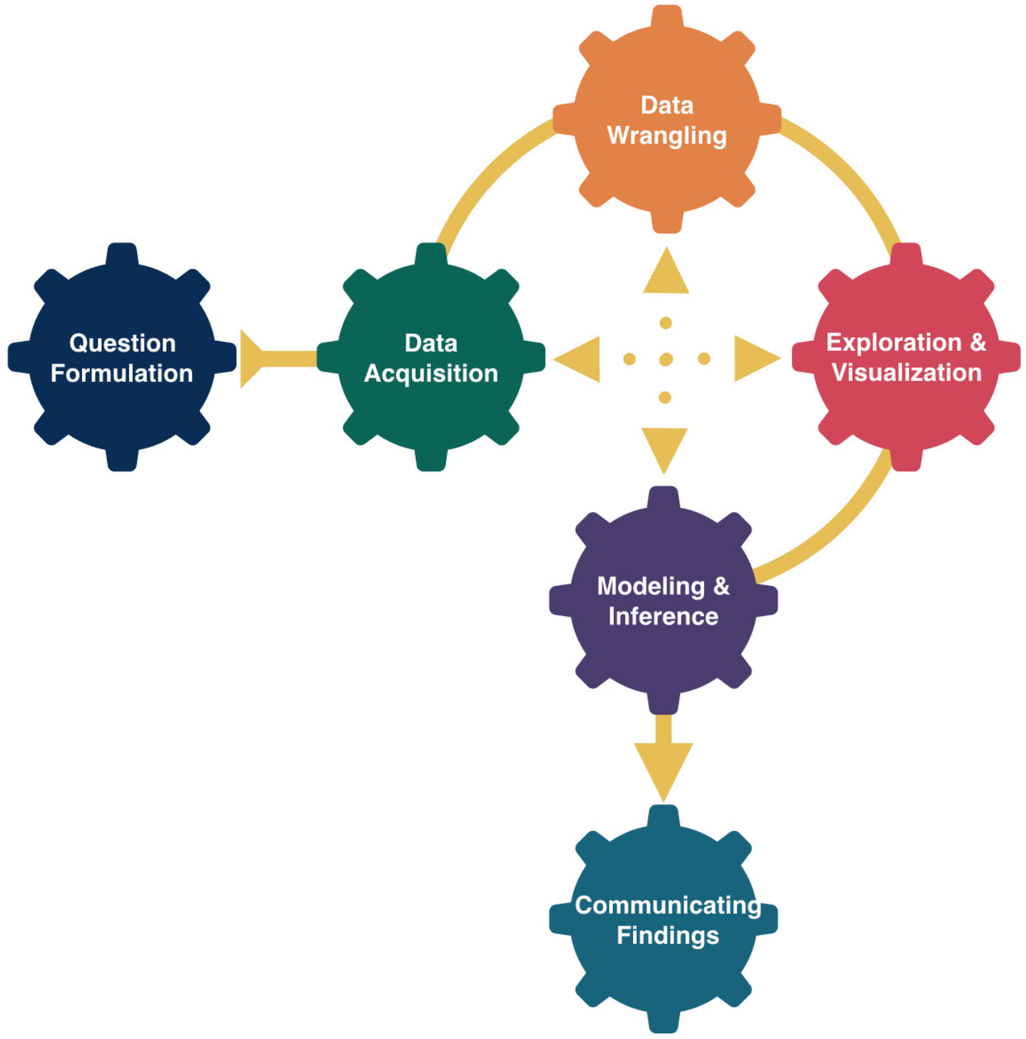
- Our sample average sleep was $\bar{x} = 7.046$.
- Based on the bootstrap distribution, this statistic has Standard Error=0.165.
- Our 95% confidence interval for the true average hours of sleep for Reed students is:

$$7.046 \pm 2 \cdot 0.165$$

  - Our best guess for average nightly sleep is that it's between 6.716 and 7.376. This method has a success rate of 95%.

Next time:

- More on changing confidence levels
- What to do with distributions that aren’t bell-shaped (percentile method)
- Confidence interval misconceptions



Confidence Intervals II

Megan Ayers

Math 141 | Spring 2026

Wednesday, Week 7

Midterm logistics

- Please be mindful of students finishing exam from the previous section when arriving for the midterm (section 2)
- Lab 5 grades are posted
- Final office hours before midterm: Megan’s office today from 3:30-5pm

Goals for today

- Review the concept of a 95% confidence interval
- Discuss (one way) of creating confidence intervals with different confidence levels.
- Interpret confidence intervals and discuss common misconceptions

Review

Setting

- There is some **population** we're interested in studying.
  - e.g., Reed College students
- There is a population **parameter** we want to know
  - e.g., average nightly hours of sleep (μ)
- We draw a **sample** from the population with sample size n
  - e.g., We survey $n = 50$ Reed students
- We provide a **point estimate** of the parameter with a statistic
  - e.g., average hours of sleep in our sample ($\bar{x} = 8.02$ hours)
- We construct an **interval estimate** centered at our statistic
  - e.g., 8.02 ± 0.16 hours

Warm-Up

1. What are the differences between a sampling distribution and a bootstrap distribution?
2. Suppose we created confidence intervals based on **distinct** samples of size $n = 10$ and $n = 100$. How might they differ?

Confidence Intervals

- A **confidence interval** gives a range of plausible values for a parameter. It usually takes the form:

$$\text{Statistic} \pm \text{Margin of Error (ME)}$$

- The Margin of Error (ME) is partially determined by our sample size:
  - Bigger samples yield smaller ME and interval (more data $\implies$ more certainty)
  - Smaller samples yield wider ME and interval (less data $\implies$ less certainty)

- Every confidence interval has a **confidence level:**
  - the percentage of samples that would yield a corresponding confidence interval that contains the true value of the parameter.

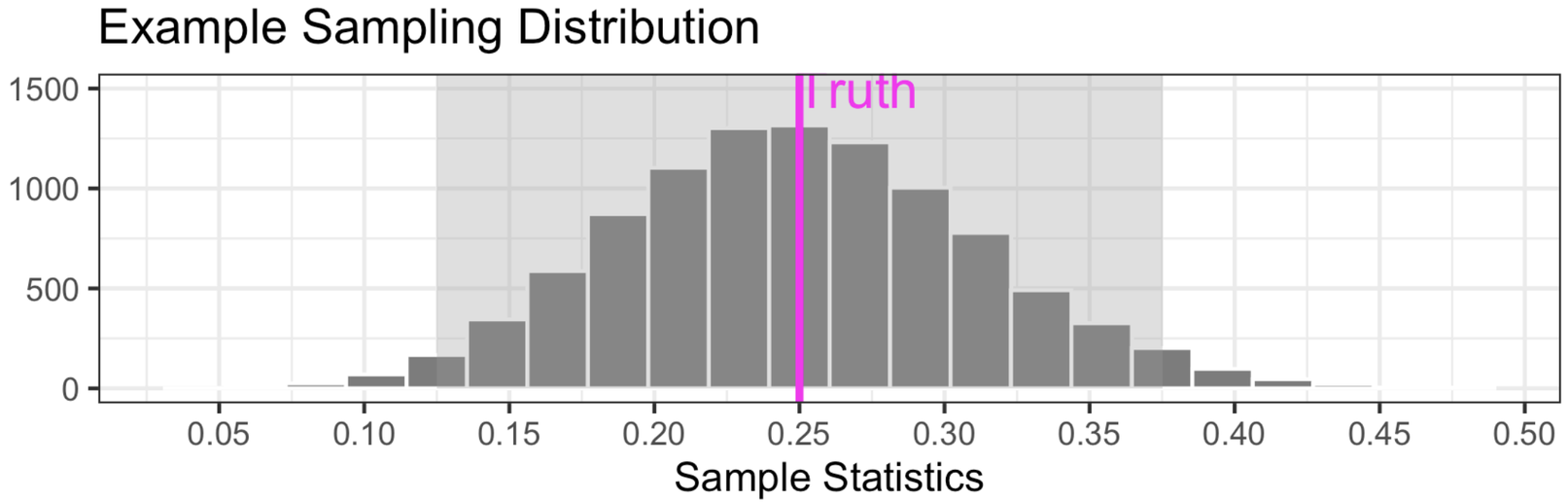
- For example, were we to:
  - Repeatedly draw samples from the population
  - Create 95% confidence intervals in each sample
  - 95% of those confidence intervals would contain the true parameter

Sampling Distributions can determine the Margin of Error

- Reminders:
  - Sampling distribution is *often* bell-shaped, with the mean equal to the parameter
  - In bell-shaped (normal) distributions, 95% of observations lie in the range:

$$\text{Mean} \pm 2 * \text{Standard Error}$$

Sampling Distributions can determine the Margin of Error



Implication: in the sampling distribution, 95% of sample statistics lie in the range:

$$\text{Parameter} \pm 2 * \text{Standard Error}$$

where the **Standard Error** is the standard deviation of the sampling distribution

- So, 95% of sample statistics are within $2 * \text{SE}$ from the parameter!
- Thus, 95% confidence intervals usually take the form:

$$\text{Statistic} \pm 2 * \text{Standard Error (SE)}$$

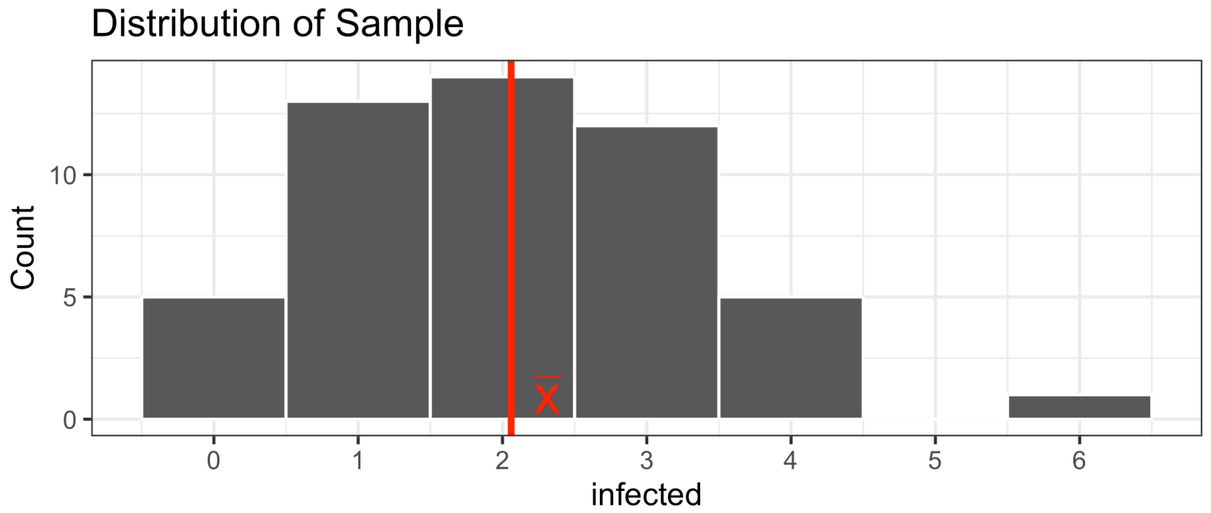
Bootstrapping Confidence Intervals

Example: Reproduction Rate for Covid-19

- Researchers are interested in the COVID-19 **reproduction rate** (the average number of individuals each infected person further infects)
- Sample 50 infected individuals and perform contract tracing.

```
  infected  n
1        0  5
2        1 13
3        2 14
4        3 12
5        4  5
6        6  1
```

```
  mean_infected
1          2.06
```



- **Goal:** Create an interval of plausible values for the reproduction rate.
- **Q:** What is the population? What is the parameter?
- **Q:** What is the sample? What is the statistic?

Bootstrap Reproduction Rate

We can use our sample to create a 95% confidence interval. **What is each step doing, and why?**

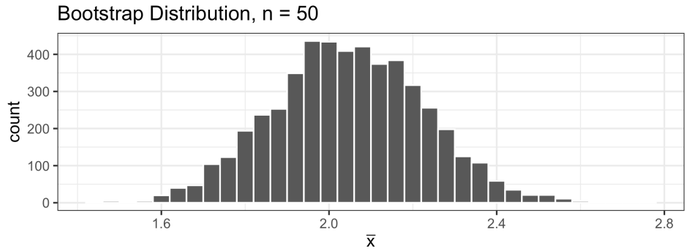
1. Step 1:

```
1  set.seed(121)
2  bootstrap_samples <- covid %>% rep_sample_n(size = 50, replace = TRUE, reps = 5000)
```

2. Step 2:

```
1  bootstrap_stats <- bootstrap_samples %>% group_by(replicate) %>% summarize(x_bar = mean(infected))
```

3. Step 3:



4. Step 4:

```
1  bootstrap_stats %>% summarize(SE = sd(x_bar))
```

```
# A tibble: 1 × 1
     SE
  <dbl>
1 0.181
```

5. Step 5:

$$\bar{x} \pm 2 \cdot SE \implies 2.06 \pm 2 \cdot 0.181$$

Bootstrap Reproduction Rate

We can use our sample to create a 95% confidence interval.

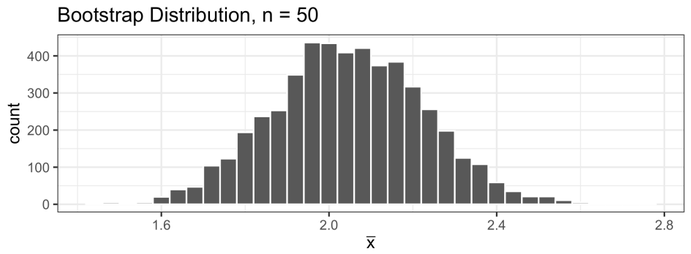
1. Create the bootstrap samples:

```
set.seed(121)
bootstrap_samples <- covid %>% rep_sample_n(size = 50, replace = TRUE, reps = 5000)
```

2. Compute bootstrap statistics within each bootstrap sample:

```
bootstrap_stats <- bootstrap_samples %>% group_by(replicate) %>% summarize(x_bar = mean(infected))
```

3. Graph the bootstrap distribution to check shape:



4. Estimate the standard error

```
bootstrap_stats %>% summarize(SE = sd(x_bar))
```

```
# A tibble: 1 × 1
     SE
  <dbl>
1 0.181
```

5. Because the bootstrap distribution is bell-shaped, we use $2\times$ the estimated SE as our **margin of error** to create a 95% confidence interval

$$\bar{x} \pm 2 \cdot SE \implies 2.06 \pm 2 \cdot 0.181$$

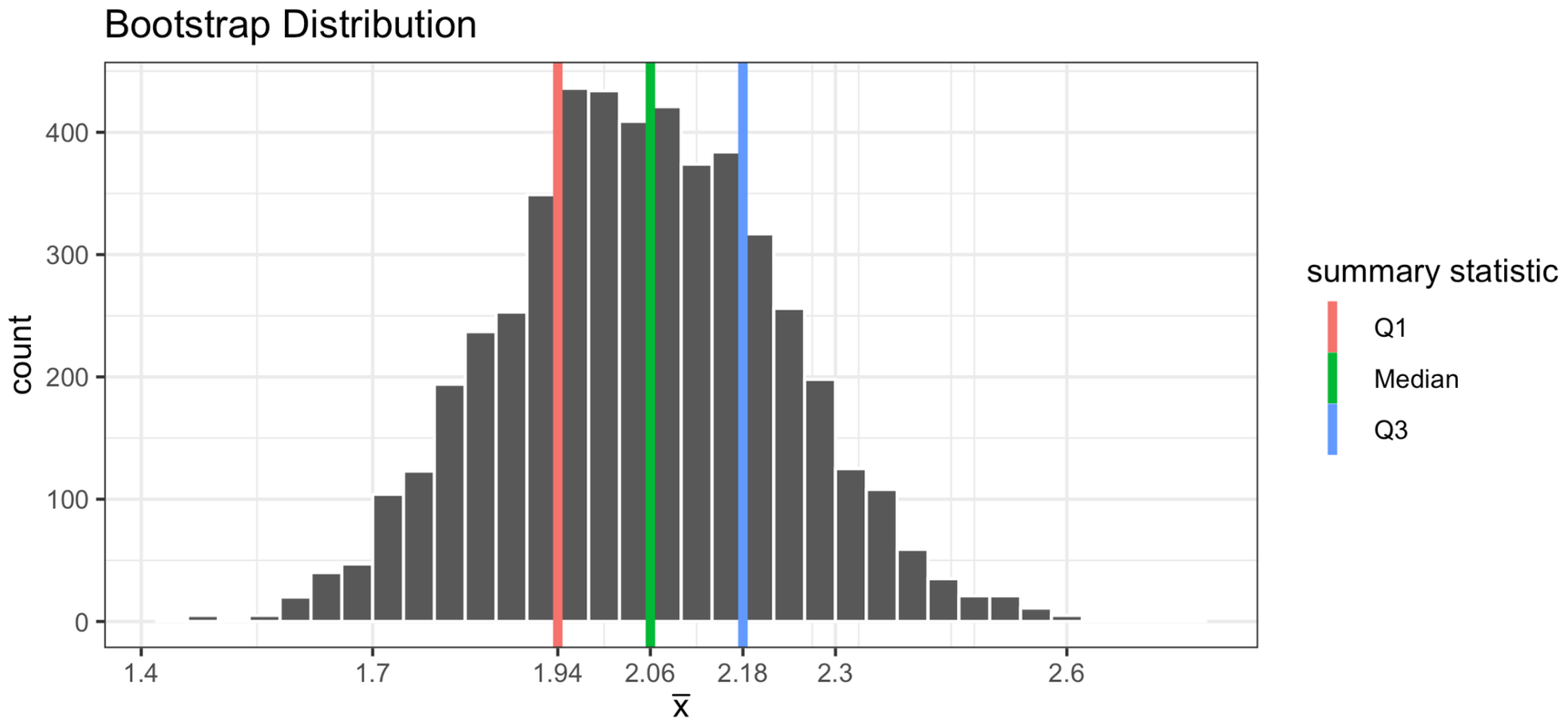
Generalized Confidence Intervals

- In the previous example, we used our knowledge that for approximately bell-shaped sampling distributions, 95% of sample statistics are within 2 SE of the population parameter
  - Suppose we instead want a **different success rate** for our estimation method
  - Or suppose we want interval estimates for **sampling distributions that are NOT bell-shaped**
- We can make these modifications again using the bootstrap approximation to the sampling distribution and make:

> **General Confidence Intervals**
>
> The $C\%$ confidence interval for a parameter is an interval estimate that is computed from sample data by a method that captures the parameter for $C\%$ of all samples.

Review: Percentiles and Quantiles

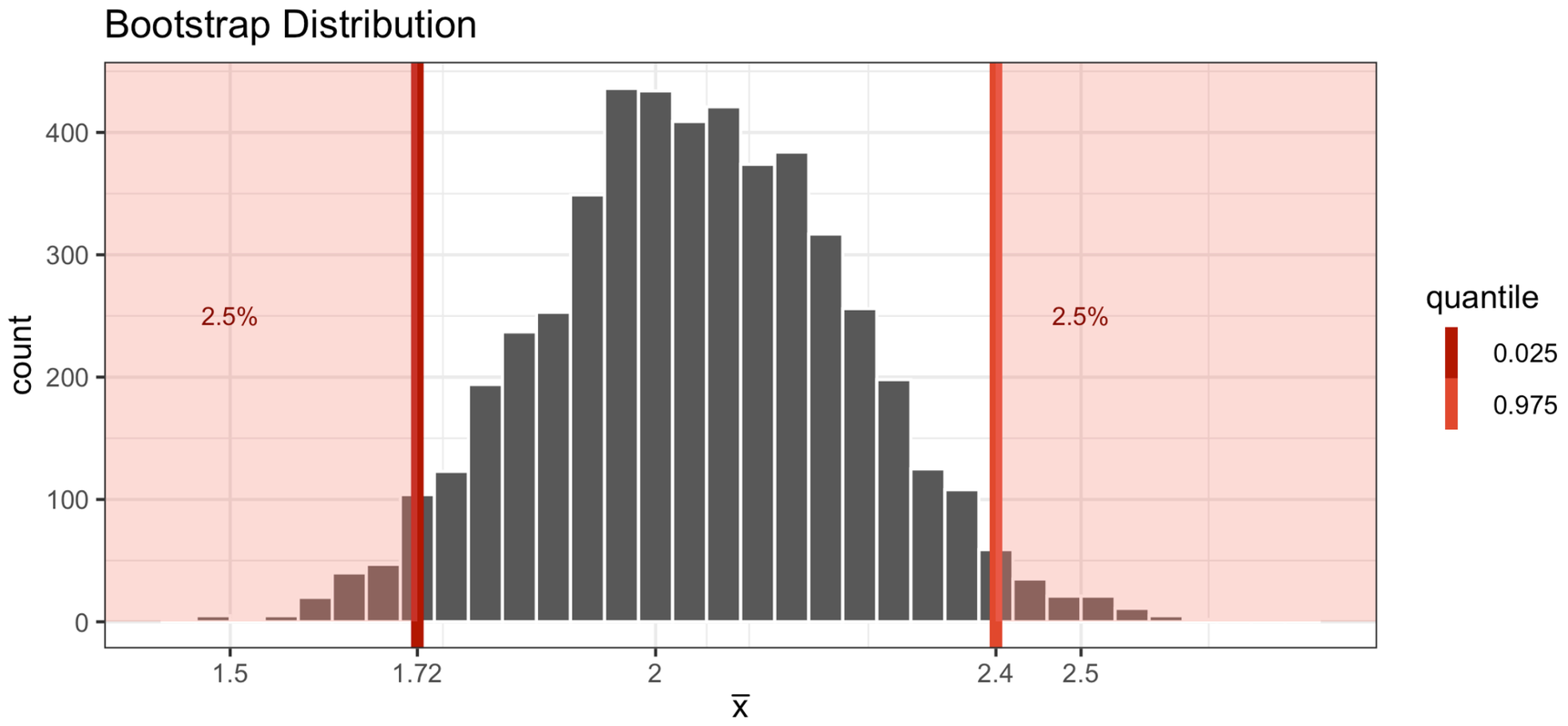
- For a number k between 0 and 100, the kth **percentile** of a distribution is the value so that $k\%$ of the data is less than or equal to that value.
  - The median is the 50th percentile of a distribution
  - 1st/3rd quartiles (Q1/Q3) are the 25th and 75th percentiles, respectively.



- For a number p between 0 and 1, the p **quantile** of a distribution is the value so that a proportion p of the data is less than or equal to that value.
  - The median is the 0.5 quantile of a distribution
  - 1st/3rd quartiles (Q1/Q3) are the 0.25 and 0.75 quantiles, respectively.

Quantiles and Percentiles

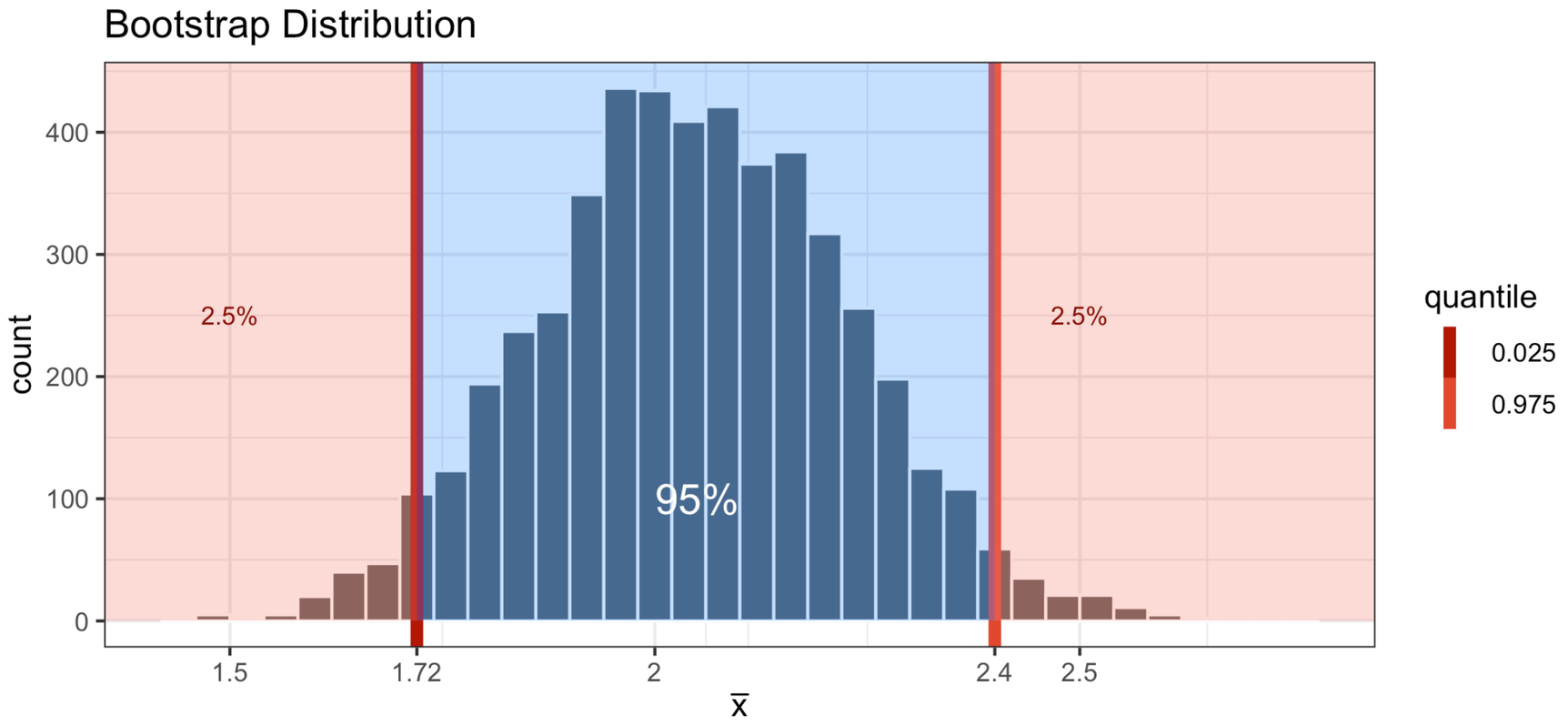
By definition, 2.5% of the data is less than the .025 quantile, and 2.5% of the data is greater than the .975 quantile



- This means that 95% of the data is between the .025 and the .975 quantiles.

Quantiles and Percentiles

By definition, 2.5% of the data is less than the .025 quantile, and 2.5% of the data is greater than the .975 quantile



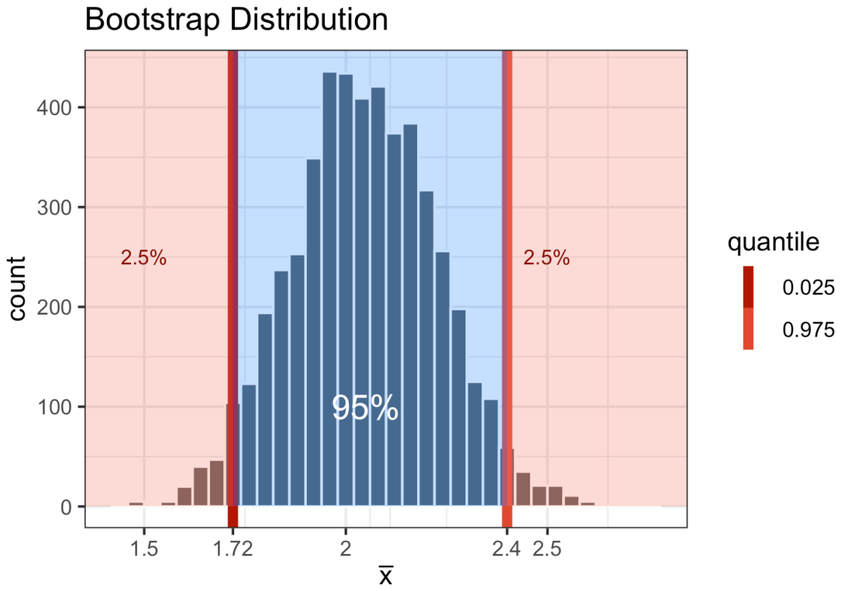
- This means that 95% of the data is between the .025 and the .975 quantiles

- *For sampling distributions that are bell-shaped*, the .025 quantile is about $2 \cdot SE$ below the mean, and the .975 quantile is about $2 \cdot SE$ above the mean
- So using the .025 and .975 quantiles is roughly equivalent to forming a 95% CI as: $\text{Statistic} \pm 2 * \text{SE}$!

95% Confidence Interval: 2 ways

1. Statistic ± 2 * SE
    - Statistic ($\bar{x}$) = 2.06
    - SE = 0.18
    - 95% CI = 1.70 to 2.42

2. Percentile Method




```
1 quantile(bootstrap_stats$x_bar, c(0.025, 0.975))
```

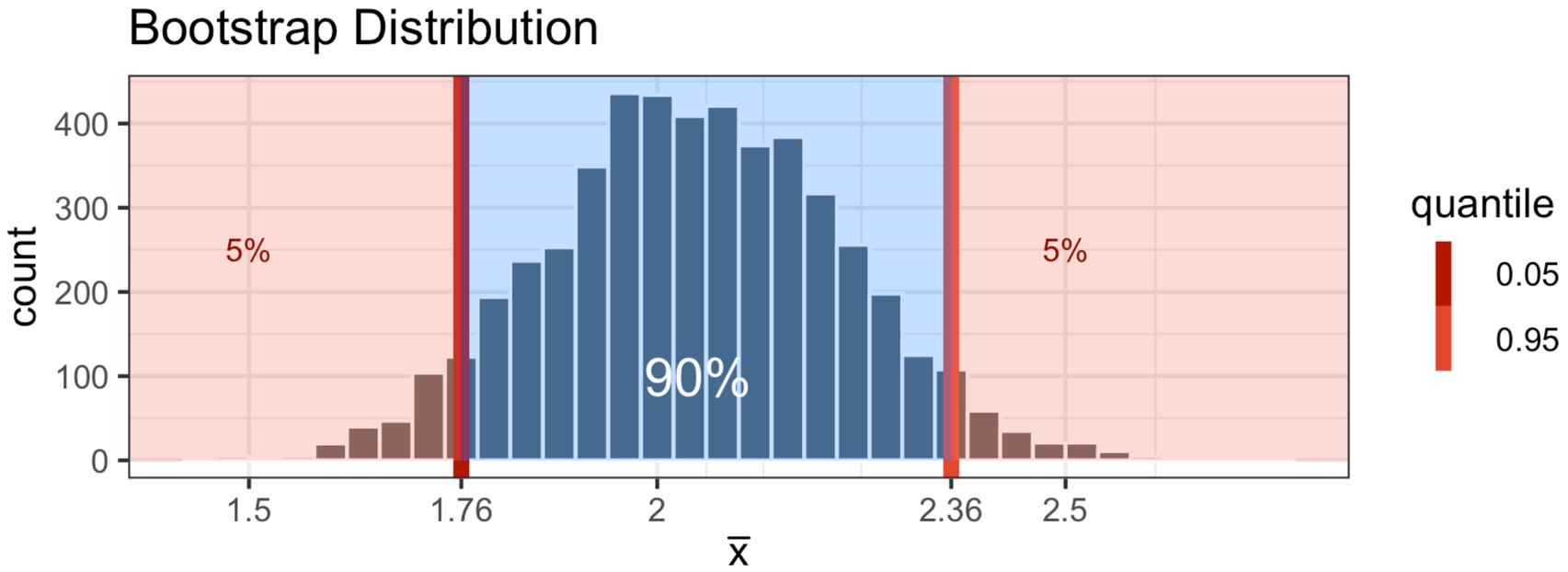
```
2.5% 97.5%
1.72  2.40
```

Percentile Method for Confidence Intervals

- **Percentile Method:** For a $C\%$ Confidence Interval, report the quantiles of the *bootstrap distribution* such that:
  - $C\%$ lies in the middle
  - $\frac{(100-C)}{2}\%$ lies on either end (i.e., "the rest" is evenly distributed on the ends)
- **Ex:** For $C\% = 95\%$ (i.e., a 95% Confidence Interval), we want
  - 95% in the middle
  - $\frac{(100-C)}{2}\% = \frac{(100-95)}{2}\% = 2.5\%$ **on either end**
- Report the **0.025 quantile** (or 2.5th percentile) and the **0.975 quantile** (or 97.5th percentile)!

The Percentile Method: Example

- Suppose we want to construct a **90% confidence interval** for the reproduction rate
  - Find the .05 and .95 quantiles in the bootstrap distribution.
  - 90% of bootstrap sample statistics will be between these values



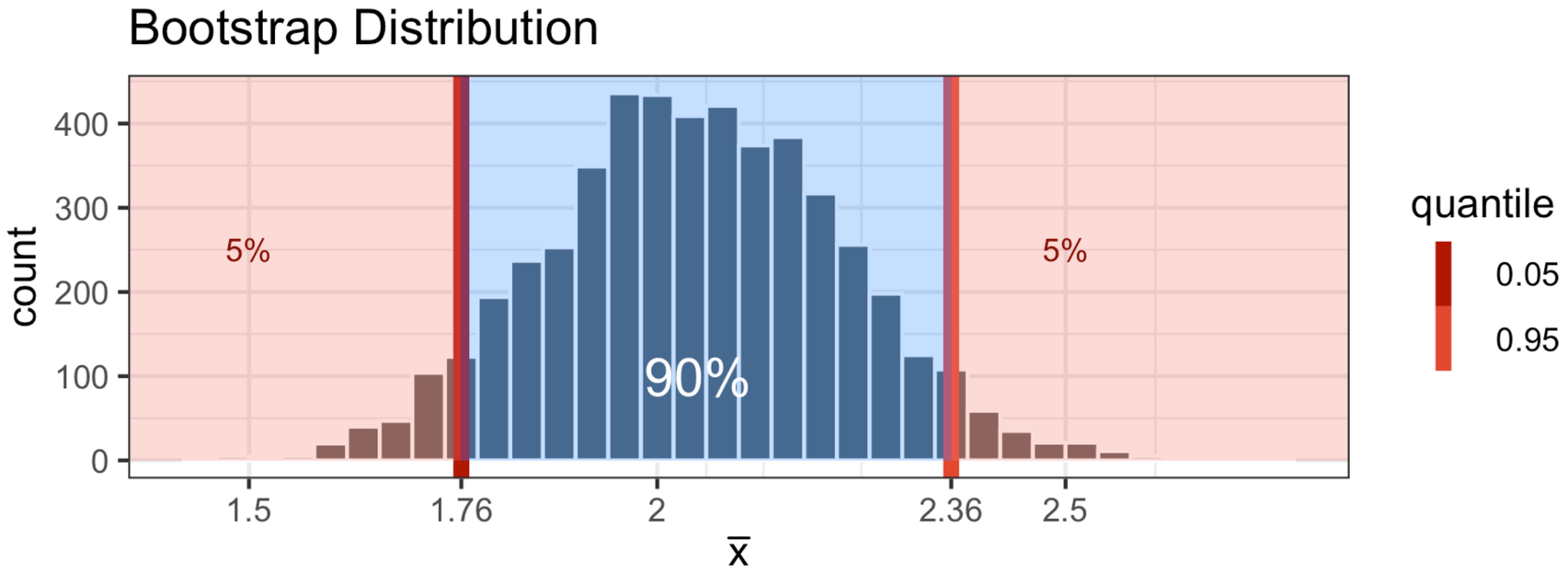
- We can use the `quantile` function in R to calculate the .05 and .95 quantiles

```r
quantile(bootstrap_stats$x_bar, c(.05, .95))
```

```
   5%   95% 
1.76 2.36 
```

The Percentile Method: Example

- Suppose we want to construct a **90% confidence interval** for the reproduction rate
  - Find the .05 and .95 quantiles in the bootstrap distribution.
  - 90% of bootstrap sample statistics will be between these values



- Our 90% confidence interval is therefore 1.76 to 2.36

```r
quantile(bootstrap_stats$x_bar, c(.05, .95))
```

```
   5%   95% 
1.76 2.36 
```

Percentile Method: Practice

With neighbor(s), name the quantiles of the bootstrap distribution you would need for:

1. 80% Confidence Interval
2. 99% Confidence Interval
3. 2% Confidence Interval *(You would never do this! This is just for fun (:)*

Answers:

1. 0.10 and 0.90 quantiles
2. 0.005 and 0.995 quantiles
3. 0.49 and 0.51 quantiles

Width of Confidence Intervals

Two factors determine the width of a confidence interval:

1. **Sample Size**
    - The Standard Error of the sampling distribution decreases as sample size increases.
    - Smaller sample size $\implies$ larger interval
    - Larger sample size $\implies$ smaller interval
2. **Confidence Level**
    - Decreasing the confidence level brings the relevant quantiles closer to the middle, decreasing the width of the interval.
    - Higher confidence level $\implies$ larger interval
    - Lower confidence level $\implies$ smaller interval

Discuss with Neighbor(s): Confidence Intervals get smaller with:

- a larger sample size
- a lower confidence level

Intuitively, why does this make sense?

Width of Confidence Intervals

Confidence Intervals get smaller with:

- a larger sample size
- a lower confidence level

Note: These reasons for getting smaller are competing in terms of certainty!

- With a larger sample size, the interval gets smaller because we're *more* certain the statistic is close to the parameter.
- With a lower confidence level, we become *less* certain the interval will contain the true parameter.

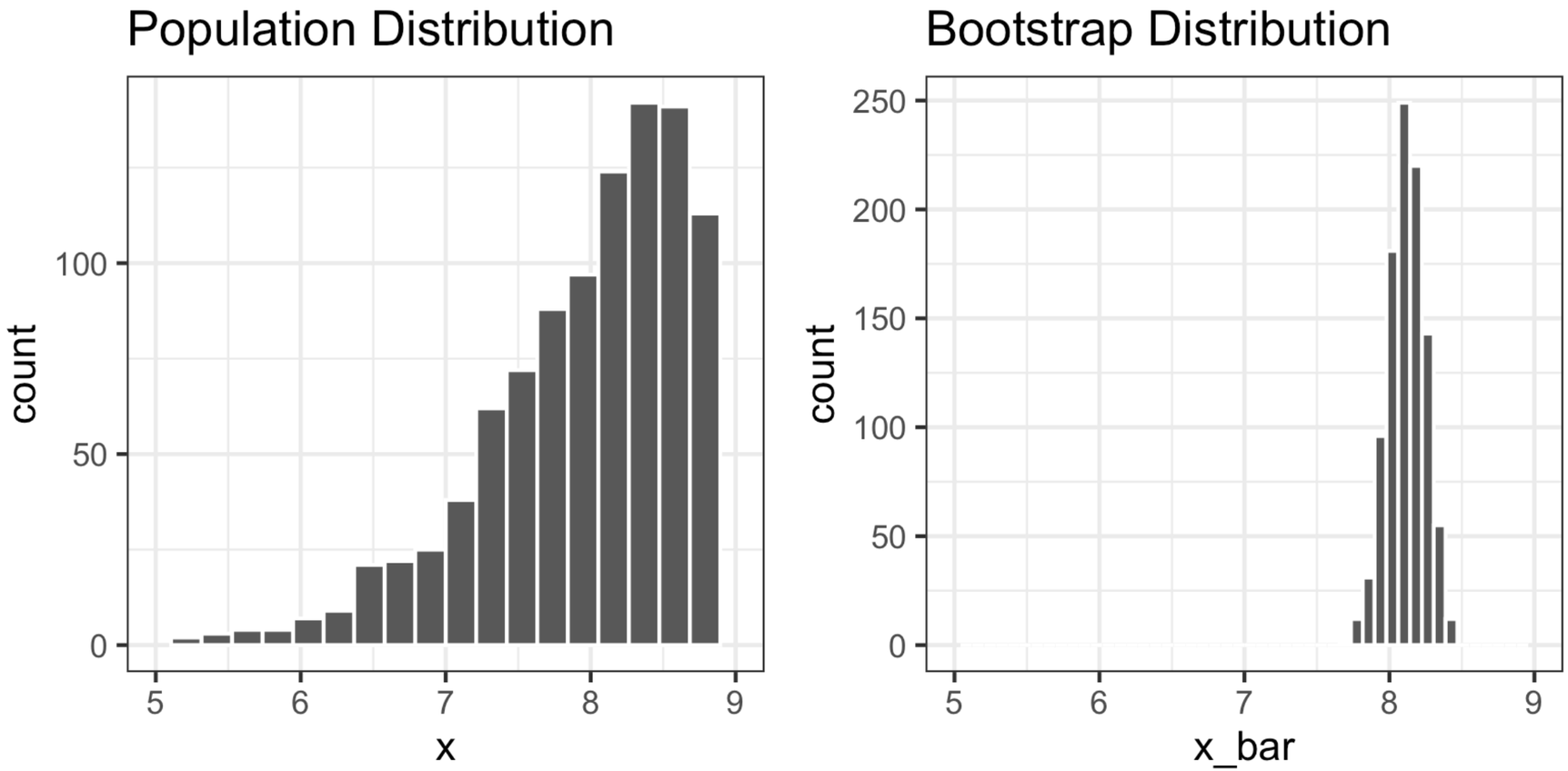
Reminder: While a lower confidence level gives you a smaller interval, there is a cost! (i.e., lower success rate)

Confidence Interval Misunderstandings

Misunderstanding 1

Suppose we wish to estimate the number of hours a Reed student sleeps on a typical night. We obtain the following 95% confidence interval: $(7.86, 8.34)$

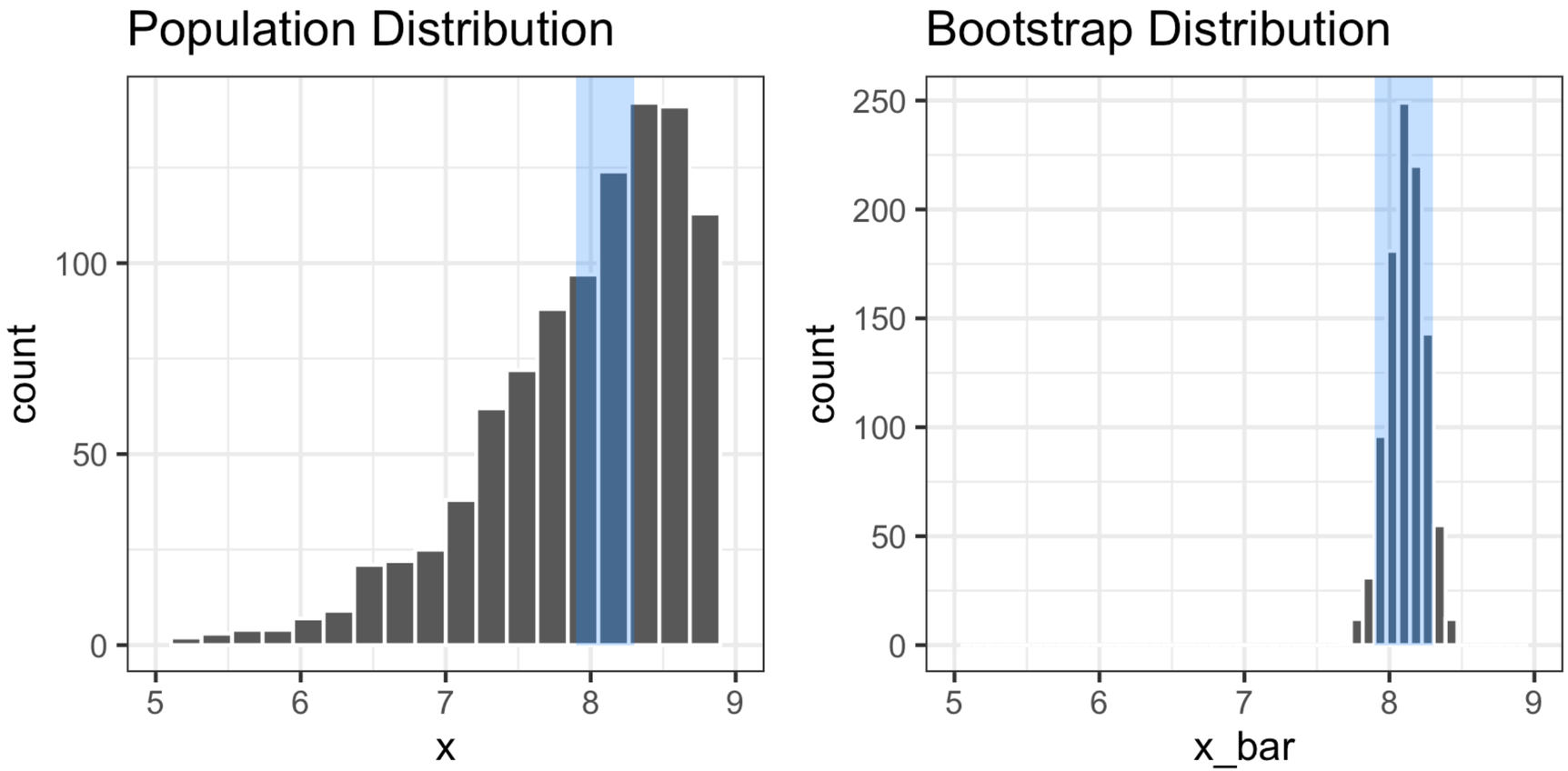
1. A 95% confidence interval **does not** contain 95% of observations in the population.



Misunderstanding 1

Suppose we wish to estimate the number of hours a Reed student sleeps on a typical night. We obtain the following 95% confidence interval: $(7.86, 8.34)$

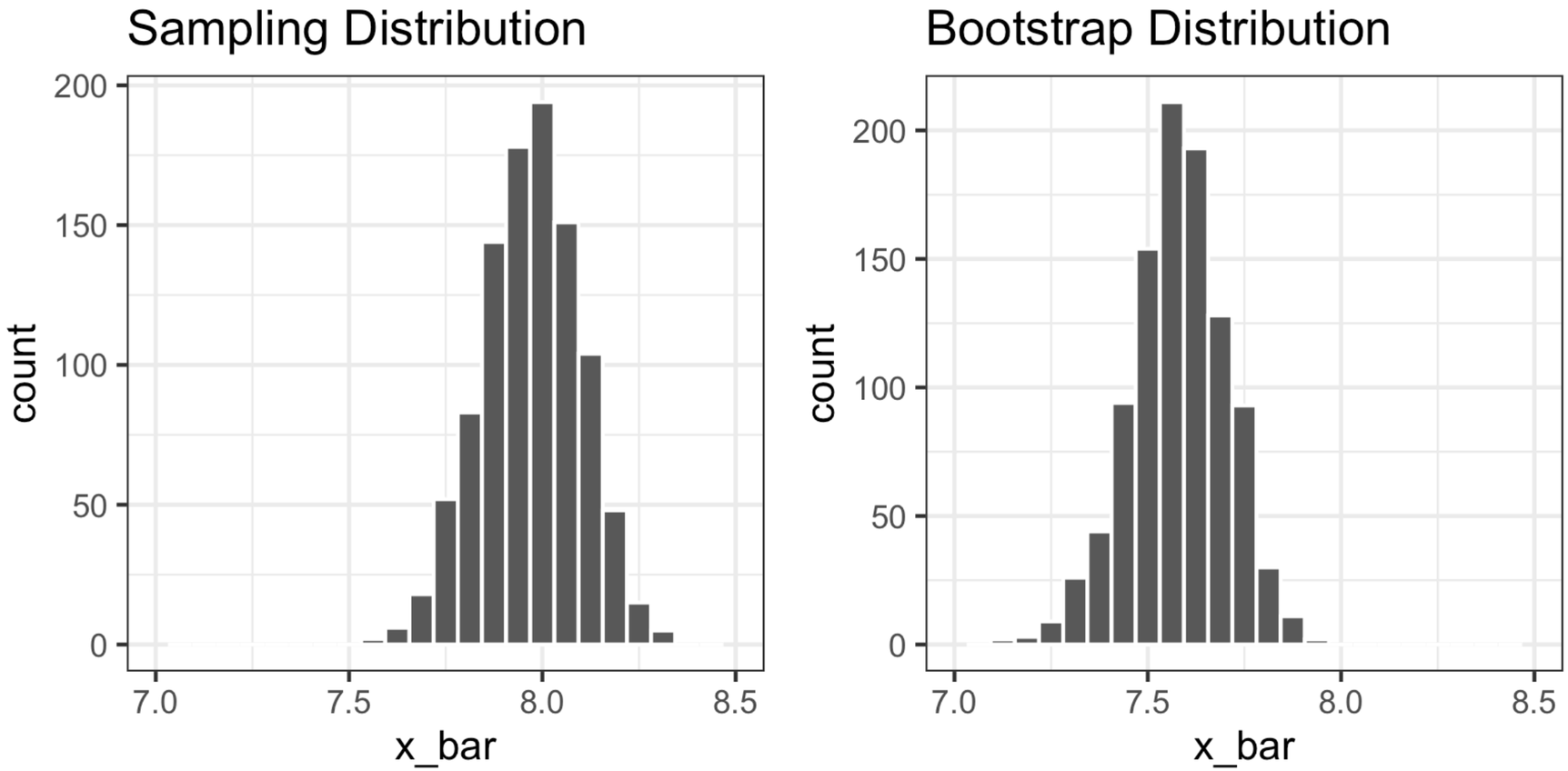
1. A 95% confidence interval **does not** contain 95% of observations in the population.



- Saying that 95% of all Reed students sleep between 7.86 and 8.34 hours should just feel wrong. That's a pretty narrow interval!

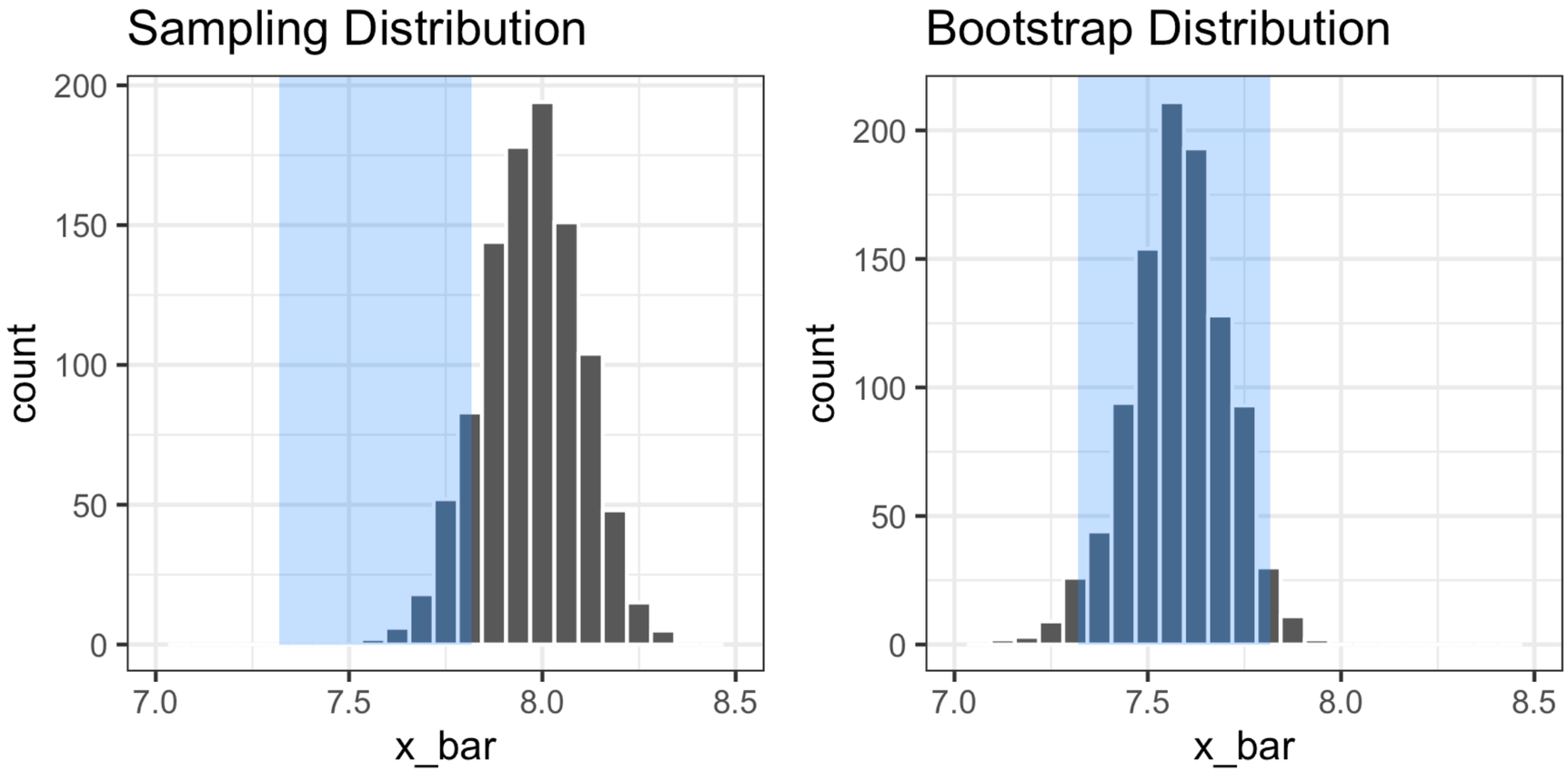
Misunderstanding 2

2. A 95% confidence interval **does not** mean that 95% of all sample means fall within the given range.



Misunderstanding 2

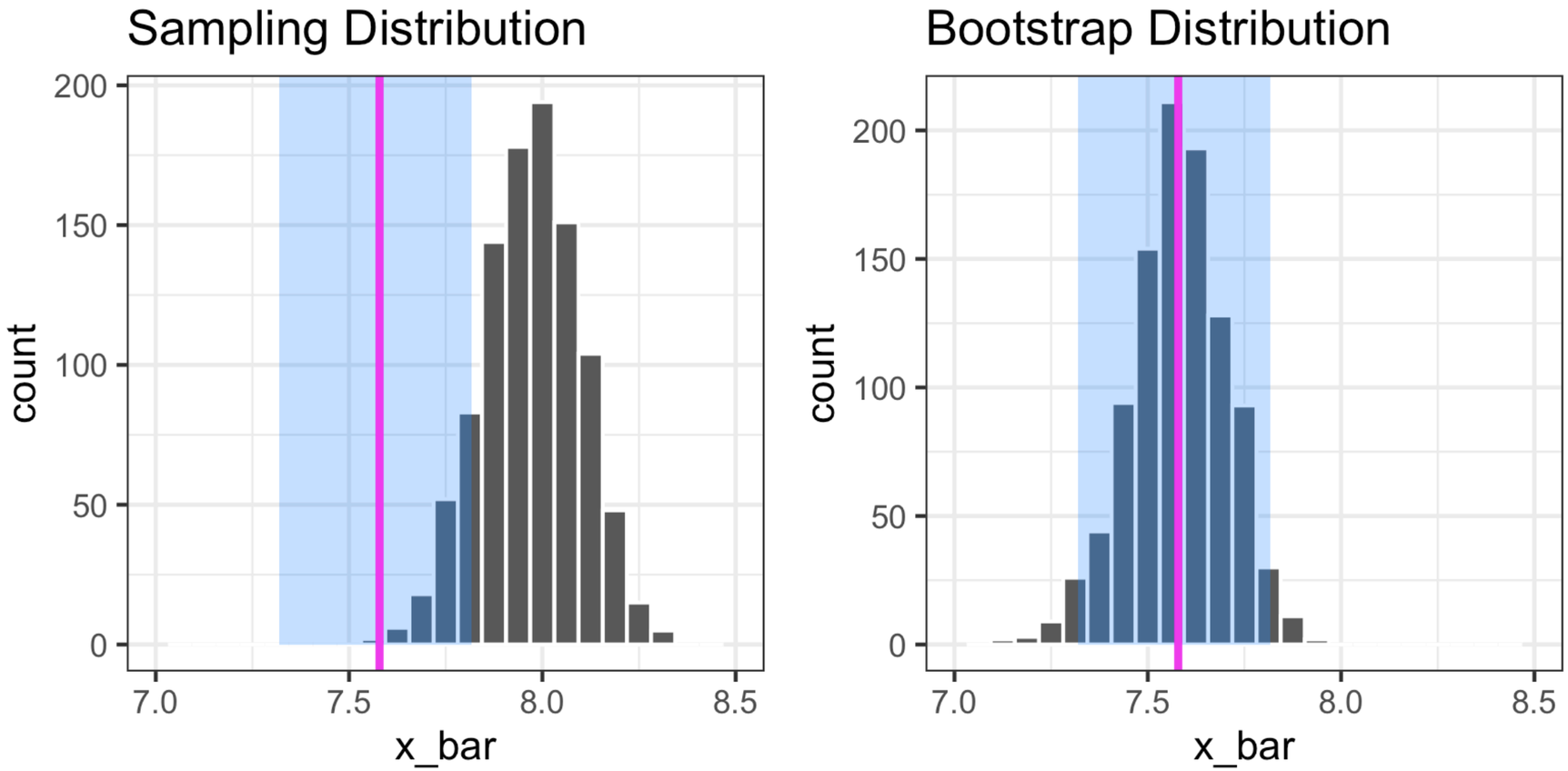
2. A 95% confidence interval **does not** mean that 95% of all sample means fall within the given range.



- **Q:** Why do the sampling distribution and bootstrap distribution look different?

Misunderstanding 2

2. A 95% confidence interval **does not** mean that 95% of all sample means fall within the given range.



- **Q:** Why do the sampling distribution and bootstrap distribution look different?

Misunderstanding 3

3. Given a 95% confidence interval, **Do Not Say:** "There is a 95% chance that the true parameter falls within my interval."

- Once we take a sample and calculate a confidence interval, there's no more randomness!
  - The interval either does or doesn't contain the (unknown) parameter.
- This is may seem like arguing over semantics – but it's an important distinction!

Instead, say either:

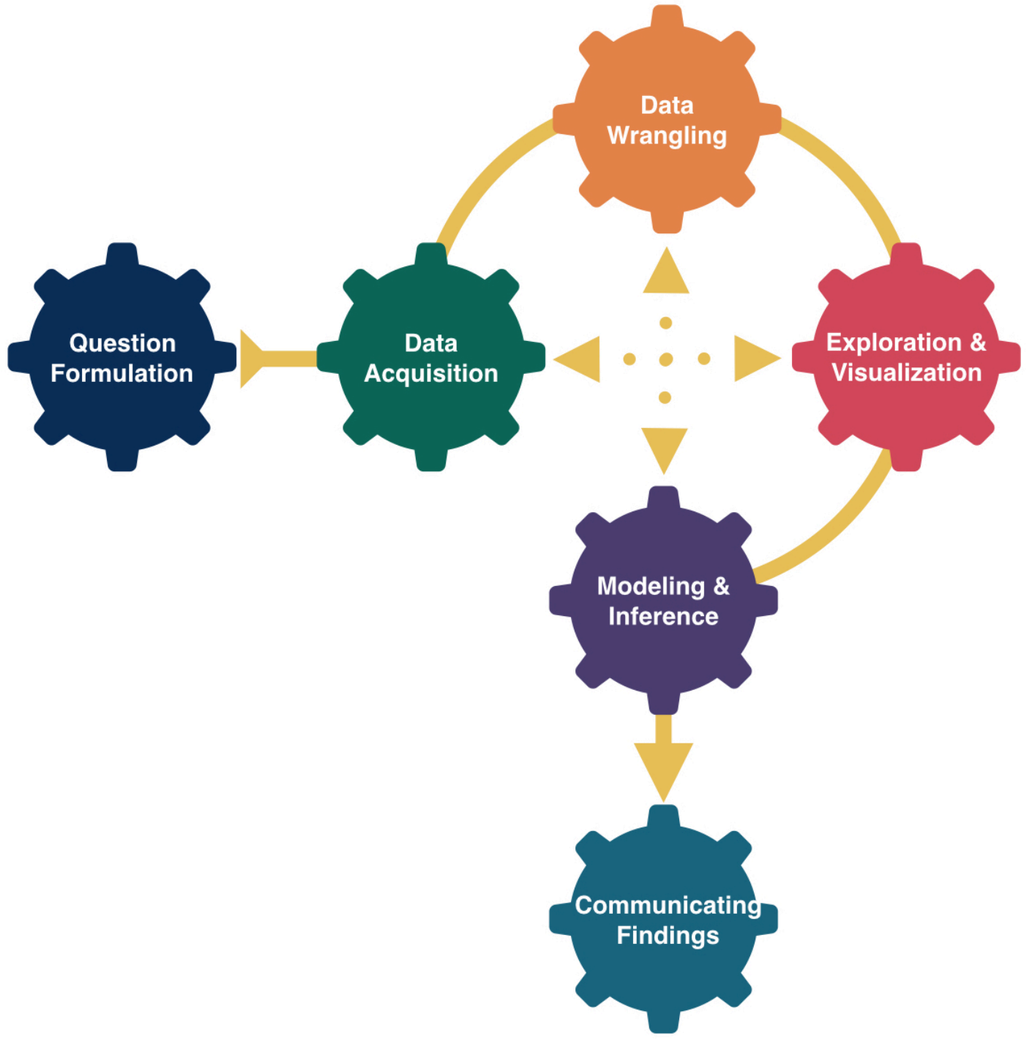
- "If we were to take many samples and calculate a 95% confidence interval for each, then 95% of them would contain the true parameter"
- "We are 95% confident that the true parameter is in our confidence interval"

Next time

- Hypothesis testing!

With any time left...

- Midterm review questions?



Hypothesis Testing I: Introduction

Megan Ayers

Math 141 | Spring 2026

Wednesday, Week 7

Reminder

- HW 6 was posted on Monday, due in 1 week (next Friday)

Goals for today

- Introduce hypothesis testing framework

Introduction

Card Guessing

- I have 5 (hearts) cards from a standard deck:
  - 10, Jack, Queen, King, Ace
- I'm going to shuffle the 5 cards and make a guess about which is on top, then reshuffle
  - Let's bet $20 that I can guess right 10 out of 10 times

- **Discuss with Neighbor(s):**
  - **Q:** If I'm guessing at random, what's the probability that I guess correctly on **one** draw?
  - **Q:** How many times out of 10 draws would you expect me to guess correctly?
  - **Q:** Which is the probability of guessing correctly 10 times in a row?
    - $\frac{1}{5}$ or a 20% chance
    - 2 times out of 10
    - $(\frac{1}{5})^{10} = 0.0000001$ or <0.001%

The guesses...

- Let's say I do this and get 10 in a row.
  - Clearly I'm cheating and you shouldn't have to pay me $20
  - But I say "Oh I just got lucky… pay me!!" … **how do you prove I'm lying?**
- **Q:** Any ideas?

- One way of doing it:
  - **Assume for a second** that I was guessing at random
  - If that's the case, then **there's a** $< 0.001\%$ **chance** of guessing 10 in a row correctly!
  - Therefore, I must be cheating (i.e., **You reject the idea that I'm guessing randomly**)

- We just (informally) did a **Hypothesis Test**
  - Framework for deciding whether or not "due to chance" is a plausible explanation
  - If the event (data, sample) that we observe is extremely unlikely, "due to chance" is not a plausible explanation.

Hypothesis Tests and "P-Values", Informally

- Say we want to disprove a hypothesis:
  - e.g., "Megan was guessing randomly, and not cheating"
- We are going to calculate something called a **P-value**:
  - **Assuming the hypothesis was true…**

$$\text{P-value} = \text{Probability of what happened in reality (or something more extreme)}$$

- e.g., Assuming Megan was guessing randomly…

$$\text{P-value} = \text{Probability guessing correctly 10 times in a row}$$

- Then we'll **reject the hypothesis if the p-value is small**

Why reject with "small" p-values?

Discuss with Neighbor(s): If the "p-value",

$$\text{P-value} = \text{Probability of what happened in reality assuming the hypothesis is true}$$

is small, **why does it make sense to reject the hypothesis?**

- In life, we tend to believe things that happen are pretty "typical" or "likely"… or else they wouldn't happen!
- With a **small p-value**, the hypothesis holding means **"You're special! You observed something really rare!"**
  - We usually don't get that lucky - so we doubt our hypothesis.
- With a **large p-value**, the hypothesis holding means **"Your observation meets expectations."**
  - In this case, we don't have much reason to doubt our hypothesis.

Hypothesis Testing Framework

Framework for Hypothesis Testing

Hypothesis Testing is a scientific experiment, and follows the general scientific method.

1. Present research question and identify **hypotheses**
    - **Null** Hypothesis (e.g., I was guessing cards randomly)
    - **Alternative** Hypothesis (e.g., I was cheating)
    - We'll express these hypotheses mathematically, *in terms of parameters*
2. Describe **Null distribution**
    - What should we expect to happen if the null hypothesis is true?
3. Obtain data, calculate relevant **Test Statistic**
    - **Test Statistic** = sample statistic based on our data
4. Calculate the **P-value**
    - **P-value** = probability of observing the Test Statistic assuming the Null Hypothesis
5. Use the P-value to **make a conclusion** on the research question

1) Identifying Hypotheses: Informal vs. Formal Hypotheses

- Before the card guessing experiment, we may have several (informal) hypotheses:
  1. Megan is guessing at random
  2. The cards are not equally likely to come up
  3. Megan is psychic
  4. Megan is not to be trusted
  5. Megan is able to do better than guessing at random
- But in order to compare these, it would be helpful to consider a set of hypotheses that:
  - Are mutually exclusive
  - Make specific statements about a **parameter**
  - Do not discuss the specific outcome of the experiment
- **Formal Hypotheses:** Let p denote the true probability that a guess is correct.

$$\textbf{Null Hypothesis: } p = 1/5 \qquad \textbf{Alternative Hypothesis: } p > 1/5$$

- The first informal hypothesis is represented by Hypothesis 1. The others are represented by Hypothesis 2, summarized in the last informal hypothesis.

1) Identifying Hypotheses

- The **null hypothesis (H_0)** is the claim we are testing. It often represents a skeptical perspective or that there is no relationship among several variables.
  - H_0: The chance of guessing correctly is 1/5, or $p = 1/5$.
  - The **null value** = value of the population parameter under the Null Hypothesis (e.g., $1/5$)
- The **alternative hypothesis (H_a)** is contrary to the null hypothesis. It is often the theory we would like to prove.
  - H_a: Megan can do better than random guessing, or $p > 1/5$.
- The Null and Alternative hypotheses are most often statements about the particular value of a population parameter
  - H_0 and H_a are **never** statements about particular values of **sample statistics**. They are **hypotheses** and should be able to be expressed before any observation of data.
    - **Incorrect** H_0: The proportion of correct guesses in 10 guesses is $\hat{p} = 1/5$.
    - **Incorrect** H_a: The proportion of correct guesses in 10 guesses is $\hat{p} > 1/5$.

1) Identifying Hypotheses: Determining the Null Hypothesis

Framework for Hypothesis Testing

1. Present research question and identify hypotheses
2. **Describe “Null” distribution - What should we expect to happen due to randomness if the null hypothesis is true?**
3. Obtain data, calculate relevant “Test Statistic”
4. Calculate the “P-value”
    - P-value = likelihood of observing the Test Statistic or something more extreme assuming the Null Hypothesis
5. Use the P-value to make a conclusion on the research question

2) Describe the Null Distribution: Likelihood of Observing Sample Statistic

- To compare Null and Alternate Hypotheses, we need to quantify how likely it is to observe a particular sample statistic, *if the null hypothesis were true*.
  - If I'm guessing at random, how many correct guesses do you expect?
  - What is the greatest number of correct guesses you would plausibly expect to see?
  - If I had 4 correct guesses, would you think that I'm guessing at random?
  - How likely is is that I would get 7 or more correct guesses, if I'm guessing at random?
- To answer questions like these, we need to know the **distribution of the statistic of interest, if the null hypothesis were true**.
- This distribution is called the **Null Distribution** and is the theoretical *sampling distribution* for the statistic if the *null hypothesis* were true.
- We can approximate the Null Distribution using simulation or theory.

2) Describe the Null Distribution: A Model of Card Guessing

We can use R to simulate one experiment of 10 guesses by…

- Creating a data frame consisting of correct and incorrect guesses

```
guesses <- data.frame(
  correct = c(0, 0, 0, 0, 1))
guesses
```

```
  correct
1       0
2       0
3       0
4       0
5       1
```

- Sampling from this data frame *with replacement* 10 times

```
guesses %>%
  rep_sample_n(size = 10, replace = T)
```

```
   replicate correct
1          1       0
2          1       0
3          1       0
4          1       0
5          1       0
6          1       0
7          1       1
8          1       0
9          1       0
10         1       1
```

This gives us a $\hat{p} = 2/10 = 0.2$

2) Describe the Null Distribution: A Model of Card Guessing

We can use R to simulate 2000 experiments of 10 guesses by putting in `reps = 2000` (`reps = 1` is the default).

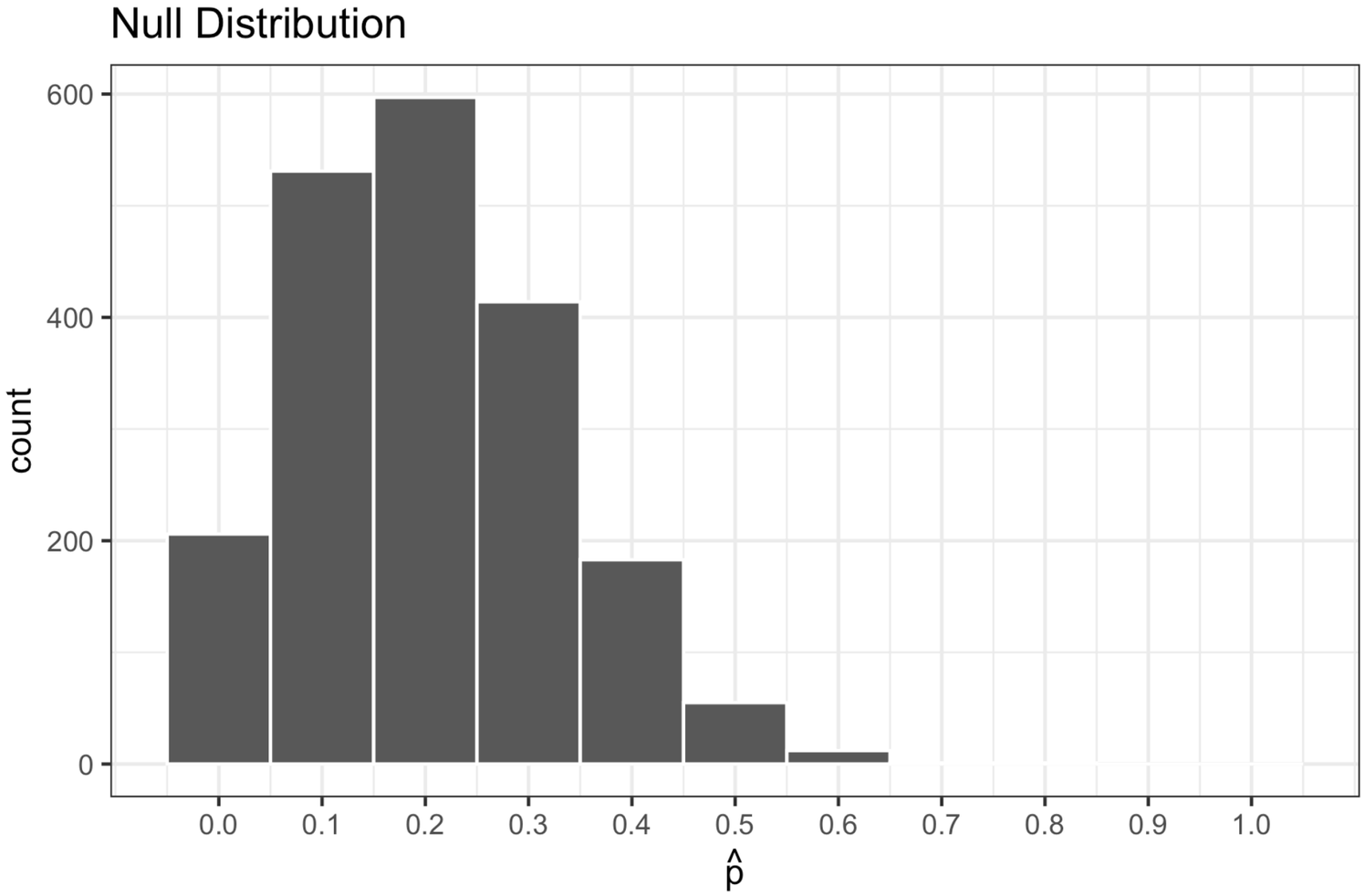
```
guesses %>%
  rep_sample_n(size = 10,
               replace = T,
               reps = 2000) %>%
  group_by(replicate) %>%
  summarize(n_correct = sum(correct)) %>%
  mutate(p_hat = n_correct/10)
```

```
# A tibble: 2,000 × 3
   replicate n_correct p_hat
       <int>     <dbl> <dbl>
 1         1         2   0.2
 2         2         5   0.5
 3         3         1   0.1
 4         4         2   0.2
 5         5         0   0
 6         6         3   0.3
 7         7         0   0
 8         8         2   0.2
 9         9         1   0.1
10        10         4   0.4
# ℹ 1,990 more rows
```

2) Describe the Null Distribution: Visualizing

- We can use a histogram to visualize the Null Distribution of the sample proportion $\hat{p}$

```
1 null_stats %>% ggplot(aes(x = p_hat))+geom_histogram(bins = 11, color = "white")
```

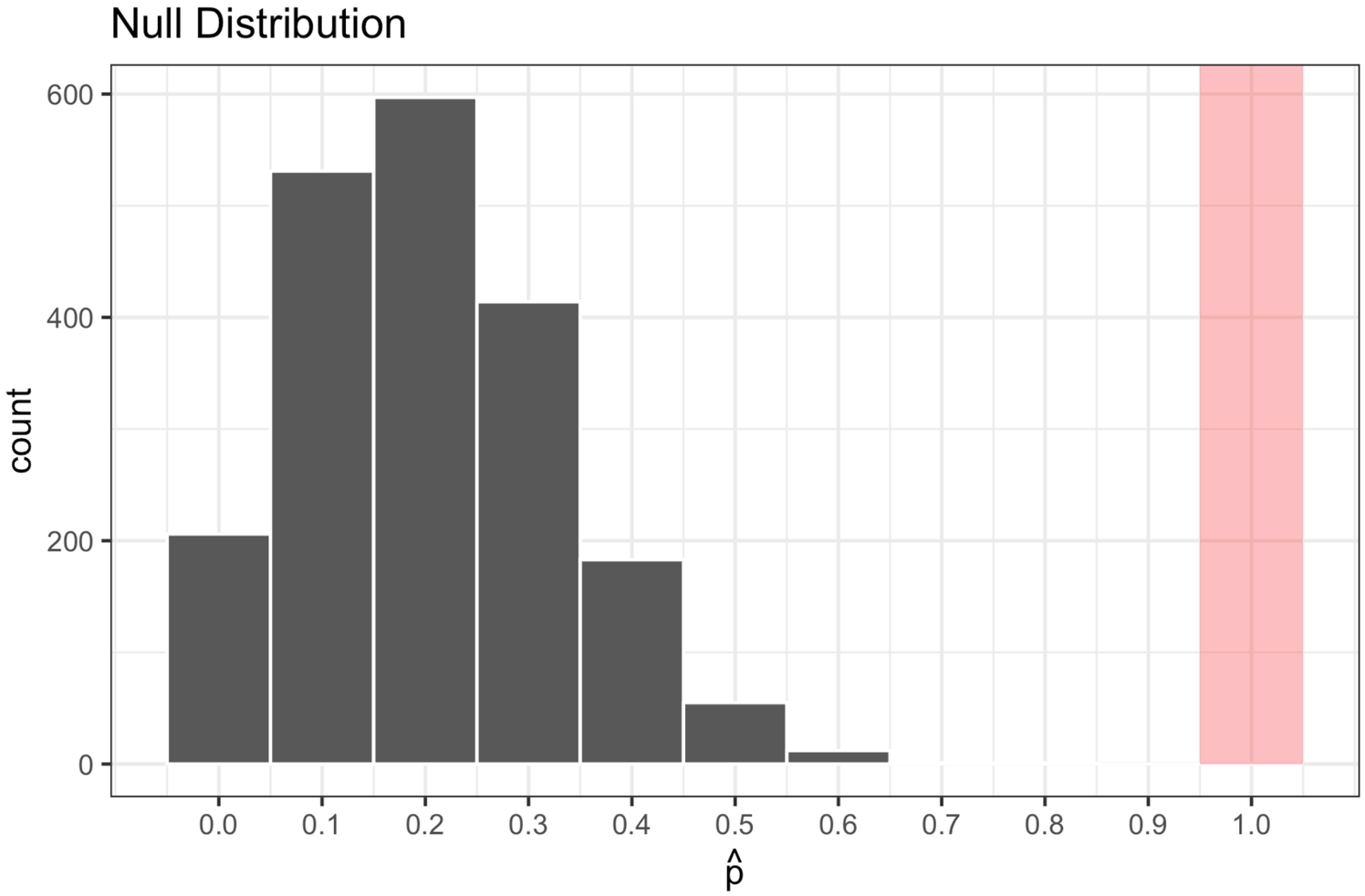


- **Q**: How often would we have observed $\hat{p} = 1.0$?

2) Describe the Null Distribution: Visualizing

- We can use a histogram to visualize the Null Distribution of the sample proportion $\hat{p}$

```
1 null_stats %>% ggplot(aes(x = p_hat))+geom_histogram(bins = 11, color = "white")
```



- **Q**: How often would we have observed $\hat{p} = 1.0$?

Framework for Hypothesis Testing

1. Present research question and identify hypotheses
2. Describe “Null” distribution
3. **Obtain data, calculate relevant “Test Statistic”**
    - **Test Statistic = sample statistic based on our data**
4. Calculate the “P-value”
    - P-value = likelihood of observing the Test Statistic or something more extreme assuming the Null Hypothesis
5. Use the P-value to make a conclusion on the research question

3) Calculate the Test Statistic

- We've already done this with my card drawing:
  - We found that $\hat{p} = 1$ (I guessed right in 10 out of 10 draws)
- In the context of hypothesis tests, we typically call the sample statistic (e.g., $\hat{p}$) the **Test Statistic**

Framework for Hypothesis Testing

1. Present research question and identify hypotheses
2. Describe “Null” distribution
3. Obtain data, calculate relevant “Test Statistic”
4. **Calculate the “P-value”**

- **P-value = likelihood of observing the Test Statistic or something more extreme assuming the Null Hypothesis**

5. Use the P-value to make a conclusion on the research question

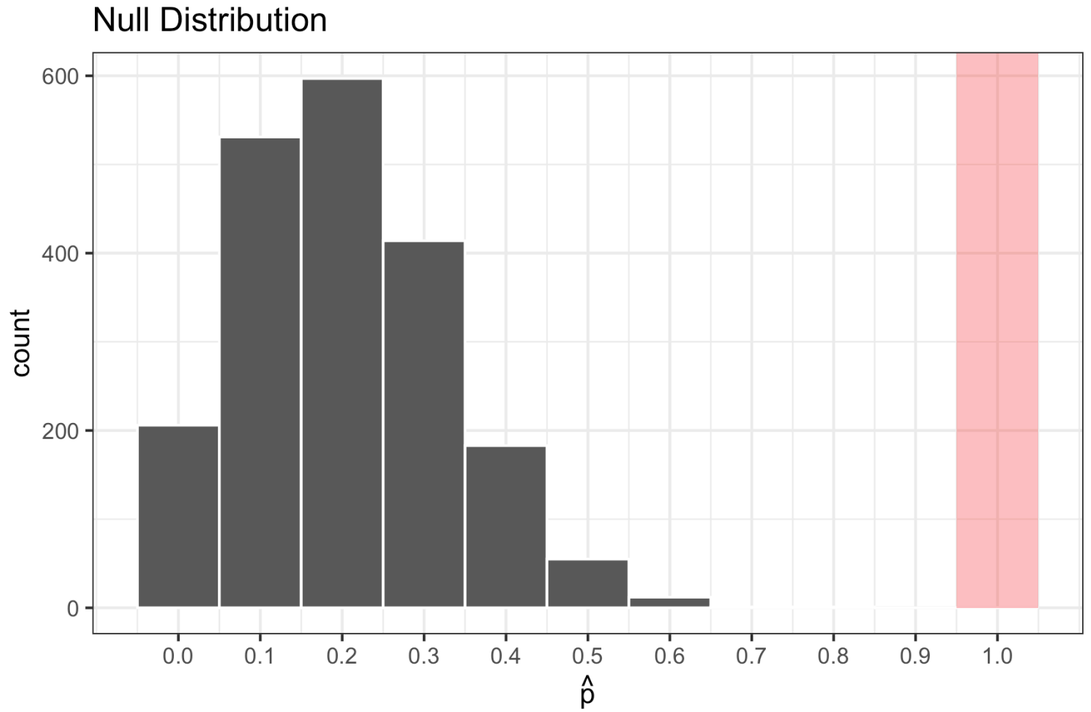
4) Calculate the P-Value

- **Reminder:** Informally,
  - P-value = Probability of what happened in reality (or something more extreme) assuming the hypothesis is true
- More formally,
  - The **p-value** is the probability of observing a sample/test statistic (e.g., $\hat{p}$) at least as favorable to the alternative hypothesis as the current statistic, if the null hypothesis (H_0) were true.
    - e.g., I had $\hat{p} = 1$, so we want the probability of $\hat{p} = 1$ **given** H_0
    - e.g., If I had $\hat{p} = 0.5$, we'd want the probability of $\hat{p} \geq 0.5$ **given** H_0
- The p-value quantifies the strength of evidence against the Null Hypothesis.
  - **Smaller p-values represent stronger evidence against H_0.**
  - $\text{P-value} \approx 0$ means test statistic was unlikely to arise by chance, if the null hypothesis were true.
  - If the p-value is "sufficiently small", we'll reject H_0.

4) Calculate the P-Value

- **Method 1:** Approximate the null distribution using simulation.
  - Then, calculate the proportion of simulated statistics at least as extreme as the test statistic.

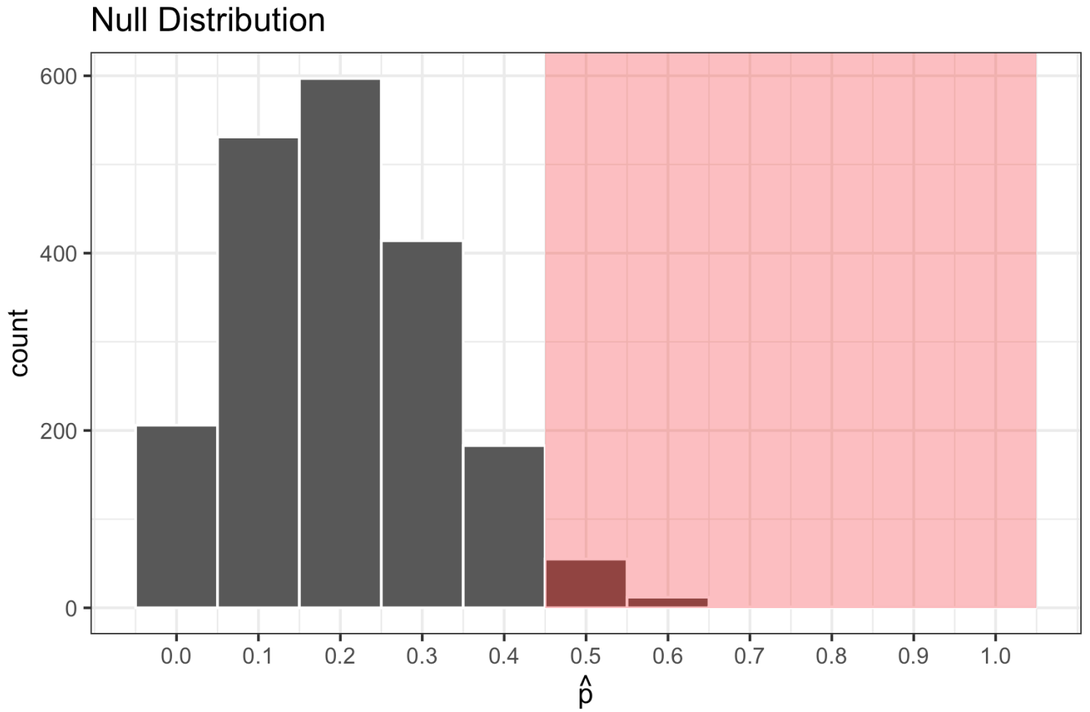
For $\hat{p} = 1$, we get a p-value of:




```
1 mean(null_stats$p_hat==1)
```

```
[1] 0
```

For $\hat{p} = 0.5$, we get a p-value of:




```
1 mean(null_stats$p_hat>=0.5)
```

```
[1] 0.0345
```

4) Calculate the P-Value

- **Method 2:** We use theory-based tools to create the theoretical null distribution.
  - Then use the model to calculate the theoretical probability of observing a sample statistic as extreme as the test statistic.
  - In the case where I get all 10 guesses right, we have already calculated the p-value as:

$$\text{P-value} = (1/5)^{10} \approx 0.0000001$$

Framework for Hypothesis Testing

1. Present research question and identify hypotheses
2. Describe “Null” distribution
3. Obtain data, calculate relevant “Test Statistic”
4. Calculate the “P-value”

- P-value = likelihood of observing the Test Statistic or something more extreme assuming the Null Hypothesis

5. **Use the P-value to make a conclusion on the research question**

5) Making a conclusion using the P-value

Null Hypothesis: $p = 1/5$ **Alternative Hypothesis:** $p > 1/5$

Discuss with Neighbor(s):

1. When $\hat{p} = 1$ (Megan guesses all 10 cards correctly), we found **P-value** ≈ 0. Do we reject the Null Hypothesis under this framework? Why?
2. Hypothetically, a $\hat{p} = 0.5$ gives **P-value** ≈ 0.04. Do we reject the Null now? Why or why not?

Answers:

Next time

- More hypothesis testing!
  - Practice framing research questions in the hypothesis testing framework
  - 1-sided vs 2-sided tests
- Graded midterms and info on revisions