Review Session

Ricky Truong

Disclaimer

- This was designed to focus on reviewing important concepts, so there won't be many "practice problems"
  - There's also no way I can summarize an entire semester in an hour!
  - All mistakes in this slideshow are entirely my own
- A **Practice Session** will be held by Jude tomorrow (**Thu, 12/05 from 1:30-3:30 PM in SC Hall E**)
  - Hopefully, these two sessions will be complementary
- I do not know what the exam will look like

Big Ideas from STAT 100

Data Visualization

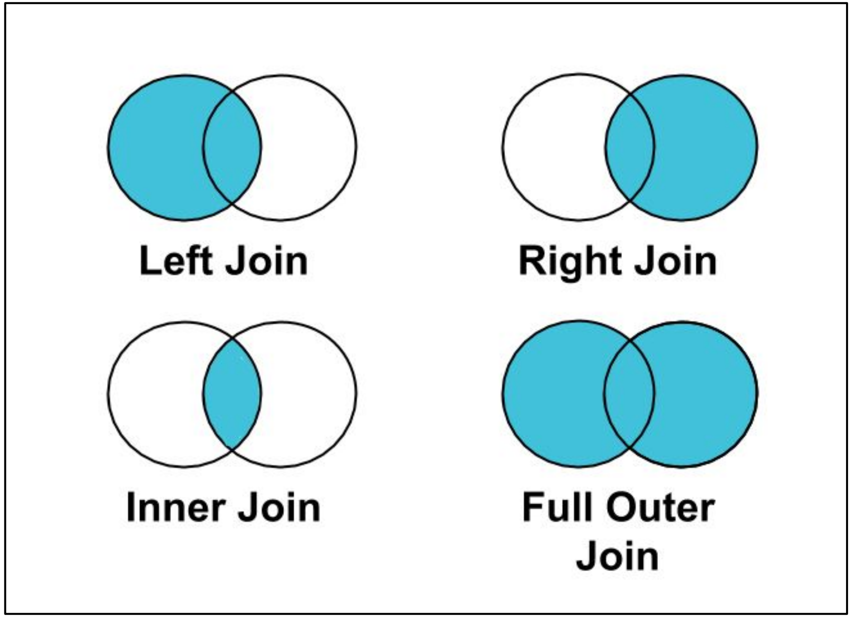
- We begin the course with **data visualization** (i.e., making graphs)
- We use the grammar of graphics as vocabulary to describe graphs
- Depending on our **variable(s)**, we need to know when to choose the right graph

Data Wrangling

- As our data is 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)

Data Joins

- **Data joins** are used to join datasets via a key (variable to link the 2 datasets)
  - The 4 types are **left join**, **right join**, **inner join**, and **full join**
- This might show up, but I doubt it'll be a big part
  - To be safe, I suggest having some notes to reference



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

right

```
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
```

left

```
houses
##         name built        area
## 1    Dunster  1930 River East
## 2  Winthrop  1931 River West
## 3    Currier  1970        Quad
## 4     Mather  1970 River East
```

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

5 matches

```
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
```

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
```

5 matches

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
```

5 matches

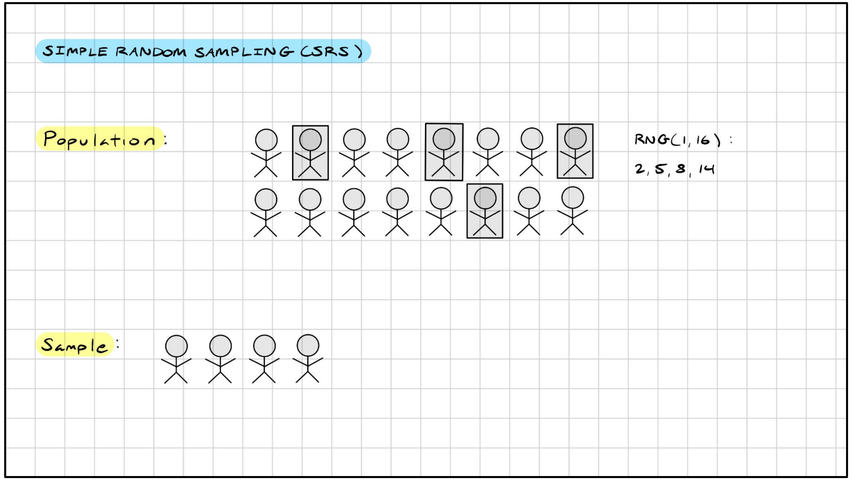
no matches, but still added

Four Sampling Methods

- There are four main methods for **random sampling**
  - **Simple random sampling**
  - **Systematic sampling**
  - **Cluster sampling**
  - **Stratified sampling**
- Again, this might show up, but it most likely won't be a big part of the exam
  - I'll go over the next 4 slides quickly, but on your own time (or during the exam), you can read them more carefully

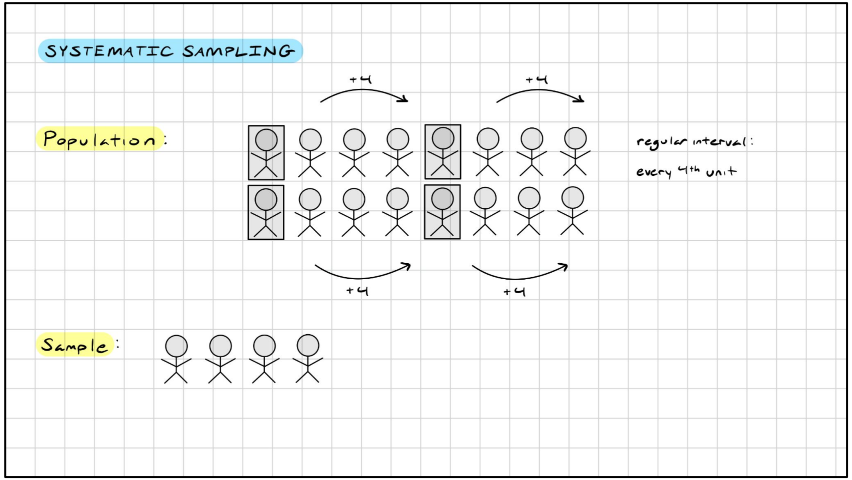
Simple Random Sampling (SRS)

- **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*



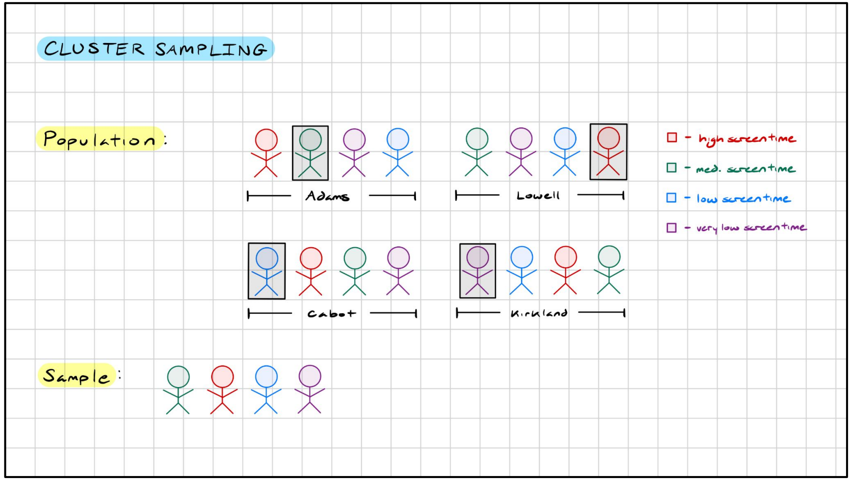
Systematic Sampling

- **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*



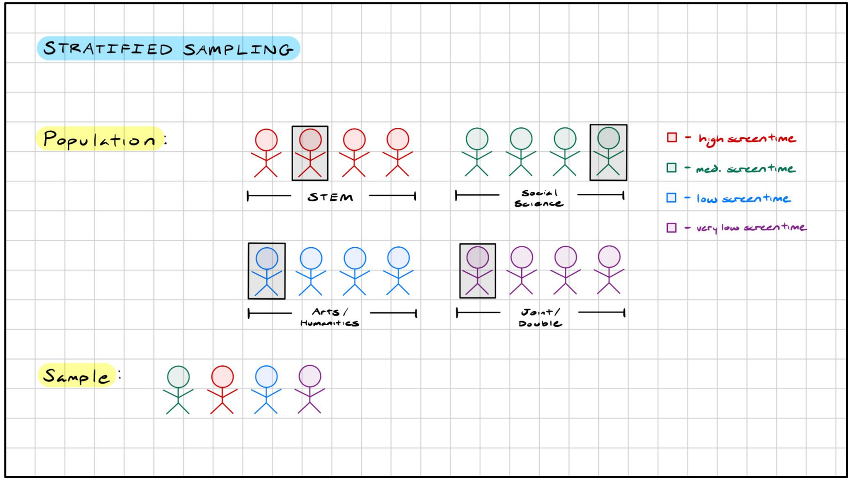
Cluster Sampling

- **Cluster sampling**: Divide **population** into **homogeneous groups/clusters** 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*



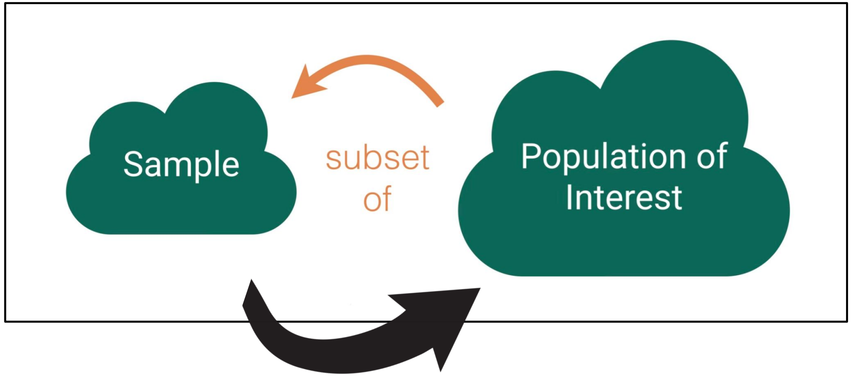
Recapping Our Probability Toolkit

- **Union**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - For **disjoint** events, $P(A \cup B) = P(A) + P(B)$ because $P(A \cap B) = 0$
- **Intersection**: $P(A \cap B) = P(A)\, P(B \mid A) = P(B)\, P(A \mid B)$
  - For **independent** events, $P(A \cap B) = P(A)\, P(B)$ because $P(A \mid B) = P(A)$
- **Complement Rule**: $P(A) = 1 - P(A^C)$, $P(A \mid B) = 1 - P(A^C \mid B)$
  - Use when you see "**at least**" (e.g., "Find the probability of rolling a 5+ at least once in 3 rolls")
- **Def. of Conditional Probability**: $P(A \mid B) = P(A \cap B) / P(B)$
- **Bayes' Rule**: $P(A \mid B) = P(B \mid A)\, P(A) / P(B)$
- **LOTP**: $P(A) = P(A \mid B)\, P(B) + P(A \mid B^C)\, P(B^C)$
  - Use for **wishful thinking** (e.g., "I really wish I knew which factory the cone came from")
- In general with probability, **start by defining events**

What Is Inference and Why Do We Care?

- With **inference**, we go **from sample to population**
  - Recall the difficulty of obtaining a **census**
- We have data from a **sample** and are interested in concluding something about the **population**
  - **Confidence intervals** estimate the **parameter**
  - **Hypotheses** test a certain "conjecture" about the **parameter**

Sampling 😃



Inference 😱

Parameter vs. Statistic

Population parameter:

- Typically **unknown** (what we're interested in finding)
- For **population proportion**, it's denoted as **p**
  - This is for **binary categorical variables**
- *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)

Examples of Parameters and Statistics

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

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 **difference of means**.

$\mathbf{H_0}$: $\mu_{Harvard} - \mu_{Other} = 0$ (Harvard students get same amount of sleep)

$\mathbf{H_A}$: $\mu_{Harvard} - \mu_{Other} < 0$ (Harvard students get less sleep)

A (Brief) Rundown on Hypothesis Testing

- **Test statistic**: Numerical summary of the **sample data** (often, but not always, equal to our **observed sample statistic**)
- **Null hypothesis (H_0)**: 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**

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 α = 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 Confidence Interval "Formula"

- **"We are {confidence level}% confident that the true {population parameter} lies between {lower bound} and {upper bound}."**

The P-Value "Formula"

- **"If {<u>null hypothesis</u>} were true, then the probability of observing {<u>test statistic</u>} or {<u>more extreme</u>} would be {<u>p-value</u>}."**
- **"Because {<u>p-value</u>} is a {<u>high/low</u>} probability compared to {<u>alpha level</u>}, we reject {<u>reject/fail to reject</u>} the null hypothesis."**

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 $\alpha = 5\%$.

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.

Statistical Power

- 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:

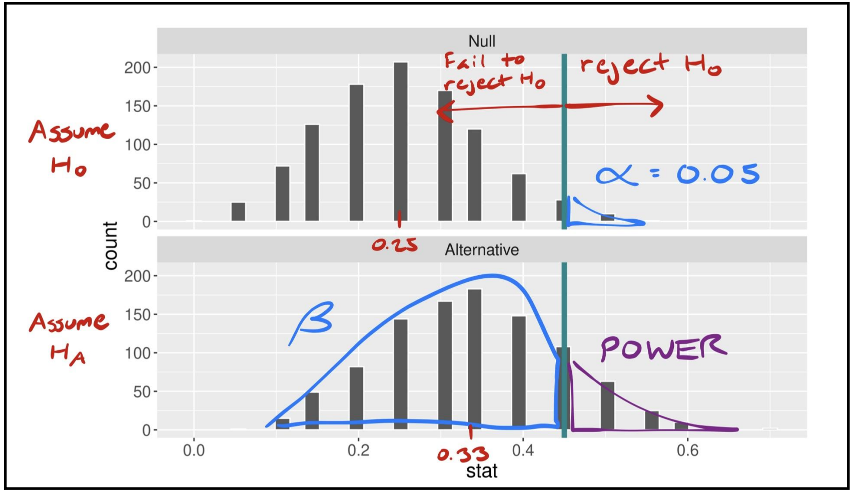
If we reject the null hypothesis, is it possible we committed a Type I Error? A Type II Error?

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).

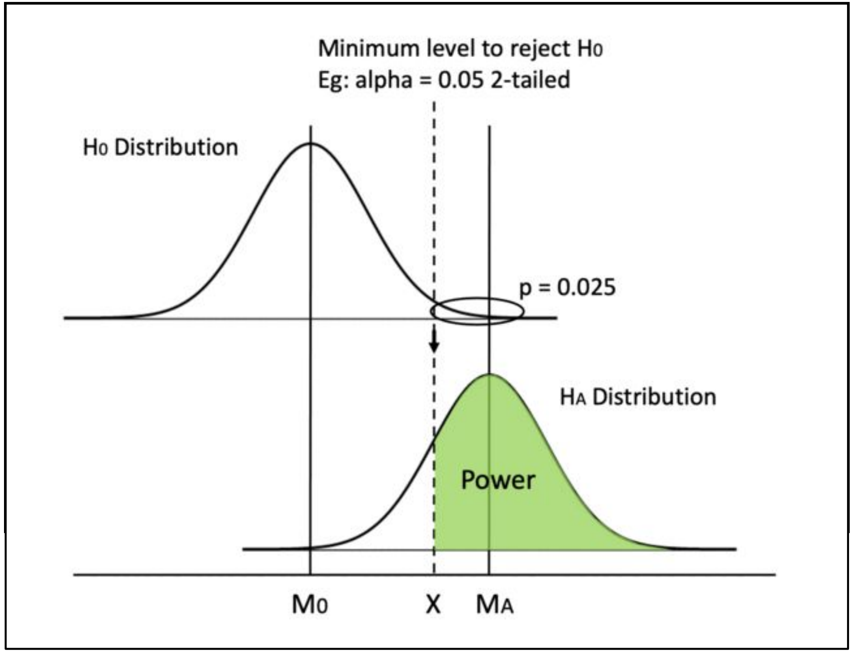
Intuition behind Power

- **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**



How to Increase Power: Increase Alpha

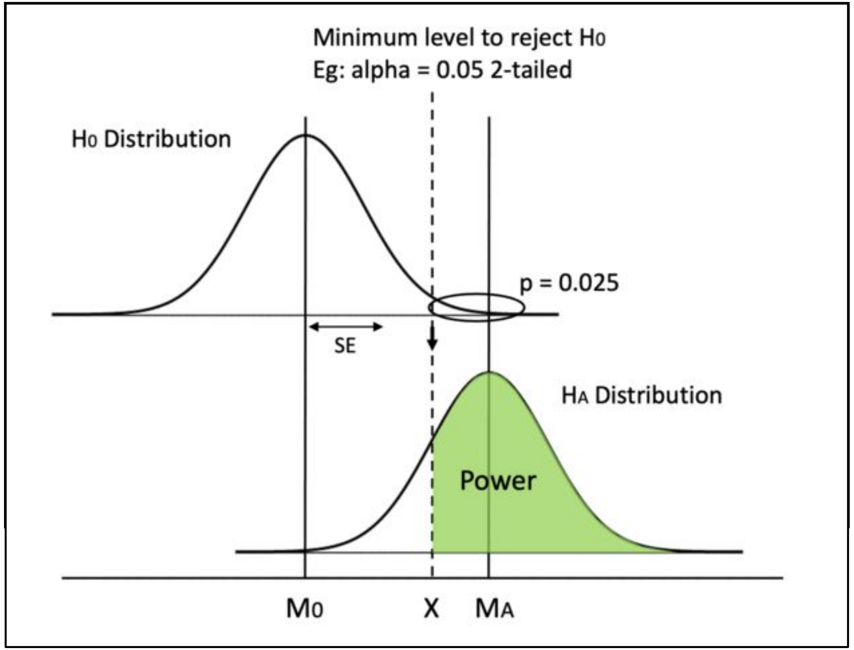
- This makes it easier to reject $\mathbf{H_o}$
- Also, this "shifts" the **critical line** to the left, leading to more area in the "**power region**" of the **alternative distribution**
- Intuitively, we now have a higher probability of rejecting $\mathbf{H_o}$, and **power** is probability of rejecting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



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

How to Increase Power: Increase Sample Size

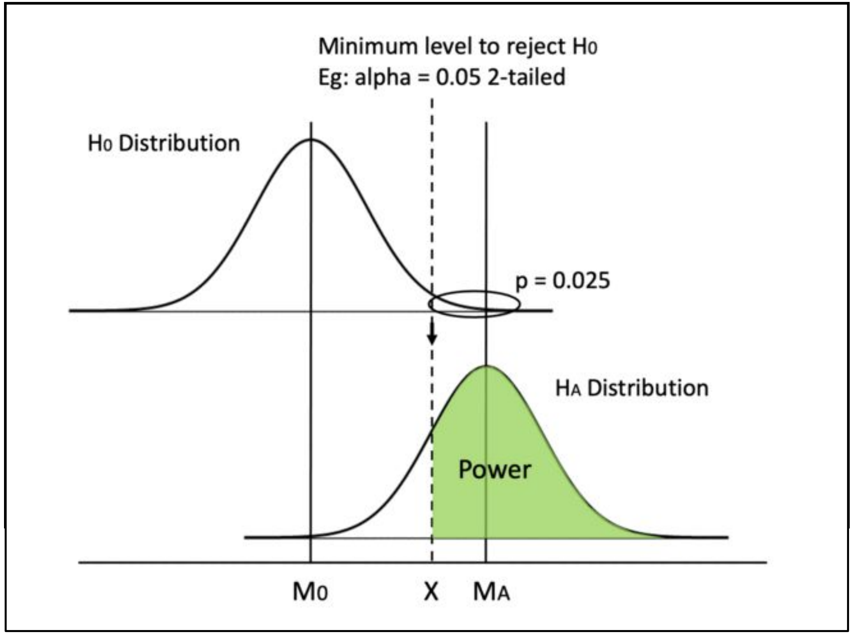
- This decreases **spread** of **histograms**, leading to less overlap between **null distribution** and **alternative distribution**



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

How to Increase Power: Increase Effect Size

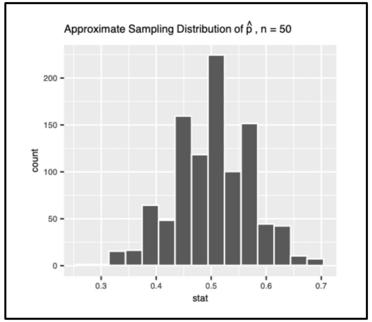
- **Effect Size**: Difference between **true value of parameter** and **null value**
- This makes it easier for us to notice a difference
- Also, this "shifts" the **center of the alternative distribution** to the right, leading to more area in the "**power region**"



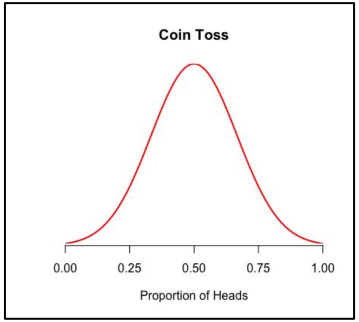
https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

Simulation-Based to Theory-Based Inference

- When the assumptions of CLT are met, we can recast our **sample statistics** as **random variables** and conduct **theory based-inference**
  - Before, we **simulated** our **null dist.**
  - Now, we use known distributions (e.g., **normal dist.**) as our **null dist.**
- The code changes, but the interpretation is (mostly) same

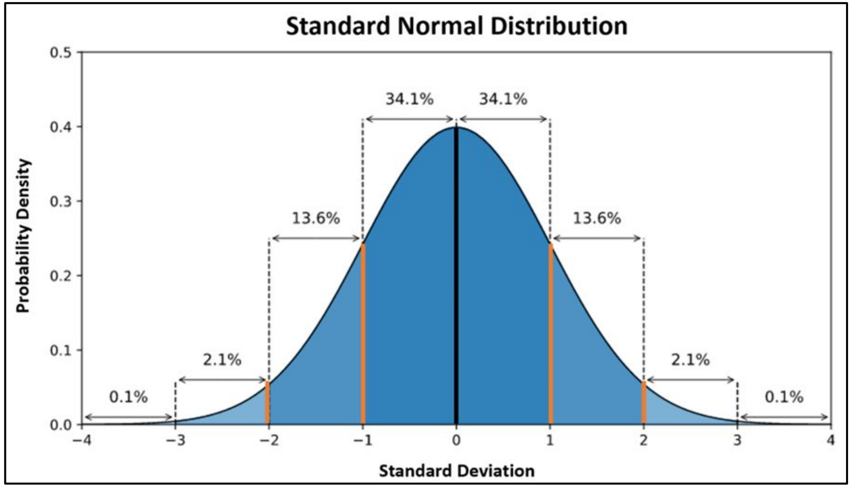


→



Standardization: Z-Score, T-Score

- Before, we used our **(observed) sample statistic** as our **test statistic**
  - *"The prob. we get our observed test stat. of 75% heads (or more extreme) is…"*
- We can use **z-score** (for normal dist.) and ***t*-score** (for *t*-dist.), which are **standardized** versions of the **sample stat.**
  - *"The prob. we get a z-score of 2.4 (or more extreme) is…"*
- They measure how many **SDs** the **sample stat.** is away from its **mean**
  - **z ~ N(0, 1)**, and the **standard normal dist.** is easy to use as our **null dist.**



"Estimate" vs. "Statistic" in R

- **Estimate** is the **observed sample statistic** (i.e., the numeric quantity calculated with the dataset)
  - *Here, the dataset had a sample correlation coefficient of -0.398*
- **Statistic** is the **standardized test statistic** (i.e., z-score or *t*-score)
  - *Here, that sample statistic is 7.07 standard errors below what we'd expect if the null hypothesis were true (i.e., if there is no correlation between age and vitamin D levels)*
  - *Here, the standardized test statistic is a t-score that's distributed t(266)*

```
## # A tibble: 1 x 8
##   estimate statistic  p.value parameter conf.low conf.high method    alternative
##      <dbl>     <dbl>    <dbl>     <int>    <dbl>     <dbl> <chr>     <chr>
## 1   -0.398     -7.07 6.89e-12       266       -1    -0.309 Pearson'~ less
```

A (Brief) Rundown on Linear Regression

- **Linear regression**: Models the **linear** relationship between **numerical response variable (y)** and **explanatory variables (x)**, which can be either **numerical** or **categorical**
  - **Simple linear regression** has one **explanatory variable**
  - **Multiple linear regression** has multiple **explanatory variables**
- **Multiple linear regression** can be **equal-slopes** or **varying-slopes**
- $\hat{B}_k$ (coefficient of **predictor** x_k) is predicted mean change in **y** corresponding to **1 unit change in** x_k when **all other predictors are held constant**
  - If x_k is **numerical**, think of **slope**
  - If x_k is **categorical**, think of **difference in means** ($\bar{y}_{other} - \bar{y}_{baseline}$)

The General “Formulas” for Equal-Slopes (When x_2 Is Categorical)

- **$\hat{B}_0$** is **y-intercept of line when x_2 = 0**
  - *Ex: For houses with central air (x_2 = 0), when living area (x_1) equals 0, the price ($\hat{y}$) is $42,595 ($\hat{B}_0$), on average*
- Since this is equal-slopes, **$\hat{B}_1$** is **slope of both lines** (a.k.a. increase in $\hat{y}$ after 1-unit increase in x_1, **controlling for x_2**)
  - *Ex: Controlling for central air (x_2), as living area (x_1) increases by 1 unit, price ($\hat{y}$) increases by $107 ($\hat{B}_1$), on average*
- $\hat{B}_0 + \hat{B}_2$ is y-intercept of line x_2 = 1, so **$\hat{B}_2$** is **difference in $\hat{y}$ between both lines ($\hat{y}_{other}$ - $\hat{y}_{baseline}$), controlling for x_1**
  - *Ex: Controlling for living area (x_1), houses without central air (x_2 = 0) cost $28,451 ($\hat{B}_2$) less than houses with central air (x_2 = 1), on average*

The General "Formulas" for Varying-Slopes (When x_2 Is Categorical)

- **$\hat{B}_0$** is **y-intercept of line when x_2 = 0**
  - *Ex: For houses with central air (x_2 = 0), when living area (x_1) equals 0, the price ($\hat{y}$) is -\$8,248 ($\hat{B}_0$), on average*
- **$\hat{B}_1$** is **slope of line when x_2 = 0**
  - *Ex: For houses with central air (x_2 = 0), as living area (x_1) increases by 1 unit, price ($\hat{y}$) increases by \$132 ($\hat{B}_1$), on average*
- $\hat{B}_0 + \hat{B}_2$ is y-intercept of line when x_2 = 1 (houses without central air), so **$\hat{B}_2$** is **difference in y-intercepts between both lines (b_{other} - $b_{baseline}$)**
  - *Ex: When living area (x_1) equals 0, houses without central air (x_2 = 1) cost \$53,226 ($\hat{B}_2$) more than houses with central air (x_2 = 0), on average*
- $\hat{B}_1 + \hat{B}_3$ is slope of line when x_2 = 1 (houses without central air), so **$\hat{B}_3$** is **difference in slopes between both lines (m_{other} - $m_{baseline}$)**
  - *Ex: Houses without central air (x_2 = 1) have a lower slope than houses with central air by \$44.6/unit ($\hat{B}_3$)*

Inference with Linear Regression

- When assumptions are met, we can conduct **inference** to learn about **population parameters** (beta coefficients in **population model**)
  - Our **model** is constructed from **sample data**—i.e., we have $\hat{B}_k$, but we want to know about B_k!
- We can conduct a hypothesis test on a **slope term** (in **equal-slopes model**)
  - $\mathbf{H_0: B_k = 0}$ (i.e., the slope is zero, so there is no association between X_k and Y after controlling for all other predictors in the model)
  - $\mathbf{H_A: B_k \neq 0}$ (i.e., there is an association between X_k and Y after controlling for all other predictors in the model)
- Or on an **interaction term** (in **varying-slopes model**)
  - $\mathbf{H_0: B_k = 0}$ (i.e., slope between Y and numerical expl. variable X_j doesn't differ by category)
  - $\mathbf{H_A: B_k \neq 0}$ (i.e., slope between Y and numerical expl. variable X_j differs by category)

Advanced Inference Scenarios

- Recall we've expanded our toolkit and can handle more **advanced inference scenarios** (in addition to **inference with linear regression**)
- If we have a **categorical variable with more than 2 categories…**
  - We use **chi-squared** as our test when both response variable and explanatory variable are 2+ categorical, **χ^2 (chi-squared)** as our test statistic
  - We use **ANOVA** as our test when **response variable** is **numerical** and **explanatory variable** is **2+ categorical**, with **F-statistic** as our test statistic
- If our dataset has **paired measurements** (i.e., each observation can be matched to another observation, like a person "before and after")…
  - We use a **paired *t*-test**, with ***t*-score** as our (standardized) test statistic

Let’s Recap Our Inference Scenarios

https://drive.google.com/file/d/1rvVsTfhaK_92yWn8DTp-f97SF3tPmkEr/view?usp=drive_link

More on Inference and Linear Regression

- These are big topics, and for the sake of time, I couldn't possibly cover everything in an hour
- If you want more (in-depth) information, you can check out…
  - Week 5 slides, Week 6 slides, and Week 9 slides for inference
  - Week 10 slides and Week 11 slides for linear regression

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…"*
  - $H_0: p = 55\%$, $H_A: p > 55\%$ ($p \stackrel{?}{=} 60\%$)
- **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?"*

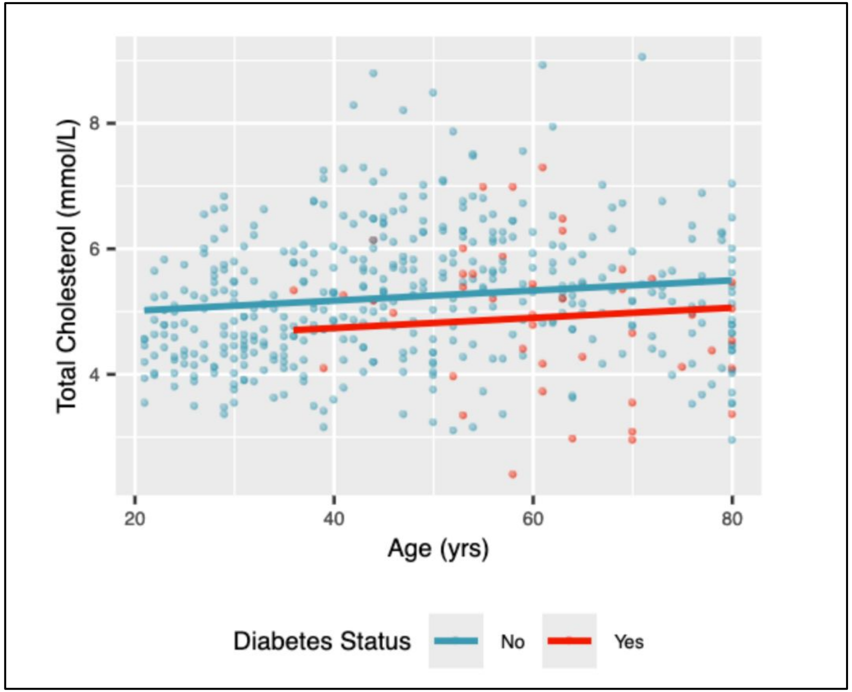
Can a Type I (or Type II) Error Occur?

- 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

When Should I Use Equal-Slopes vs. Varying-Slopes?

- Consider your goal with the model (and what's being asked of you)
- With **varying-slopes**, certain questions (like the average difference in cholesterol between diabetic groups, controlling for age) can't be answered
- With **equal-slopes**, certain questions (like whether or not the relationship/slope differs between groups) can't be answered

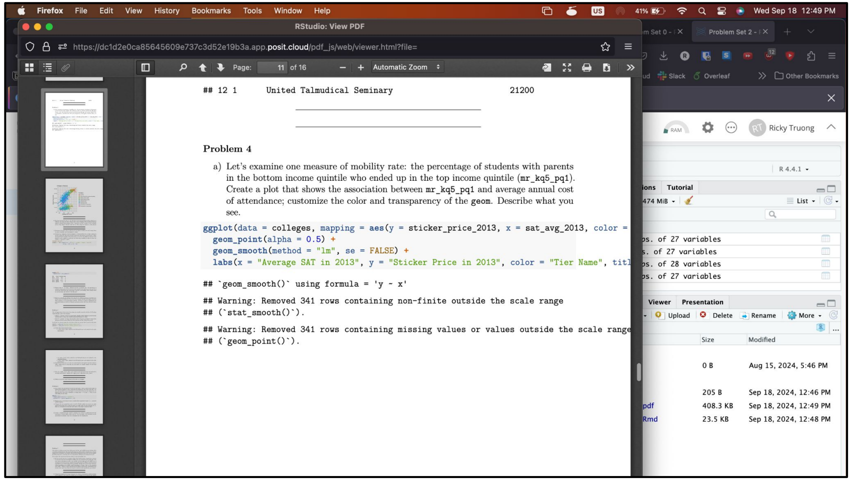


Debugging Code: Comment Out, Partial Credit

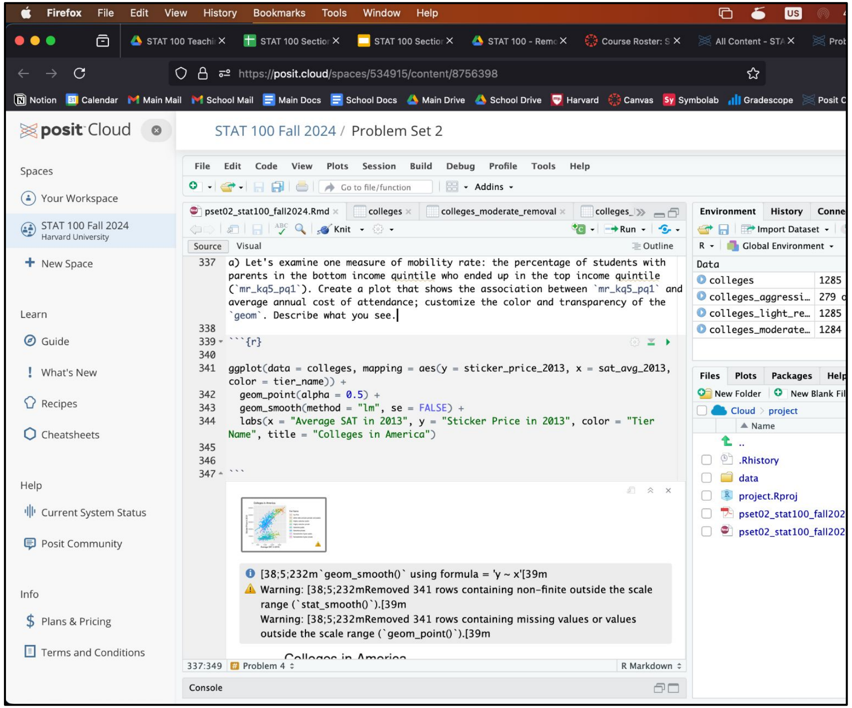
- To debug code, consider commenting out (#) the possibly-problematic lines
  - If the code runs without the line, you know it's the problem
  - R will often tell you which line(s) are causing an issue
- Do NOT delete all your code! You may get partial credit even if your code doesn't run
  - Either 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!

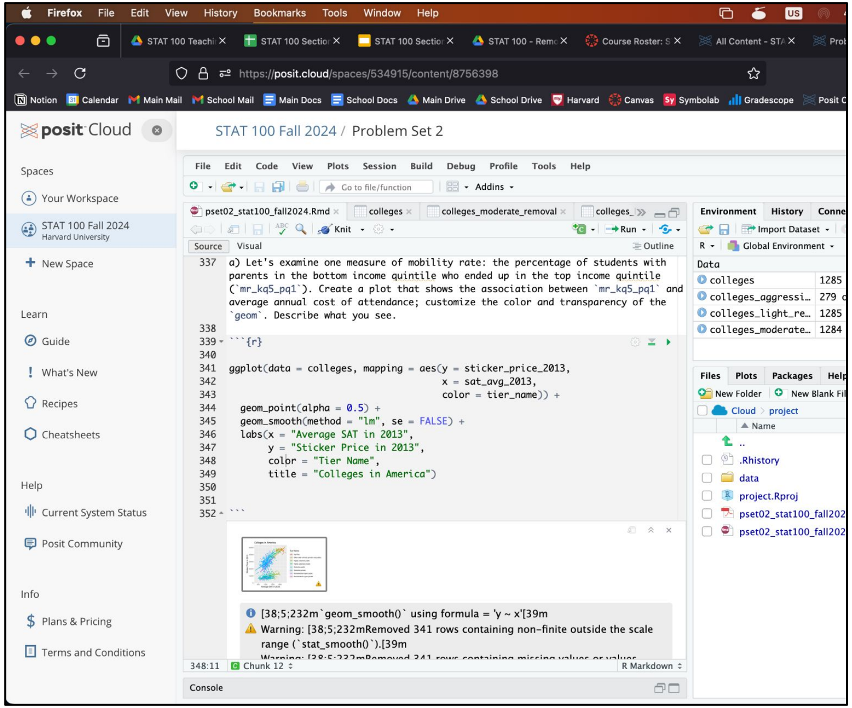
- Make sure your code isn't running off the screen in your PDF
  - If the grader can't read your code, you might get points off
- Hit "Return" to start a new line
  - Best to do this after commas (,) and plus signs (+)



Messy Code

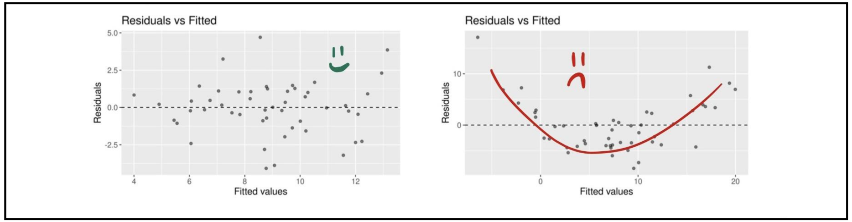


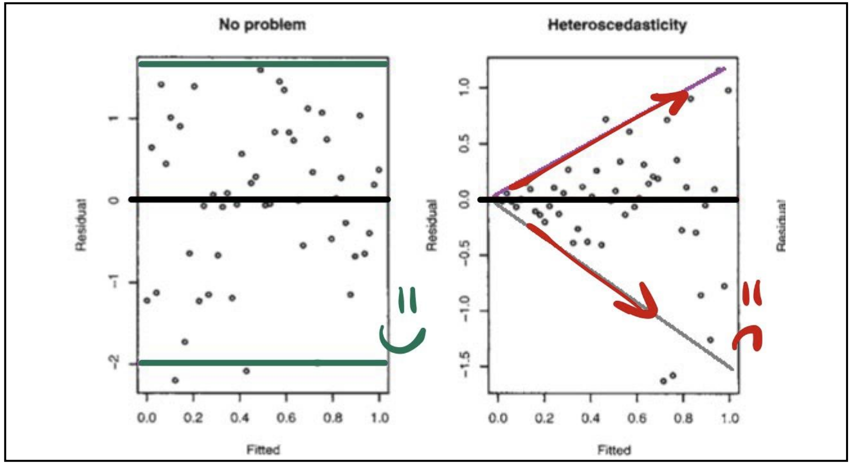
Clean Code



Assumptions for Linear Regression: Interpretations are Different

- Recall the assumptions to conduct **inference with linear regression**
  - This applies to **simple** and **multiple**
- #**1** (**Linearity**) and #**2** (**Constant Variability**) use **residual plots**, but their interpretations differ
  - For #**1**, cite the random scatter about y = 0
  - For #**2**, cite upper and lower bounds to show there is no "fanning"





Interpreting the "Tibble" from Linear Regression

- Top to bottom, the **beta coefficients** are listed in **ascending order** starting at 0
  - The top-most number is $\hat{B}_0$, the next one is $\hat{B}_1$, ...
- The **baseline group** is the **opposite** of the **group shown**
  - "centralAir: Yes" (a.k.a. houses WITH central air) is our **baseline group**
- **Interaction term** has 3 "things" (numerical variable, categorical variable, and category)
  - "livingArea: centralAirNo" is our **interaction term**

```
## # A tibble: 4 x 2
##   term                        estimate
##   <chr>                          <dbl>
## 1 intercept                     -8248.
## 2 livingArea                      132.
## 3 centralAir: No                53226.
## 4 livingArea:centralAirNo        -44.6
```

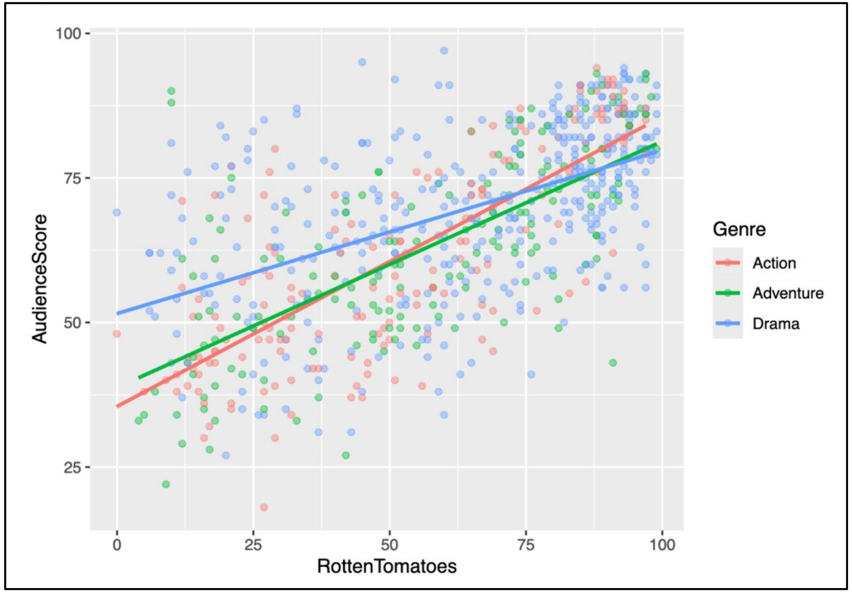
difference in intercepts

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$ $\hat{\beta}_3$

coefficient on interaction term

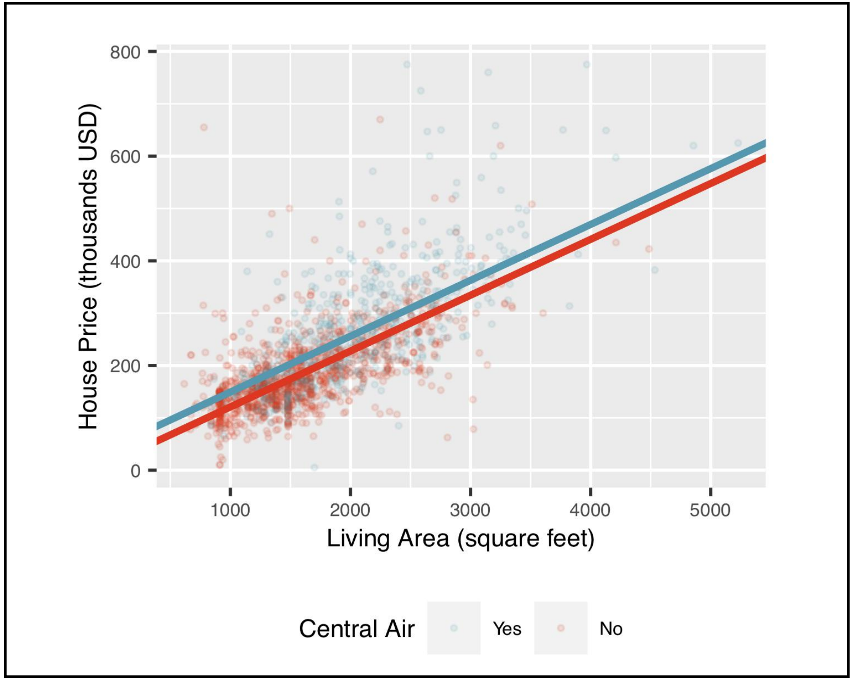
“Association” and “Relationship” Are Slope

- **Association** is another word for **slope**, which is the measure of the **relationship** between **2 numerical variables**
  - *Ex: “Assess whether there’s evidence that in the population, the relationship between audience score and critic score differs between action movies and drama movies”*
  - This is basically asking whether or not the slopes are different, which sounds like a **varying-slopes model** is needed (along with some inference)



Holding Everything Else Constant

- When interpreting the coefficients of an **equal-slopes model**, don't forget to mention that we're holding everything else constant
- We're controlling for the other **variable(s)**
  - We're "slicing" at some living area
  - *Ex: "Controlling for living area, houses without central air cost $28,451 less than houses with central air, on average"*
  - We're "picking" one of the lines
  - *Ex: "Controlling for central air, as living area increases by 1 unit, price increases by $107, on average"*



pnorm(), qnorm(), pt(), qt()...

- Want **probability**?
  - Use pnorm(), pt()
  - This is often done for **p-value** in **hypothesis testing**
- Want **quantile** (i.e. percentile)?
  - Use qnorm(), qt()
  - This is often done to find **z*** or **t*** (critical values) in **confidence intervals**

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)

Closing Remarks

- Thank you all very much for coming!
- Slack the teaching team if you have any questions
  - Slack me specifically with questions related to this slideshow (if you have any)
- The best thing you can do is practice
  - Consider going to the **Practice Session** tomorrow!
- Best of luck! You're all going to do amazing 😁