Midterm Review

Ricky Truong

Midterm

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

Logistics and Disclaimer

- This will be half "lecture"/content review (by Ricky) and half hands-on practice (by Sarah)
- We couldn't fit every single detail from the past 6 weeks into 2 hours, so these are the main/important ideas!
- We don't know what the exam looks like!

Before we start, what topics do you want me to spend the most time covering?

Week 2: Data Visualization

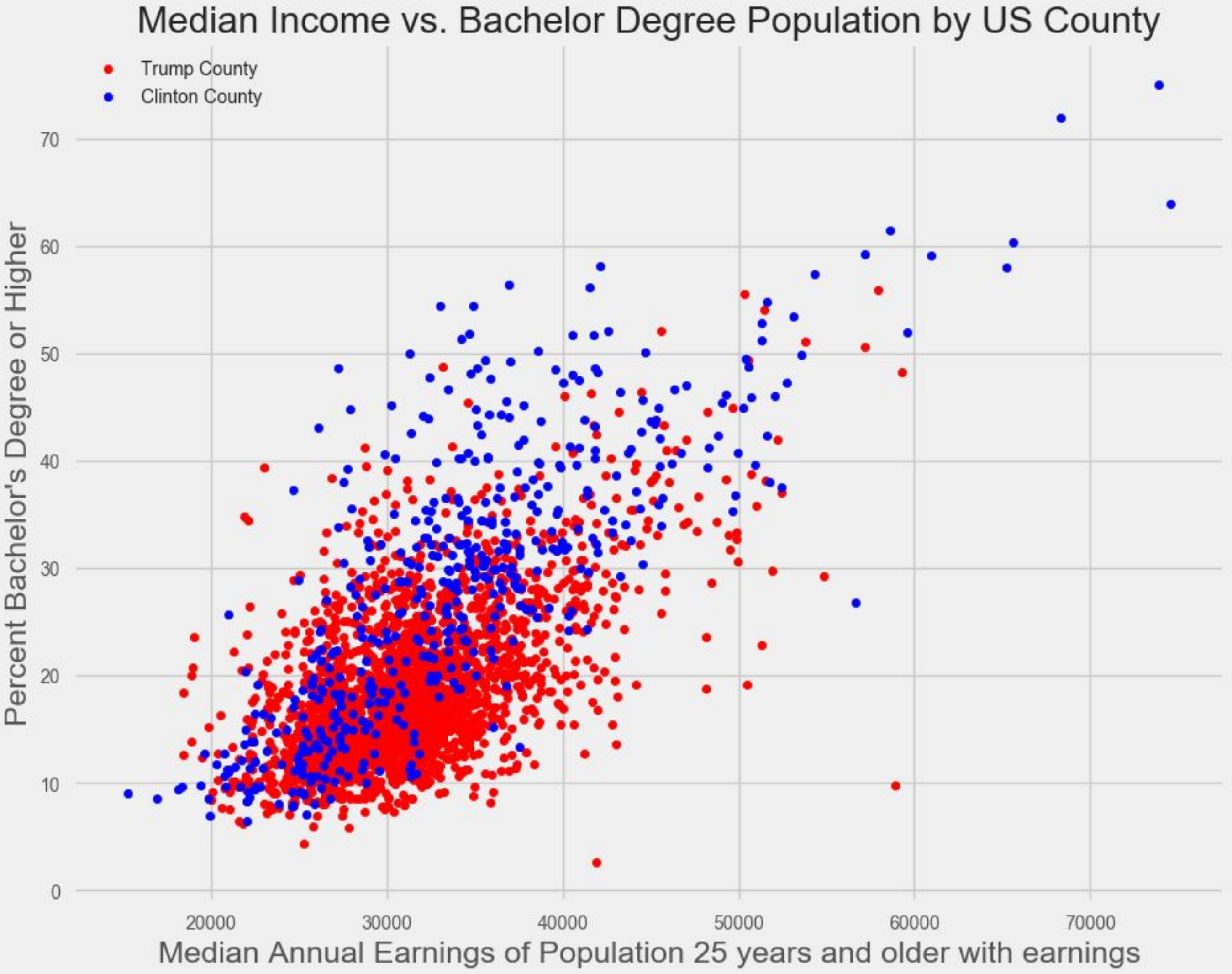
Data Visualization

- We begin the course with **data visualization** (i.e., making graphs to see our data)
- We use the grammar of graphics as vocabulary to describe graphs
- Depending on our **variable(s)**, we need to know when to choose the right graph

Grammar of Graphics

- **Dataset**: "Spreadsheet" containing data/info
- **Geom**: Geometric representation of data (as a shape), such as bars or points
- **Aesthetic**: Visual properties of geoms that mean something in the context of the data
  - Examples include color, scale, x-direction/y-direction

In the following graph, does the visual property of color mean/represent anything? If so, what? How about size?



Median Income vs. Bachelor Degree Population by US County

Question:

In the following graph, does the visual property of color mean/represent anything? If so, what? How about size?

Color represents (the variable of) political leaning. Thus, it is an aesthetic—we say "the variable of political leaning is mapped to the aesthetic of color."

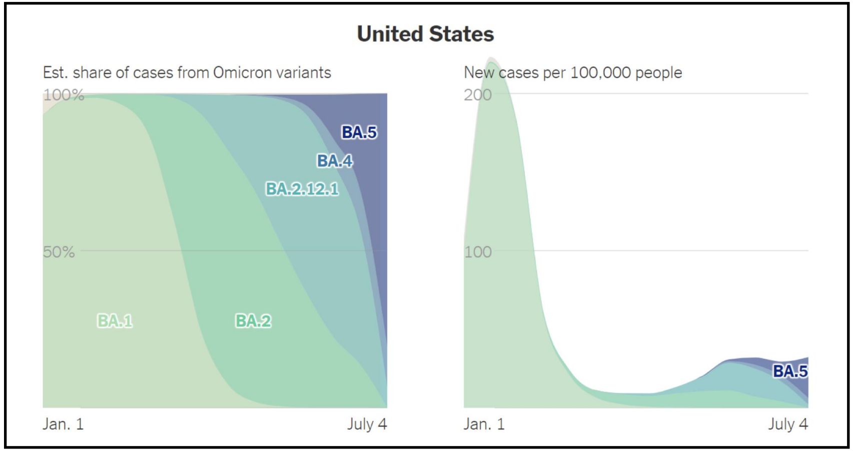
In contrast, we note the size of these dots doesn't mean anything in the context of this graph.

Color Palettes

- **Sequential**: Color progresses from low to high value
- **Diverging**: Color splits in opposite directions, departing from a meaningful middle
- **Qualitative**: Color only distinguishes cases from each other, with no inherent order

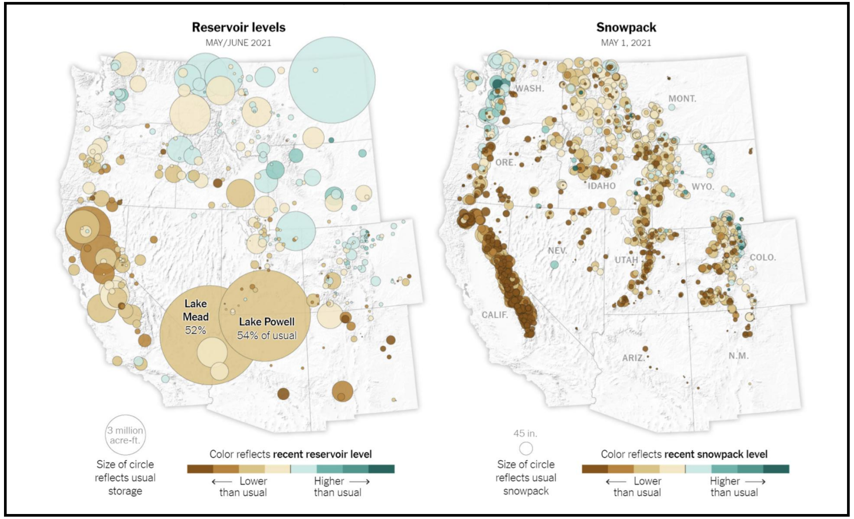
Sequential

Sequential: Color progresses from low to high value



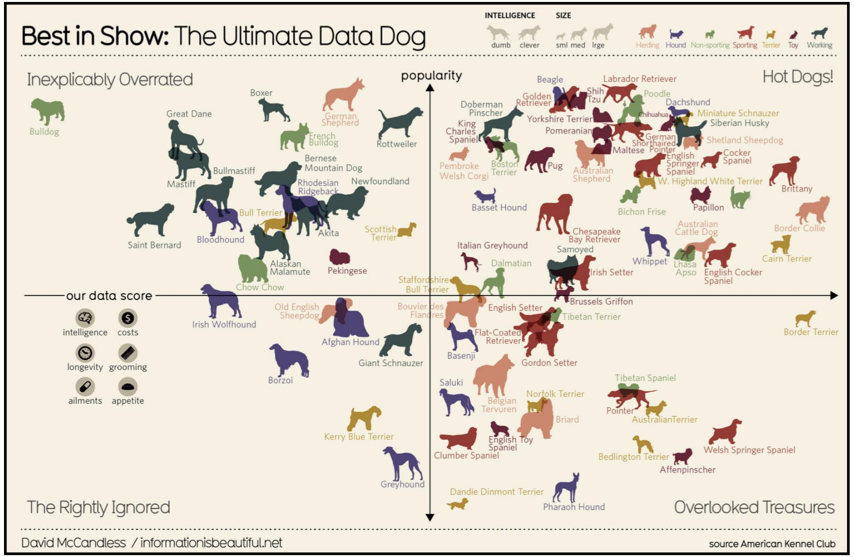
Diverging

Diverging: Color splits in opposite directions, departing from a meaningful middle



Qualitative

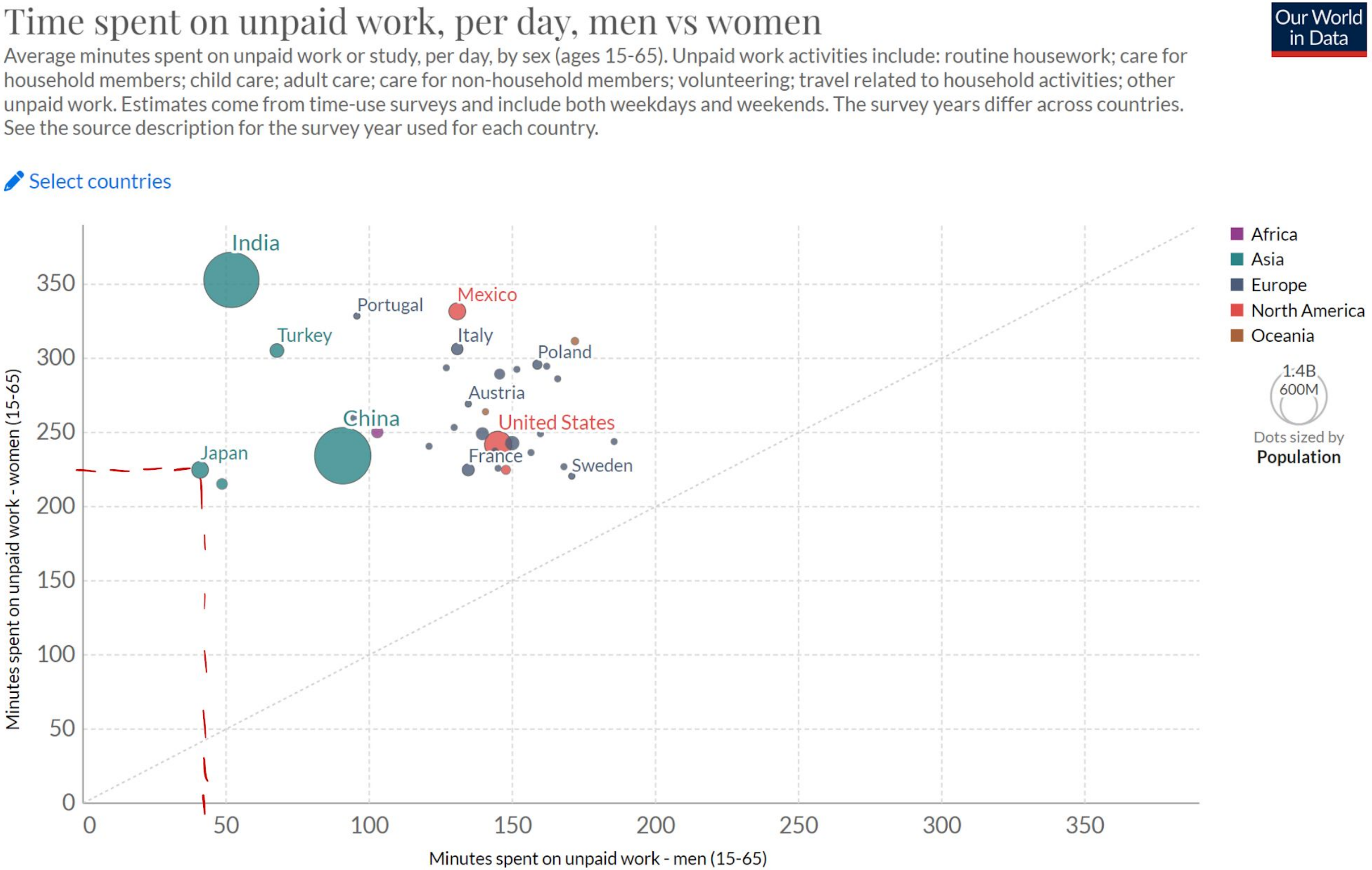
Qualitative: Color only distinguishes cases from each other, with no inherent order



David McCandless / informationisbeautiful.net

source American Kennel Club

What color palette is employed in the following graph?



Time spent on unpaid work, per day, men vs women

Average minutes spent on unpaid work or study, per day, by sex (ages 15-65). Unpaid work activities include: routine housework; care for household members; child care; adult care; care for non-household members; volunteering; travel related to household activities; other unpaid work. Estimates come from time-use surveys and include both weekdays and weekends. The survey years differ across countries. See the source description for the survey year used for each country.

Question:

What color palette is employed in the following graph?

This is using a **qualitative color palette**.

These colors are meant to distinguish points from one another, but they have no order

Europe is a darker shade of blue than Asia, but that doesn't mean Europe is "more/less ___" than Asia.

Types of Variables

- **Numerical variables**: Take on numerical values, which you can measure and "do math"
  - *For example, salary: $100k, $50k, $70k, …*
- **Categorical variables**: Take on values that are labels, which you use to divide into groups
  - *For example, income level: low, middle, high, …*

Explanatory vs. Response

- **Explanatory variable**: Expected cause ("input")
- **Response variable**: Expected result of explanatory variable ("output")
  - *For example, measuring the effect of education level (explanatory) on salary (response)*

I want to see the effect of education level on political party. What are my variables, and what type are they? What should my graph be?

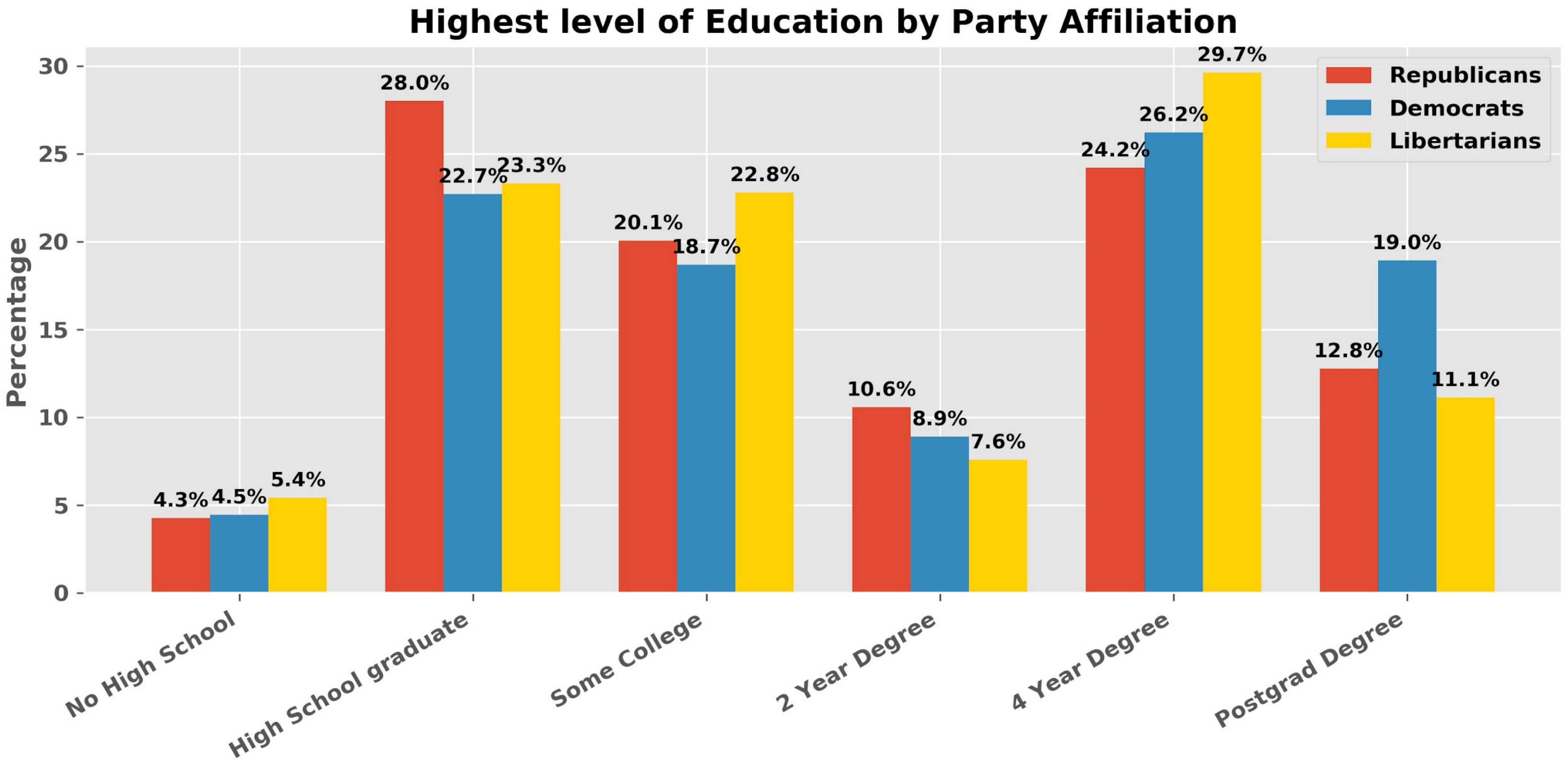
Question:

I want to see the effect of education level on political party. What are my variables, and what type are they? What should my graph be?

My **explanatory variable** is educational level, which is **categorical**.

My **response variable** is political party, which is **categorical**.

By the “choosing the right graph” handout, we would use a **segmented barplot**.

Highest level of Education by Party Affiliation
Percentage
0
5
10
15
20
25
30
Republicans
Democrats
Libertarians
No High School
High School graduate
Some College
2 Year Degree
4 Year Degree
Postgrad Degree
4.3%
4.5%
5.4%
28.0%
22.7%
23.3%
20.1%
18.7%
22.8%
10.6%
8.9%
7.6%
24.2%
26.2%
29.7%
12.8%
19.0%
11.1%

Week 3: Data Wrangling

Data Wrangling

- As our data are often messy, **data wrangling** (i.e., cleaning) is a recurring topic
- Understand the different ways to handle missing values
- Understand the important wrangling functions, which is best shown by examples
  - The really big ones are `mutate()` and `summarize()`
  - By no means is this an exhaustive list! Consider creating your own (if you haven't already)

In my `harvardstudents` dataset, I want to compare how many more/less students there are from the East Coast than from the West Coast. What functions would I use?

Question:

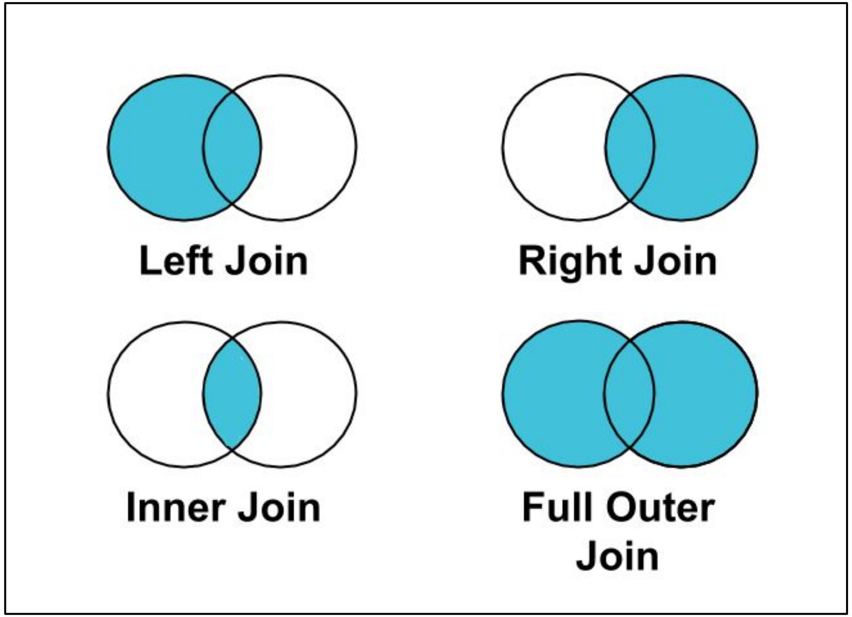
In my `harvardstudents` dataset, I want to compare how many more/less students there are from the East Coast than from the West Coast. What functions would I use?

First, we'd create a new variable, `coast`, using `mutate()` and `case_when()` (i.e., `coast` would be "`east`" in the case when `homestate %in% c("NH", "MA", …)`).

Then, we'd `group_by(coast)` and calculate via `summarize()`. Equivalently, we could use `count()`.

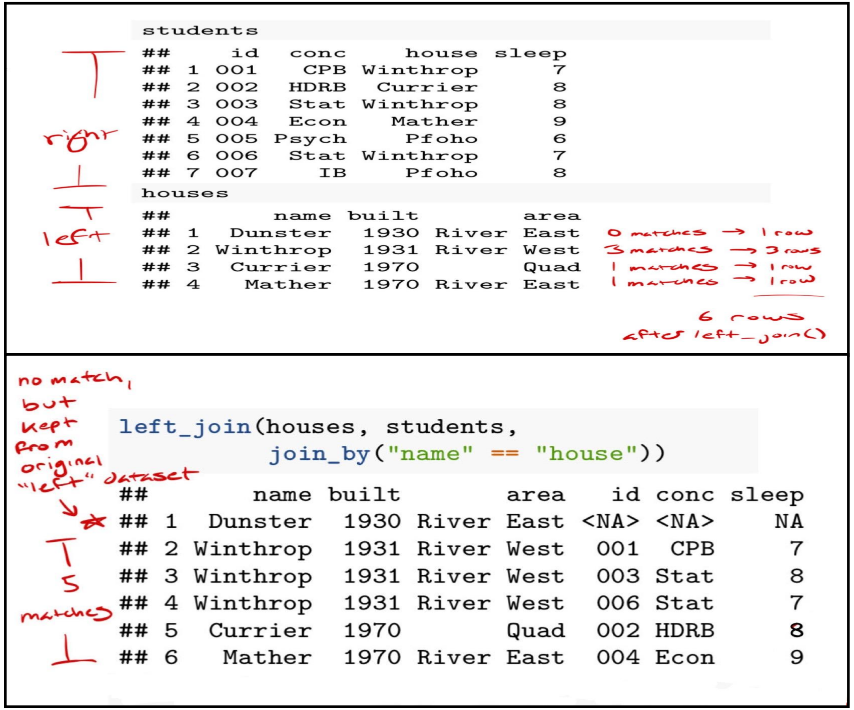
Data Joins

- Use to join **datasets** via a **key** (variable to link the 2 datasets)
- Left, Right, Inner, and Full



Left Join

- `left_join(houses, students, join_by("name" == "house"))`
- Combine 2 datasets via key, keeping all original observations from LEFT-HAND dataset while adding matching observations from RIGHT-HAND dataset



```
students
## id conc house sleep
## 1 001 CPB Winthrop 7
## 2 002 HDRB Currier 8
## 3 003 Stat Winthrop 8
## 4 004 Econ Mather 9
## 5 005 Psych Pfoho 6
## 6 006 Stat Winthrop 7
## 7 007 IB Pfoho 8
houses
## name built area
## 1 Dunster 1930 River East
## 2 Winthrop 1931 River West
## 3 Currier 1970 Quad
## 4 Mather 1970 River East
```

right

left

0 matches → 1 row
3 matches → 3 rows
1 matches → 1 row
1 matches → 1 row

6 rows
after left_join()

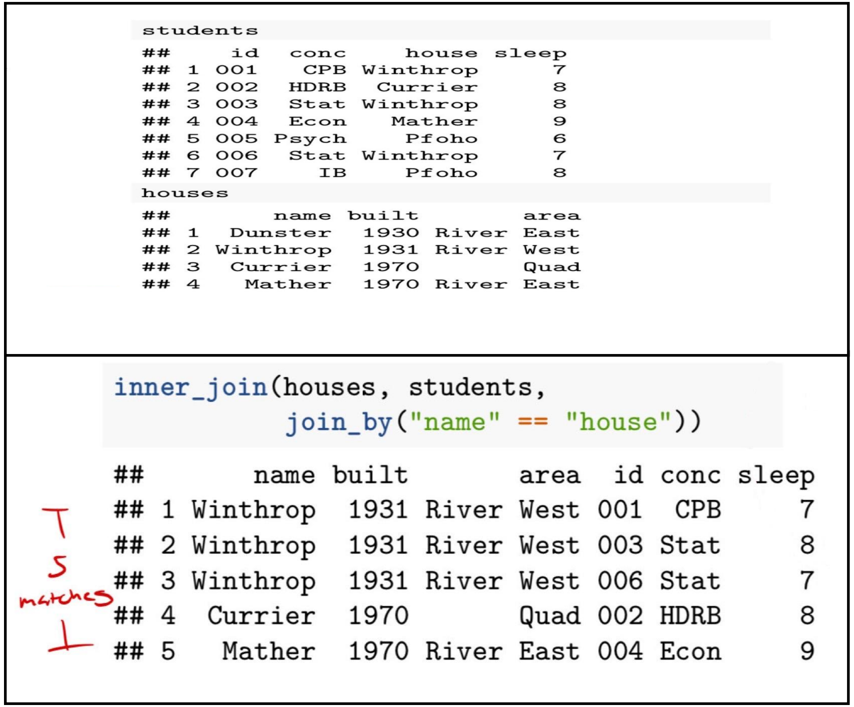
no match, but kept from original "left" dataset

```
left_join(houses, students,
          join_by("name" == "house"))
##        name built        area   id conc sleep
## 1   Dunster  1930 River East <NA> <NA>    NA
## 2  Winthrop  1931 River West  001  CPB     7
## 3  Winthrop  1931 River West  003 Stat     8
## 4  Winthrop  1931 River West  006 Stat     7
## 5   Currier  1970        Quad  002 HDRB     8
## 6    Mather  1970 River East  004 Econ     9
```

5 matches

Inner Join

- `inner_join(houses, students, join_by("name" == "house"))`
- Combine 2 datasets via key, keeping only matching observations between BOTH datasets (most constrained)



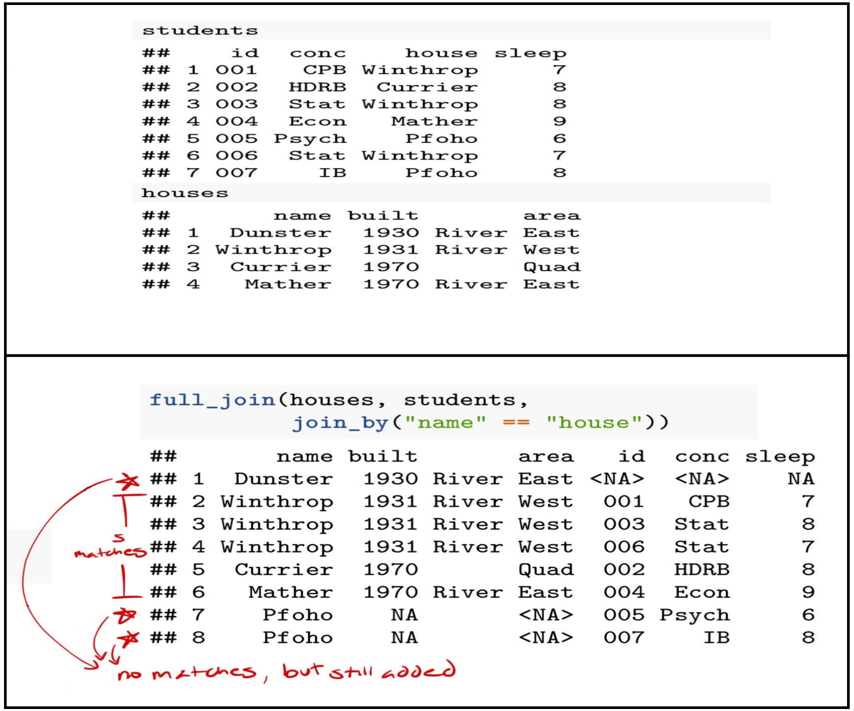
```
students
##      id   conc    house sleep
## 1 001    CPB Winthrop     7
## 2 002   HDRB  Currier     8
## 3 003   Stat Winthrop     8
## 4 004   Econ   Mather     9
## 5 005  Psych    Pfoho     6
## 6 006   Stat Winthrop     7
## 7 007     IB    Pfoho     8
houses
##          name built        area
## 1    Dunster  1930 River East
## 2   Winthrop  1931 River West
## 3    Currier  1970       Quad
## 4     Mather  1970 River East
```

```
inner_join(houses, students,
           join_by("name" == "house"))
```

```
##          name built        area  id conc sleep
## 1 Winthrop   1931 River West 001  CPB     7
## 2 Winthrop   1931 River West 003 Stat     8
## 3 Winthrop   1931 River West 006 Stat     7
## 4  Currier   1970       Quad 002 HDRB     8
## 5   Mather   1970 River East 004 Econ     9
```

Full Join

- `full_join(houses, students, join_by("name" == "house"))`
- Combine 2 datasets via key, keeping all observations between BOTH datasets and putting N/A if an observation didn't have corresponding value for a variable (most expansive)



```
students
##      id   conc     house sleep
## 1 001    CPB Winthrop     7
## 2 002   HDRB  Currier     8
## 3 003   Stat Winthrop     8
## 4 004   Econ   Mather     9
## 5 005  Psych    Pfoho     6
## 6 006   Stat Winthrop     7
## 7 007     IB    Pfoho     8
houses
##         name built        area
## 1   Dunster   1930 River East
## 2 Winthrop    1931 River West
## 3   Currier   1970        Quad
## 4    Mather   1970 River East
```

```
full_join(houses, students,
          join_by("name" == "house"))
```

```
##        name built       area   id  conc sleep
## 1   Dunster  1930 River East <NA>  <NA>    NA
## 2  Winthrop  1931 River West  001   CPB     7
## 3  Winthrop  1931 River West  003  Stat     8
## 4  Winthrop  1931 River West  006  Stat     7
## 5   Currier  1970       Quad  002  HDRB     8
## 6    Mather  1970 River East  004  Econ     9
## 7     Pfoho    NA       <NA>  005 Psych     6
## 8     Pfoho    NA       <NA>  007    IB     8
```

Pipe

- %>%: Takes dataset and "pipes" it as the first argument in the next line
  - The first argument of most wrangling verbs is a dataset
  - This is read as "and then" when reading code aloud
- 'Command' + 'Shift' + 'M'

Pipe: These Are Equivalent Statements

```
mythbusters %>%
    summarize(count =
n())
```

```
summarize(mythbusters,
count = n())
```

Practice Translating: What Does This Code Do (In English)?

```
people %>%
   drop_na(pay) %>%
   filter(gender == "Female", jobtitle ==
"Financial Analyst") %>%
   slice_max(pay, n = 10) %>%
   select(pay, education, name)
```

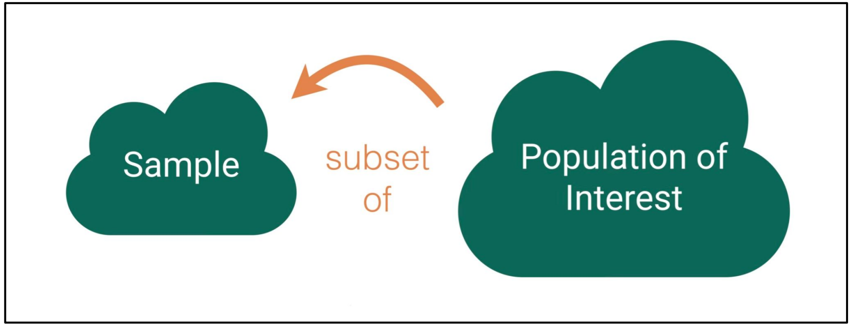
Solution

- We take the **dataset** “people”…
- **And then** we **drop** all observations that have **NA** for the variable “pay”…
- **And then** we **filter** for the gender of “female” **and** the job title of “financial analyst”
- **And then** we **slice** for the observations with the top 10 **maximum** values for the variable “pay”
- **And then** we **select** (to display) the variables “pay,” “education,” and “name”

Week 4: Data Collection

What Is Sampling?

- **Sample**: Subset of **population of interest**, whatever that may be (ideally, it's **representative** of the population)
- **Census**: When there is data for whole population (**everyone is represented**)
  - Often, it's hard to get a census



What Is Bias?

- **Sampling bias**: When **sampled** units are different from **non-sampled** units on the **variable(s) of interest**
  - *Ex: If I ask Harvard students for their screen time via Instagram poll, those who are sampled probably have higher screen times*
- **Nonresponse bias**: When **respondents** are different from the **non-respondents** on the **variable(s) of interest**
  - *Ex: If I ask Harvard students for their screen time, those with higher screen times may be embarrassed and decline to answer*

Observational Study vs. Experiment

- **Experiment**: Researchers directly influence how the data arise
  - Causal relationship can be established with random assignment
- **Observational study**: Researchers only observe and record data without interfering
  - “Correlation does not mean causation”

Principles of an Experiment

- **Control group**: Group of subjects who get **no treatment**
- **Experimental group**: Group that does get **treatment**
- **Random assignment**: Subjects are randomly assigned to either the **control group** or the **experimental group**
- **Confounding variable**: Third variable that is associated with both the **explanatory variable** and **response variable** *(e.g., genetics on sunscreen use and skin cancer)*

Principles of an Experiment

- **Placebo**: Fake treatment to control for **placebo effect**
  - *If given a sugar pill (placebo), someone may start to feel better because they believe it is medicine*
- **Blinding**: When **subjects** don't know the **group assignments (control** vs. **experimental)**
  - *If given a pill, the subject wouldn't know whether it's medicine or sugar/placebo*
- **Double blinding**: When **both subjects and researchers** don't know (not always possible)
  - *All the pills are mixed, so researchers can't tell whether they're giving out medicine or sugar/placebo*

Why can experiments establish causal relationships?

Question:

Why can experiments establish causal relationships?

Due to **random assignment**, those in **control group** should be very similar to those in **experimental group**. Thus, **confounding variables** have been eliminated/minimized.

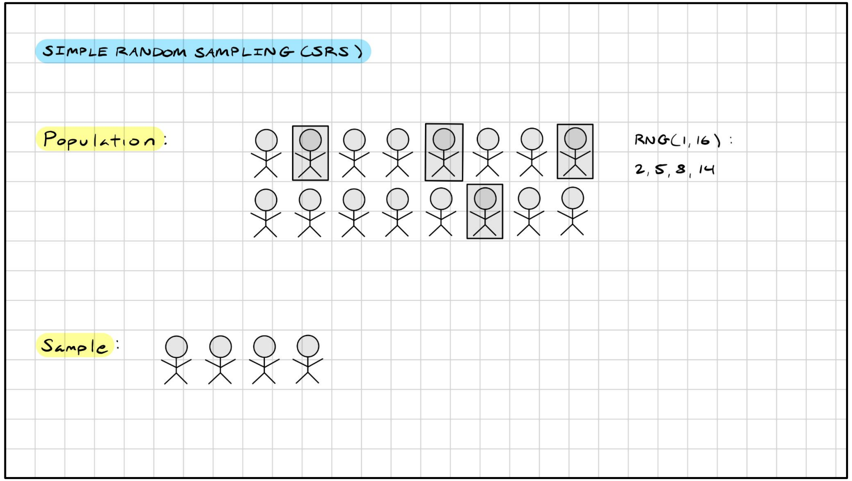
The differences between the two groups after the **experiment** must have been caused by the **treatment/explanatory variable**.

Four Sampling Methods

- There are four main methods for **random** sampling:
  - **Simple random sampling**
  - **Systematic sampling**
  - **Cluster sampling**
  - **Stratified sampling**

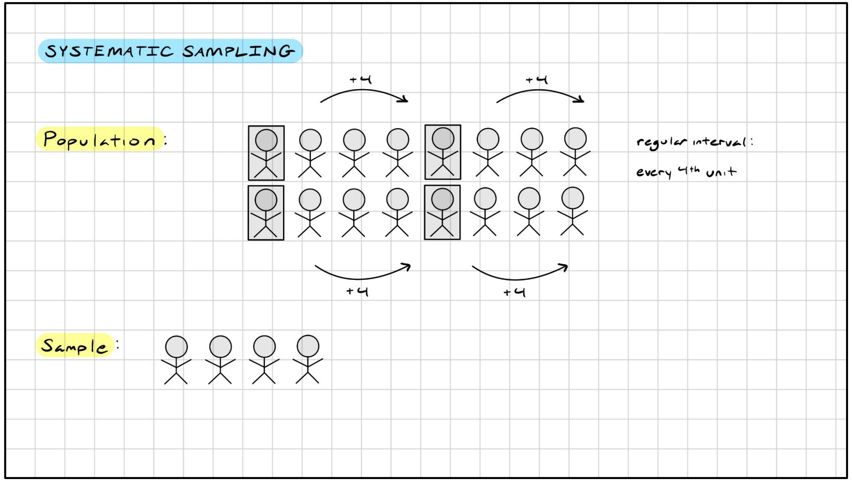
Simple Random Sampling (SRS)

- **Simple random sampling**: Every unit has an equal chance of being selected via **random mechanism** (all units must be listed out in a **sampling frame**)
  - *Ex: To determine smartphone usage within Harvard students, number every student (HUID) and then draw random numbers to determine which ones to sample*



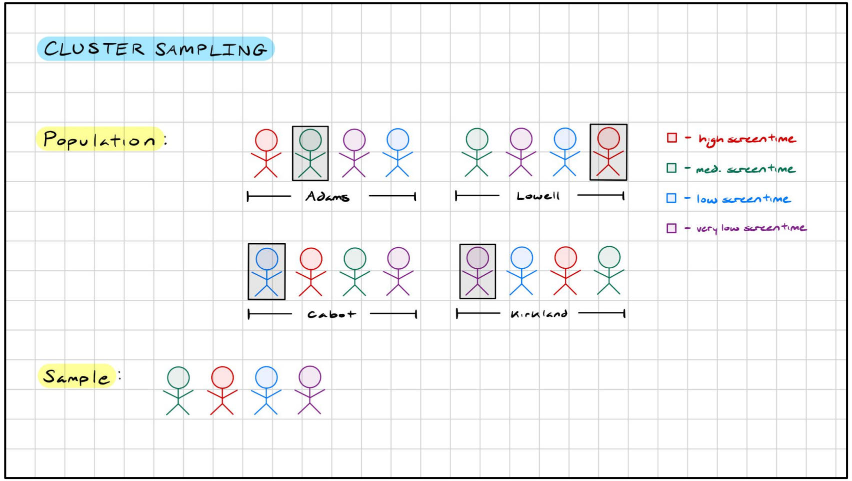
Systematic Sampling

- **Systematic sampling**: Starting point is randomly chosen, and then units are sampled at a **regular interval**
  - *Ex: To determine smartphone usage within Harvard students, number every student (HUID) and then sample every fourth student*



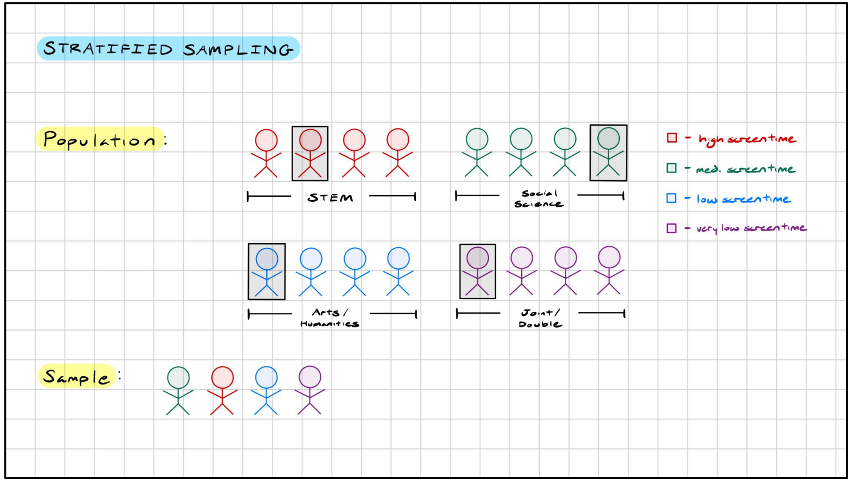
Cluster Sampling

- **Cluster sampling**: Divide **population** into **homogeneous groups/clusters** and take a **random sample** within **SOME** of the **clusters** (to be chosen randomly)
    - *Ex: To determine smartphone usage within Harvard students, sample students within four randomly-selected houses*
    - *Here, houses should be homogeneous (in terms of screen time) because houses are randomly assigned*



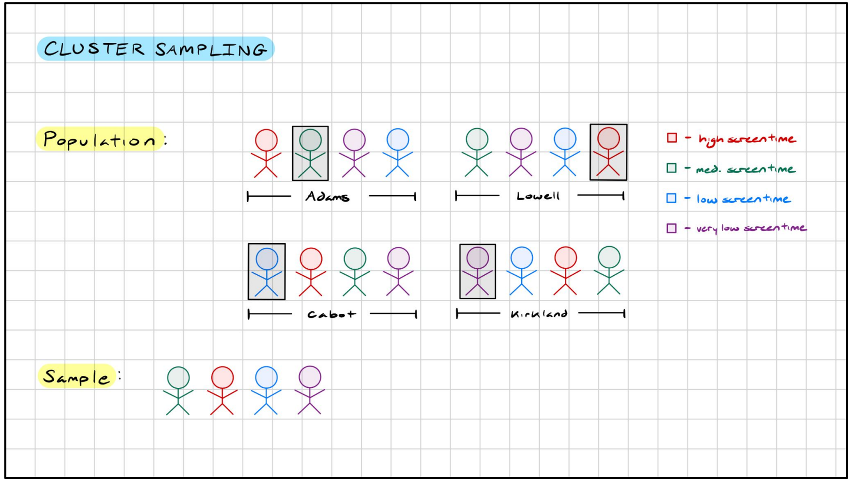
Stratified Random Sampling

- **Stratified random sampling**: Divide **population** into **heterogeneous groups/strata** and take a **random sample** within **EVERY stratum**
  - *Ex: To determine smartphone usage within Harvard students, sample students within each concentration*
  - *Here, concentrations should be heterogeneous (in terms of screen time) because STEM fields require more technology*



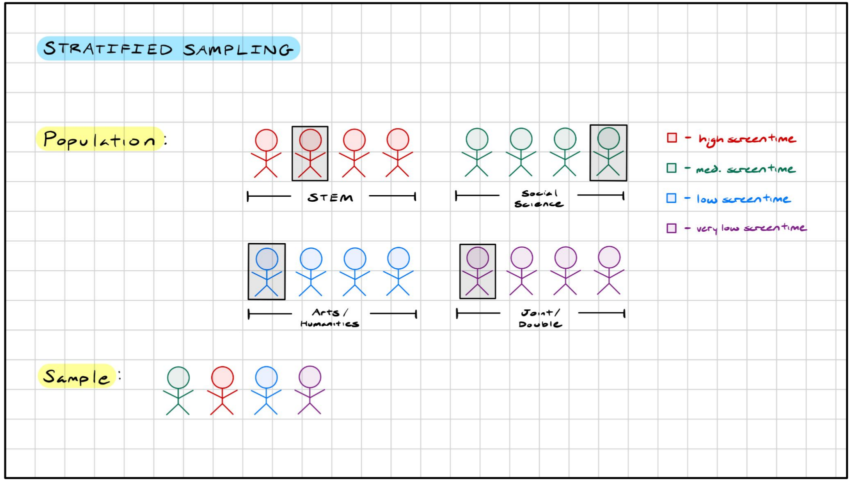
A Clarifying Note...

- "Hetero-" means different, "homo-" means same
- When we say **homogeneous groups**, we mean the GROUPS are homogeneous/similar to EACH OTHER
    - Notice how the PEOPLE within the groups are pretty different from each other, but that's not what we are referring to!



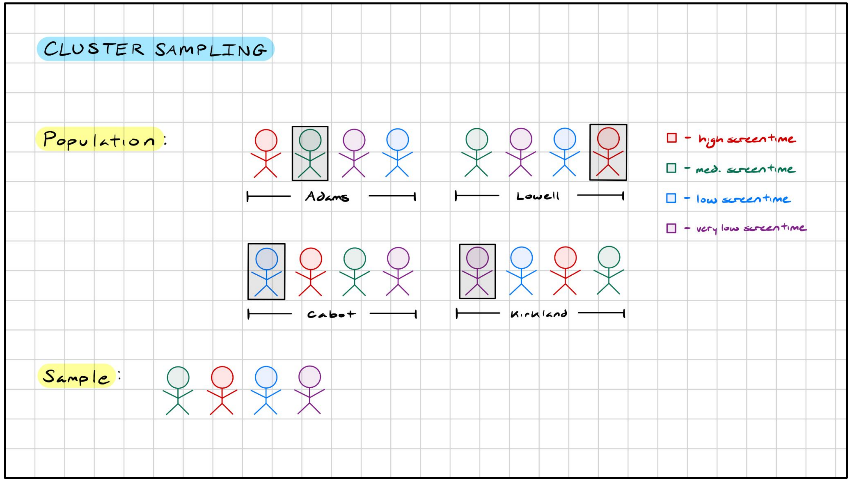
A Clarifying Note…

- Similarly, these are **heterogeneous groups** as in each GROUP is different from the others (in terms of our variable of screen time)
    - For example, the STEM group is primarily high screen time while the arts/humanities is primarily low screen time



And One Last Thing...

- **"Homogeneous in terms of our variable"** is context dependent
- Instead of screen time, let's say our variable of interest is "Hours Spent at the MAC per Week"
    - Before, houses were homogeneous groups (in terms of screen time), so we can appropriately treat them as clusters
    - Now, they aren't. Why?



Intuitively, why do we NOT need to sample every cluster?

Question:

Intuitively, why do we NOT need to sample every cluster?

Clusters are relatively **homogeneous** in terms of our **variable**. For example, houses are similar to each other in terms of screen time. Thus, we don't need to sample Leverett if we already sampled Cabot, Adams, and Pfoho.

Conversely, **strata** are defined to be relatively **heterogeneous**, so all groups must be accounted for.

We divide the population of Harvard students based on their home country and (randomly) sample 10 countries. Am I treating countries as clusters or as strata?

Question:

We divide the population of Harvard students based on their home country and (randomly) sample 10 countries. Am I treating countries as clusters or as strata?

Clusters!

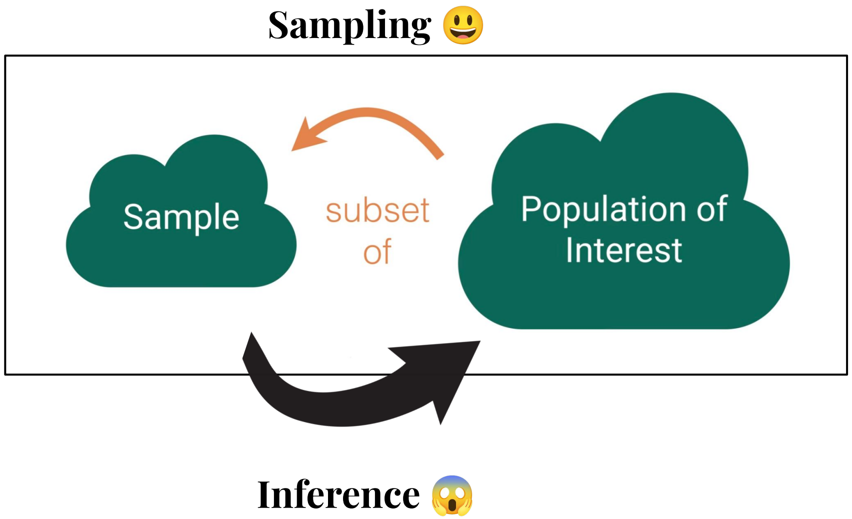
This is because we only sample SOME of the countries (if we treated countries as strata, we would need to sample ALL of them).

Treating countries (or any group) as clusters when it's not appropriate to do so results in non-representative data.

Week 5: Confidence Intervals

Introduction to Inference

- Week 4, we went **from population to sample**. Moving forward, we'll go **from sample to population**!
- Why? Recall the difficulty of obtaining a **census**
- We have data from a **sample** and are interested in concluding something about the **population**



Sampling 😃

Sample

subset of

Population of Interest

Inference 😱

Parameter vs. Statistic

Population parameter:

- Typically **unknown** (what we're interested in finding)
- For **population proportion**, it's denoted as p
- *Ex: Out of all 67 million viewers of the debate, how many believed Harris won? I don't know!*

Sample statistic:

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

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

Other Parameters and Statistics

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

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

Question:

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

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

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

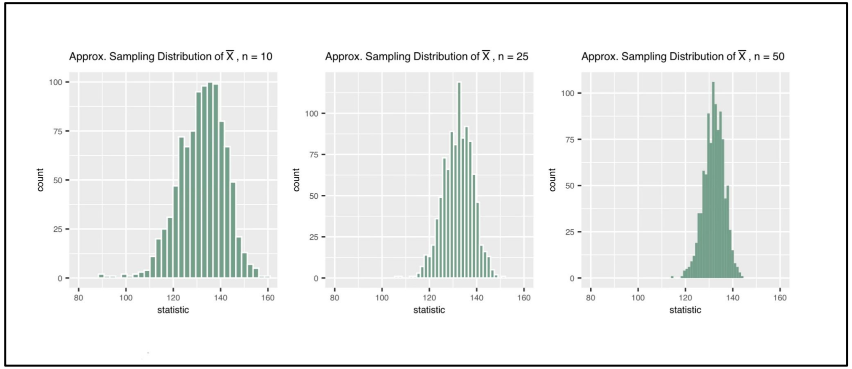
Sampling Variability

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

Sampling Distribution

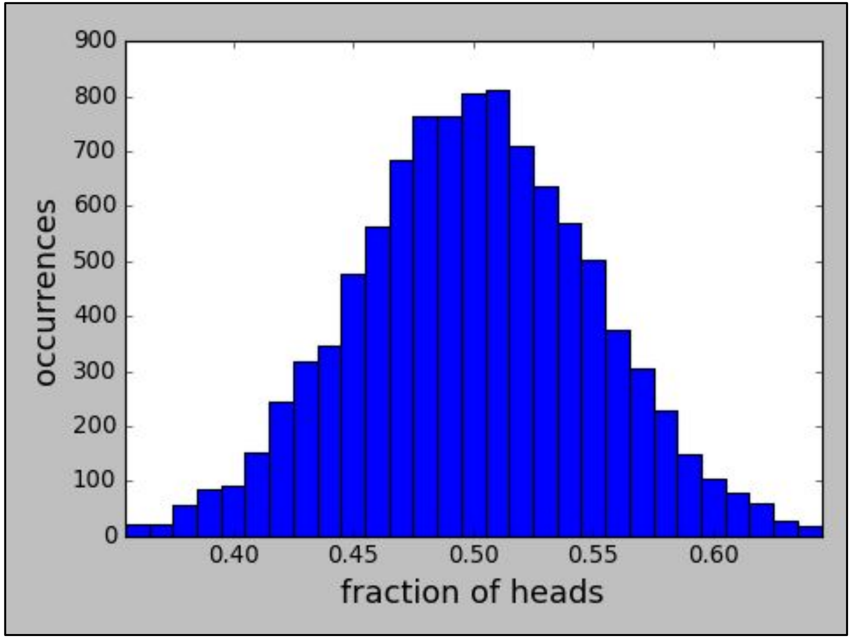
Sampling distribution of a statistic:

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



Coin Flips: An Intuition behind Sampling Distributions

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



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

Question:

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

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

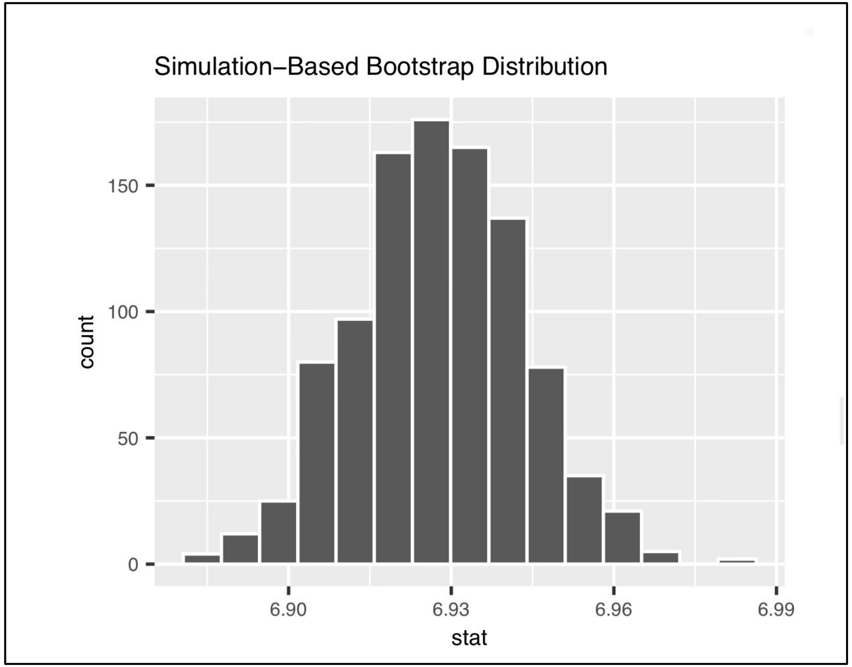
This is not always practical.

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

Bootstrap Distribution

Bootstrap distribution of a sample statistic:

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



Example of Bootstrapping

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

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

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

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

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

and so on...

Sampling Distribution vs. Bootstrap Distribution

Sampling distribution:

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

Bootstrap distribution:

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

Confidence Interval

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

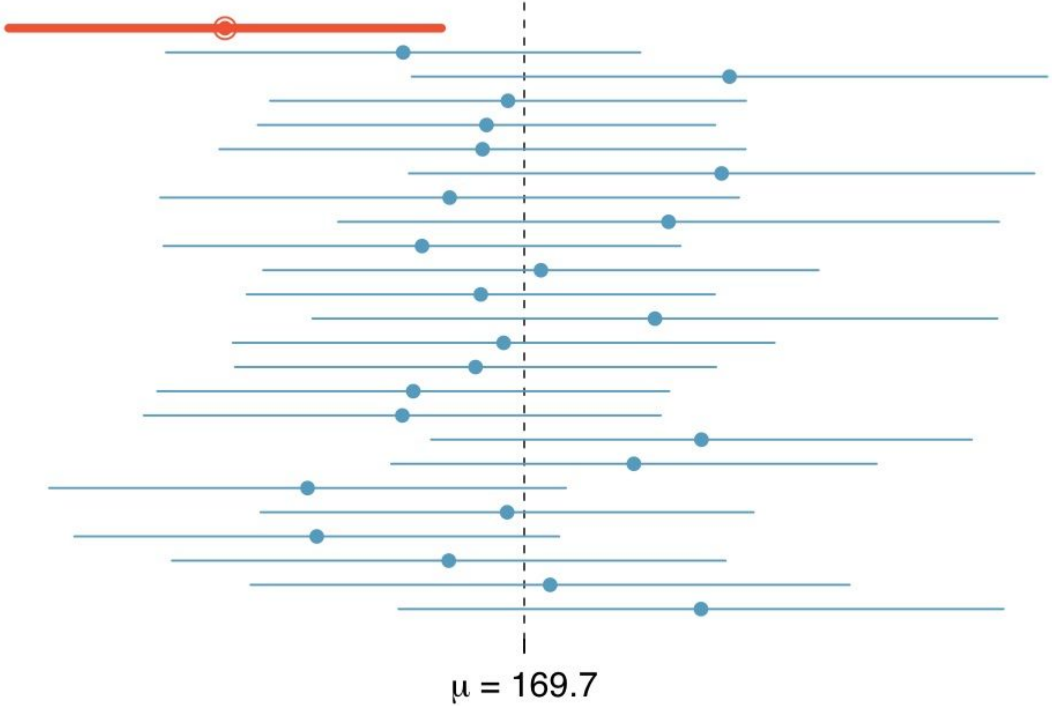
- **SE method: CI = statistic ± z* × (ŜÊ)**
  - z* is critical value, ŜÊ is **standard deviation** of bootstrapped statistics (spread of **bootstrap distribution**)
  - *Ex: 95% CI = statistic ± 1.96(ŜÊ)*
- **Percentile method: CI = the middle (CL)% of the bootstrap distribution**
  - CL = confidence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

Interpreting Confidence Intervals

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

The meaning of confidence. . .

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



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

Question:

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

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

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

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

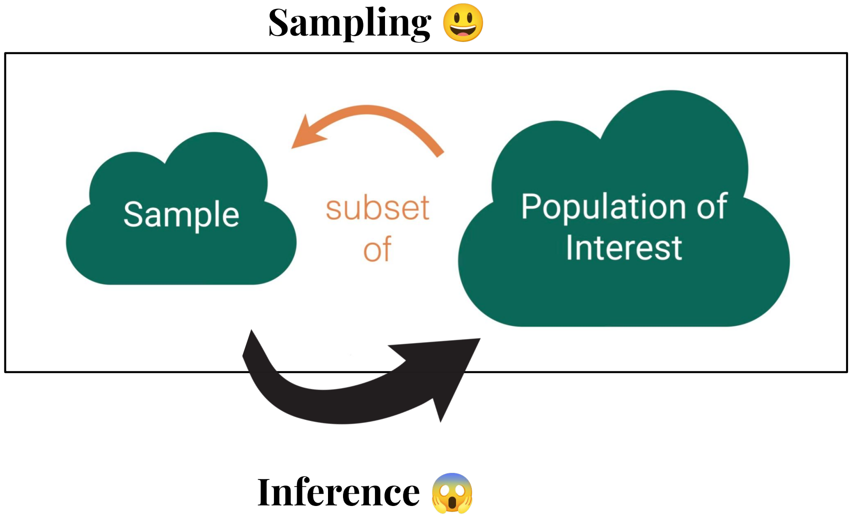
Important Code for Week 5

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

Week 6: Hypothesis Tests

Recap of Inference

- Week 5, we started **inference** with **confidence intervals**
- Now, we'll continue with **hypothesis testing**
- Though complementary, they are different
  - **Confidence intervals** estimate the **parameter**
  - **Hypotheses** test a certain "conjecture" about the **parameter**



A Tale of Two Hypotheses

- **Test statistic**: Numerical summary of the **sample data** (often, but not always, equal to our **observed sample statistic**)
- **Null hypothesis (H_o)**: World where **research conjecture is false** ("no change, status quo")
  - **Null distribution** is **sampling distribution** of **test statistic** assuming **null hypothesis is true**
- **Alternative hypothesis (H_A)**: World where **research conjecture is true**
  - **Alt. distribution** is **sampling distribution** of **test statistic** assuming **alt. hypothesis is true**
- **P-value**: Probability of getting the **observed test statistic OR MORE EXTREME** if **null hypothesis is true**, represented by area under curve of **null distribution**

A Note on Test Statistic

- **Test statistic**: Numerical summary of the **sample data** (often, but not always, equal to our **observed sample statistic**)
  - Last week, our **observed sample statistic** was $\hat{p} = 0.5$ from a random sample of 600 viewers
  - We note this as "**observed**" because if we took a different sample of 600 viewers, we probably would've gotten a different sample statistic, such as $\hat{p} = 0.4$ (i.e., **sampling variability**)
- As the name implies, this is the statistic we'll be using in our **(hypothesis) test**!
  - More on this later, but there will be other test statistics

Can the null and alternative hypotheses both be true?

Question:

Can the null and alternative hypotheses both be true?

No!

The **null hypothesis** and **alternative hypotheses** are mutually exclusive. That is, they CANNOT coexist.

Only 1 can be true. Either this drug works, or it doesn't. Either this coin is rigged, or it's not. And so on.

Essentials of Hypothesis Testing

- **Step 1**: State **hypotheses** (in terms of **population parameter**)
  - Null hypothesis posits the coin is normal. Alternative hypothesis argues it's rigged. H_0: $p = 0.5$, H_A: $p > 0.5$
- **Step 2**: Specify a **significance level**, α (usually $\alpha = 0.05$)
- **Step 3**: Generate **null distribution**
  - If I were to repeatedly sample under the null hypothesis (assuming the coin has a normal 50% chance of heads), what would my sampling distribution look like?
- **Step 4**: Compute **observed test statistic** and **compute p-value**
  - Let's say, with $n = 50$, I observe 30 heads, so $\hat{p} = 0.6$. Under our null distribution, this has a p-value of 0.103.
- **Step 5**: Draw conclusions **in the context of the problem**
  - The probability of seeing 30 or more heads when flipping a fair coin 50 times is equal to 0.103. Since our p-value is high ($0.103 > 0.05$), we fail to reject the null hypothesis. There is little evidence the coin is rigged.

The “P-Value Formula”

- **“If {null hypothesis} were true, then the probability of observing {test statistic} or {more extreme} would be {p-value}.”**
  - This is “interpreting the p-value”
- **“Because {p-value} is a {high/low} probability compared to {alpha}, we reject {reject/fail to reject} the null hypothesis.”**
  - This is “drawing a relevant conclusion”

If I want to see whether or not the majority of Harvard students agree with a new bill, what should my hypotheses be (in terms of my pop. parameters)?

Other Parameters and Statistics

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

Question:

If I want to see whether or not the majority of Harvard students agree with a new bill, what should my hypotheses be (in terms of my pop. parameters)?

We have a **binary categorical response variable** (fraction of students that agree). This is a **one-tailed proportion**.

H_o: p = ½ (There is no majority)

H_A: p > ½ (The majority agree)

If I want to see if Harvard students get less sleep than other college students, what should my hypotheses be (in terms of pop. parameters)?

Question:

If I want to see if Harvard students get less sleep than other college students, what should my hypotheses be (in terms of pop. parameters)?

We have a **binary categorical explanatory variable** (Harvard or not) and **numerical response variable** (hours of sleep). This is a **one-tailed difference of means.**

H$_0$: $\mu_{\text{Harvard}} - \mu_{\text{Other}} = 0$ (Harvard students get same amount of sleep)

H$_A$: $\mu_{\text{Harvard}} - \mu_{\text{Other}} < 0$ (Harvard students get less sleep)

If I observe a difference of means of -2.7 hours (and a p-value of 0.003), what is an interpretation of the p-value and a conclusion? Assume $\alpha = 5\%$.

Question:

If I observe a difference of means of -2.7 hours (and a p-value of 0.003), what is an interpretation of the p-value and a conclusion? Assume α = 5%.

Using the p-value formula...

If **there was no difference in mean hours of sleep between Harvard and non-Harvard students**, then the probability of observing our **test statistic, a difference of -2.7 hours**, or **less** would be 0.3%.

Because 0.3% is a **low** probability (0.3% < 5%), we **reject** the null hypothesis.

Decisions, Decisions

- There are 4 potential outcomes of a **hypothesis test** (shown below), depending on what we do and what's actually true
- **α** - Probability of Type I Error (rejecting H_0 when it's true)
- **β** - Probability of Type II Error (failing to reject H_0 when H_A is true)
  - As **α** decreases, **β** increases (but they DON'T add up to 1)
- **Power**: Probability of rejecting H_0 when H_A is true (best outcome 😁)
  - **Power = 1 - β**

	We Reject H_0	**We Fail to Reject H_0**
H_0 is true	Type I Error	Correct Decision 🙂
H_A is true	Correct Decision 😁	Type II Error

If we reject the null hypothesis, is it possible we committed a Type I Error? A Type II Error?

Question:

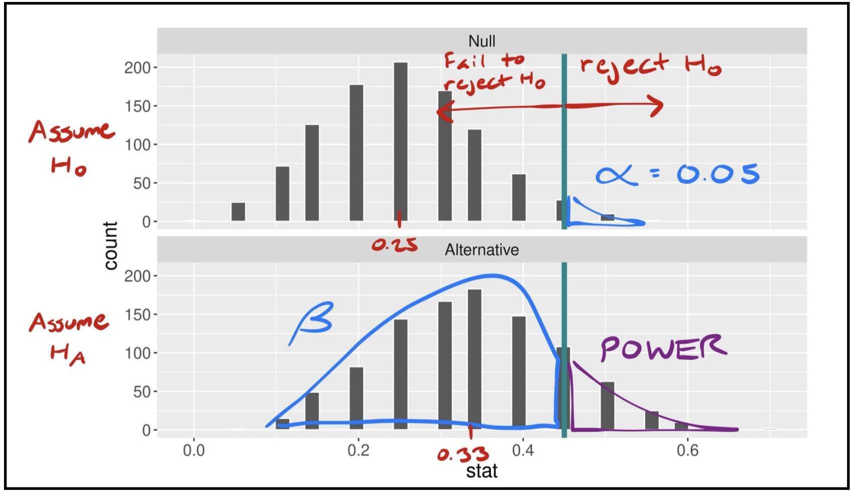
If we reject the null hypothesis, is it possible we committed a Type I Error? A Type II Error?

I remember Type I Error as a "delusional scientist" and Type II Error as a "missed opportunity."

If we reject the null hypothesis, there's a possibility we committed a Type I Error but no possibility we committed a Type II Error (by definition, this would require FAILING to reject the null hypothesis).

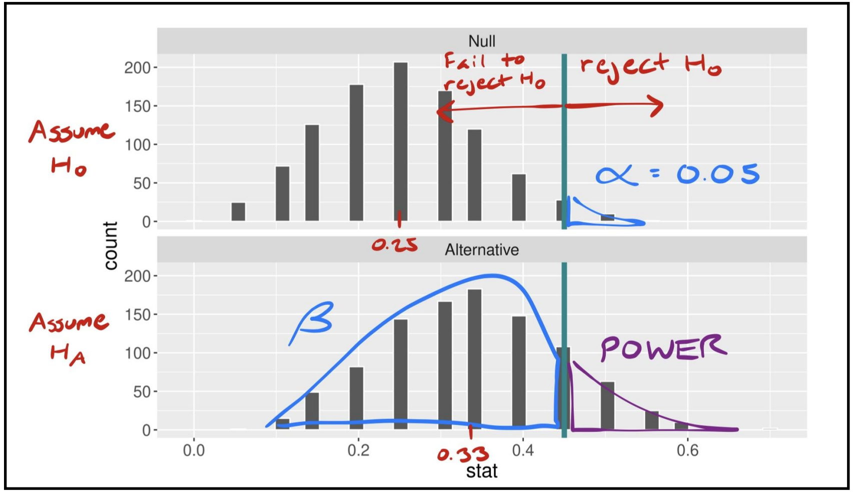
More on Power…

- **Power**: Assuming $\mathbf{H_A}$, what is the probability we reject $\mathbf{H_o}$?
- Think of **power** as a thought experiment—it helps us better understand **hypothesis testing**
  - In real life, we don't know if H_A is true… or where it's centered at!
  - There is an infinite number of alternative distributions that could exist… let's pick just one



Intuition behind Power

- **Power**: Assuming $\mathbf{H_A}$, what is the probability we reject $\mathbf{H_o}$?
  - Given $\mathbf{H_A}$ is true, we look at the **alternative distribution** (which, now, is the true state of the world)
  - The **alpha level** is the probability of rejecting $\mathbf{H_o}$ in the **null distribution**
    - The **critical region** (to the right of **α**) is where we reject $\mathbf{H_o}$
  - Thus, in the **alternative distribution**, the region to the right of the **alpha level** is **power**

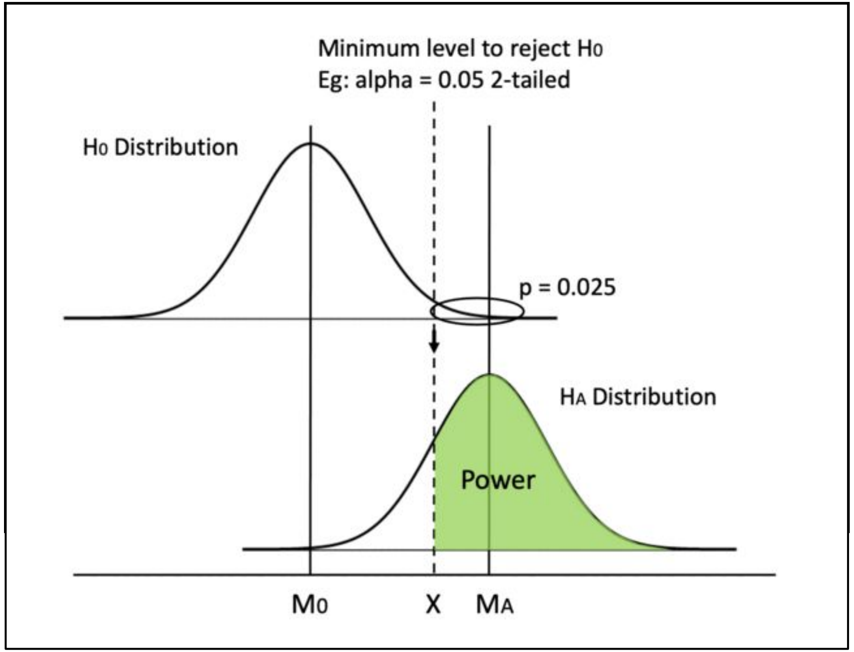


Example: Baseball

- An avid baseball player has been a **0.250** career hitter, but, magically, he improves to be a **0.333** hitter
- He wants a raise, but he has to convince his manager he genuinely improved
- The manager offers to examine his performance in **20 trials**
- **H_0: $p = 0.250$, H_A: $p > 0.250$ ($p \overset{?}{=} 0.333$)**
  - Because the **alternative hypothesis** is **$p > 0.250$**, there is an infinite amount of **alternative distributions** that could exist… specifically, I'm interested in the one centered at 0.333
- He wants his test to be "powerful"
  - When **$\alpha = 0.05$**, he needs to get **9 or more hits** to get a small enough **p-value** to reject **H_0**
  - Unfortunately, at **$\alpha = 0.05$**, the **power of this test** is **0.211** (only a 21% probability of being in the best outcome), so how can we improve the **power of this test**?

How to Increase Power: Increase Alpha

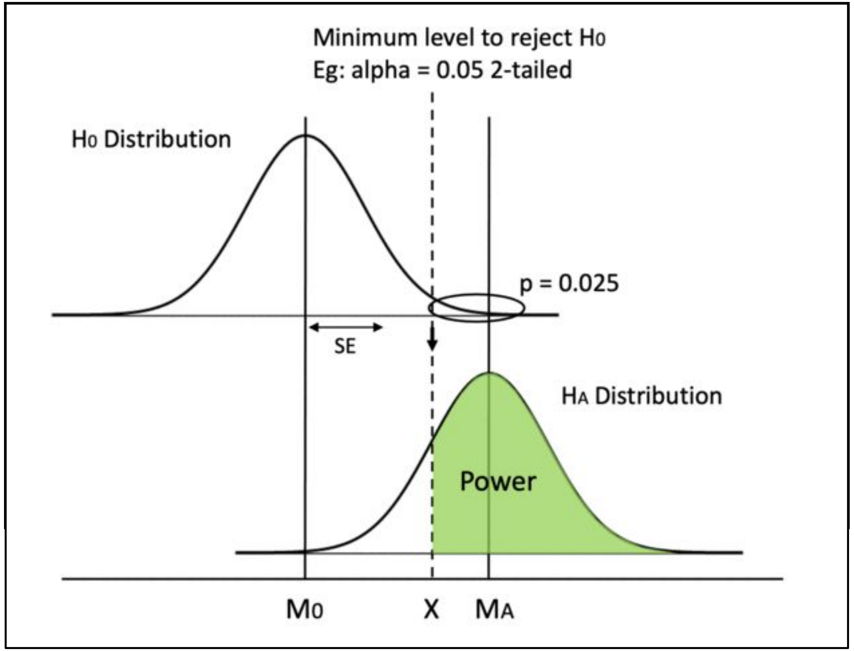
- This makes it easier to reject $\mathbf{H_o}$
- Also, this "shifts" the **critical line** to the left, leading to more area in the "**power region**" of the **alternative distribution**
- Intuitively, we now have a higher probability of rejecting $\mathbf{H_o}$, and **power** is probability of rejecting $\mathbf{H_o}$ when $\mathbf{H_A}$ is true



https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

How to Increase Power: Increase Sample Size

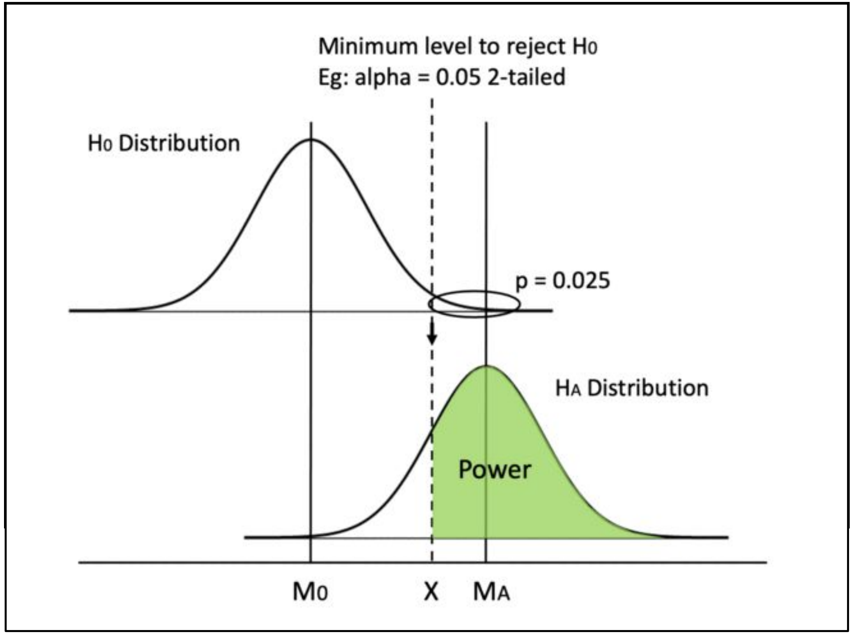
- This decreases **spread** of **histograms**, leading to less overlap between **null distribution** and **alternative distribution**



https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

How to Increase Power: Increase Effect Size

- Effect Size: Difference between **true value of parameter** and **null value**
- This makes it easier for us to notice a difference
- Also, this "shifts" the **center of the alternative distribution** to the right, leading to more area in the "**power region**"



https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

More on Effect Size…

- **Statistical Significance ≠ Practical Significance**
- Let's say, magically, he actually improved to **0.400**
- Now, the **effect size (0.400 - 0.250)** is larger than before
  - Intuitively, it should now be more noticeable if he actually improved from before, so our **hypothesis test** is more "powerful"

What are the 3 ways to increase the power of your hypothesis test?

Question:

What are the 3 ways to increase the power of your hypothesis test?

Increasing alpha, increasing sample size, and increasing effect size.

I encourage you memorize these methods (along with their "intuitive" explanation)!

What is the problem with increasing the alpha level?

Question:

What is the problem with increasing the alpha level?

Though increasing the **alpha level** leads to higher **power**, it also leads to more **false positives** (a higher probability of a **Type I Error**).

There are a lot of trade offs, so these important choices depend on the context of the study.

Important Code for Week 6

https://drive.google.com/file/d/1SD1xhBFjU2KxpbW_cXUNsOSrw74HCcr7/view?usp=drive_link

Questions?

Problem Solving Strategies and Common Mistakes

First, Load All Relevant Libraries

- `library(tidyverse)`
- `library(infer)`
- `library(ggplot2)`
- `library(gglm)`
- `library(moderndive)`
- `library(dplyr)`
- `library(broom)`
- `library(knitr)`
- There might be more I’m forgetting… it doesn’t hurt to load more than you need!

When Should I Know to Calculate Power?

- **Hint 1**: The problem is about a hypothesis test
  - *Ex: "Consider a scenario where at least 55% of voters must approve"*
  - *Here, we're interested in the population proportion of voters*
- **Hint 2**: The problem gives you a SPECIFIC value for the alternative hypothesis (in addition to a null value)
  - *Ex: "If 60% of U.S. adults actually think marijanua should be legal…"*
  - $H_0: p = 55\%,\ H_A: p > 55\%\ (p \overset{?}{=} 60\%)$
- **Hint 3**: You want to "test" something about your hypothesis test (e.g., if there is a sufficient sample size)
  - *Ex: "Would n = 400 be a reasonable sample size to demonstrate, with a one-sided test, that more than 55% of U.S. adults are in favor of legalization?"*

Can a Type I (or Type II) Error Occur?

- Recall the **definitions** and the **table of outcomes**
- **Type I Error**: Rejecting **H_o when it's actually true** (delusional scientist)
  - This can only occur if we reject the null hypothesis (i.e., our p-value is small)
- **Type II Error**: Failing to reject **H_o when H_A is actually true** (missed opportunity)
  - This can only occur if we FAIL to reject the null hypothesis (i.e., our p-value is large)

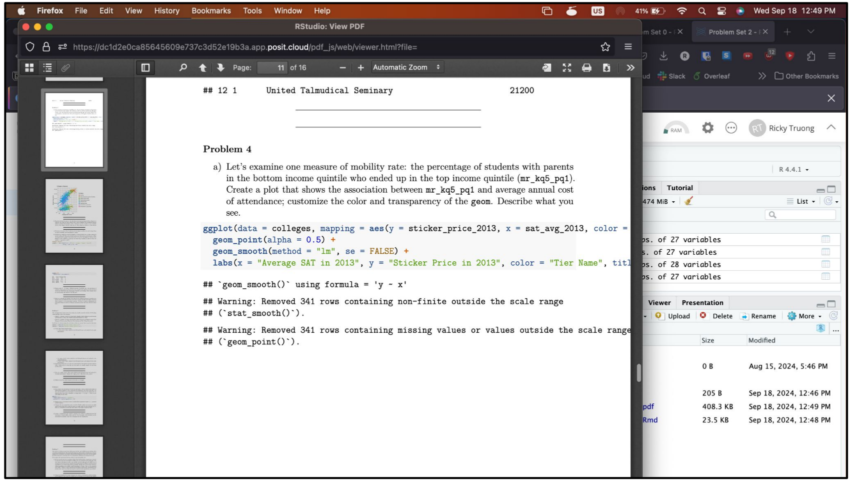
	We Reject H_o	**We Fail to Reject H_o**
H_o is true	Type I Error	Correct Decision 🙂
H_A is true	Correct Decision 😁	Type II Error

Debugging Code: Comment Out, Partial Credit

- To debug code, consider commenting out (#) the possibly-problematic lines
    - If the code runs without the line, you know it’s the problem
    - R will often tell you which line(s) are causing an issue
- Do NOT delete all your code! You may get partial credit even if your code doesn’t run
    - Either set eval = FALSE or comment out the code
    - To comment out, highlight a line and hit “Command” + “Shift” + “C”

Related, Your Code Should Be Readable!

- Make sure your code isn't running off the screen in your PDF
  - If the grader can't read your code, you might get points off
- Hit "Return" to start a new line
  - Best to do this after commas (,) and plus signs (+)



```
## 12 1      United Talmudical Seminary                    21200
```

Problem 4

a) Let's examine one measure of mobility rate: the percentage of students with parents in the bottom income quintile who ended up in the top income quintile (`mr_kq5_pq1`). Create a plot that shows the association between `mr_kq5_pq1` and average annual cost of attendance; customize the color and transparency of the `geom`. Describe what you see.

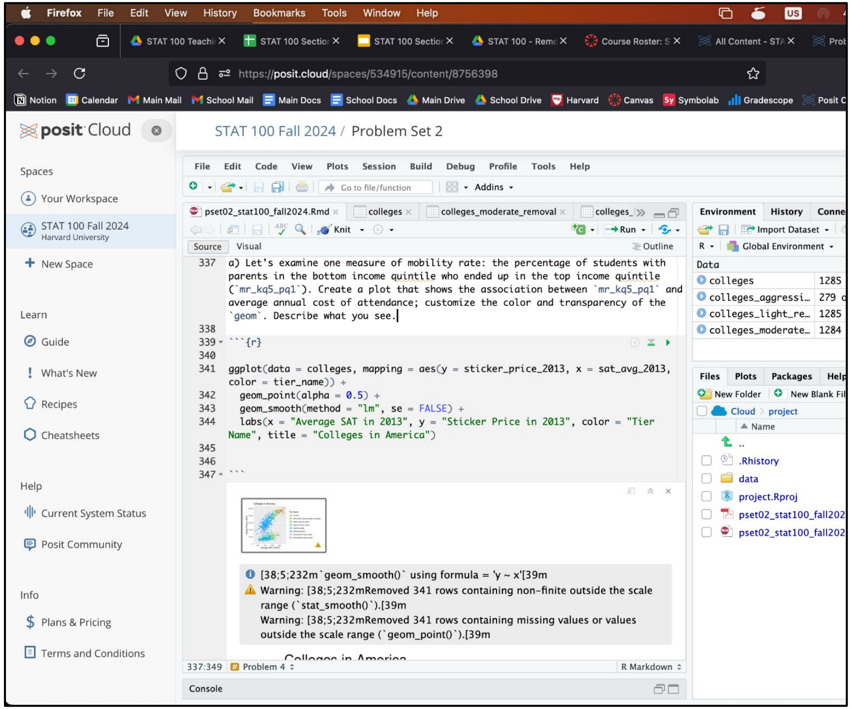
```
ggplot(data = colleges, mapping = aes(y = sticker_price_2013, x = sat_avg_2013, color =
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Average SAT in 2013", y = "Sticker Price in 2013", color = "Tier Name", titl
```

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 341 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 341 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

Messy Code



a) Let's examine one measure of mobility rate: the percentage of students with parents in the bottom income quintile who ended up in the top income quintile (`mr_kq5_pq1`). Create a plot that shows the association between `mr_kq5_pq1` and average annual cost of attendance; customize the color and transparency of the `geom`. Describe what you see.

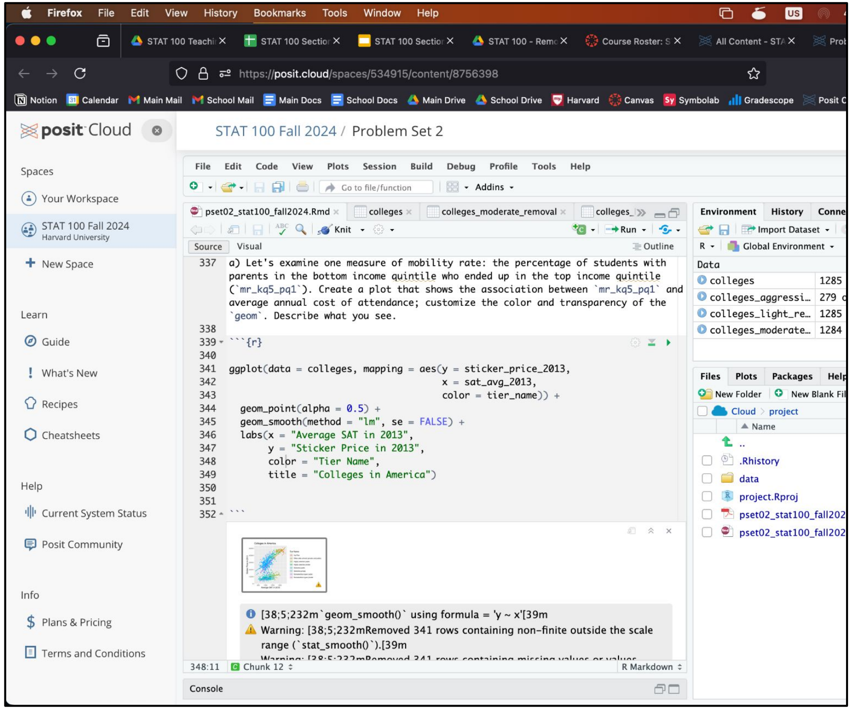
~~~
```{r}

ggplot(data = colleges, mapping = aes(y = sticker_price_2013, x = sat_avg_2013,
color = tier_name)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Average SAT in 2013", y = "Sticker Price in 2013", color = "Tier
Name", title = "Colleges in America")


```
~~~

```
[38;5;232m`geom_smooth()` using formula = 'y ~ x'[39m
Warning: [38;5;232mRemoved 341 rows containing non-finite outside the scale
range (`stat_smooth()`).[39m
Warning: [38;5;232mRemoved 341 rows containing missing values or values
outside the scale range (`geom_point()`).[39m
```

Clean Code



a) Let's examine one measure of mobility rate: the percentage of students with parents in the bottom income quintile who ended up in the top income quintile (`mr_kq5_pq1`). Create a plot that shows the association between `mr_kq5_pq1` and average annual cost of attendance; customize the color and transparency of the `geom`. Describe what you see.

~~~
```{r}

ggplot(data = colleges, mapping = aes(y = sticker_price_2013,
                                      x = sat_avg_2013,
                                      color = tier_name)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Average SAT in 2013",
       y = "Sticker Price in 2013",
       color = "Tier Name",
       title = "Colleges in America")


```
~~~

```
[38;5;232m`geom_smooth()` using formula = 'y ~ x'[39m
Warning: [38;5;232mRemoved 341 rows containing non-finite outside the scale
range (`stat_smooth()`).[39m
```

Tips for Oral Exam

Set a Timer!

- This is probably the best thing you can do for the Oral Exam
- 10 minutes goes by quickly, so use your time responsibly
- In general, try to spend around 3 minutes per question
  - There will be 3 questions, and each can have multiple parts

Don’t Feel the Need to “Ramble”

- Say what you need to say to answer the question
  - Nothing less, nothing more
- If you feel like your answer is enough, move on
- If you have time at the end, you can go back to any questions to elaborate (or even change your answer)

Let’s get some practice coding!