ORIE 5355: Applied Data Science - Decision-making beyond Prediction

Lecture 3: Survey weighting methods

Nikhil Garg

Reminder: The task

- Each person j has some complex thoughts about the course, C_j
- This is reduced to a binary opinion, $Y_j \in \{0, 1\}$
- We want to measure $\mu = E[Y_j]$, the population mean opinion on some issue
- Each person also has covariates, X_j
- We also may care about *conditional* means

$$\mu_{\mathrm{ORIE}} = E[Y_j \,|\mathrm{ORIE\ program}]$$

Example:
"Do you like the class so far?"

Options: "yes" and "no"

What fraction of people like the class so far?

Degree program, whether you like afternoon classes, etc

Fraction of people in ORIE who like the class

Challenge: Sample doesn't represent population

- For each person j, let $O_j \in \{0,1\}$ be whether they answered
- You have data $\{Z_j \mid j \text{ responds}, O_j = 1\}$ Again, you do

$$\hat{\mu} = \frac{1}{n_{obs}} \sum_{j \in \{j \mid O_j = 1\}} Z_j$$

Uncorrelated: Whether you answered is unrelated to what your true opinion is

For now, let's assume $\mathrm{Y_j} = Z_j$

When does $\hat{\mu} \to \mu$? When Y_j and O_j are uncorrelated

- Suppose someone flipped a coin before deciding to answer
- Suppose someone only answered if their opinion was $Y_j = 1$

Plan for rest of the day

Methods to tackle sample representation issues

- Stratifying sample *before* you poll
- Weighting techniques *after* you have responses

Differential response on *known* covariates

- Suppose we have a single binary covariate $X_j \in \{0,1\}$ indicating whether they graduated from college
  Half the population went to college

Whether MEng or MS degree

- Suppose whether people answer is correlated with education

$$\Pr(O_j = 1 | X_j) = \begin{cases} 0.1 \text{ if } X_j = 0 \\ 0.4 \text{ if } X_j = 1 \end{cases}$$

Whether you answer is correlated with degree program

- Education is also correlated with opinion Y_j in some unknown manner
- We want to measure $\mu = E[Y_j]$, the population mean
- *No other correlations between whether they answer and opinion:*
  Opinion Y_j *is independent of whether they respond* O_j, *conditional on* X_j

Given your degree program, whether you respond is uncorrelated with your opinion

Notation with features (groups)

- Number of people *called*: n — # of people I poll from class
- Population response rate for group ℓ: O^ℓ — Response fraction in degree ℓ
- Population mean opinion for group ℓ: μ^ℓ — "Likes class" fraction in degree ℓ
- Population fraction for group ℓ: P^ℓ — Fraction of class in degree ℓ
- Corresponding *sample* values are: $\hat{O}^\ell, \hat{\mu}^\ell, \hat{P}^\ell$

  (i.e., $\mathrm{n}\hat{P}^\ell\hat{O}^\ell = \mathrm{n}^\ell_{\mathrm{obs}}$) — # who answered in degree ℓ

and so:

$$\mu = \frac{P^0\mu^0 + P^1\mu^1}{P^0 + P^1} = P^0\mu^0 + P^1\mu^1 = 0.5\,\mu^0 + 0.5\,\mu^1$$

in example — True mean opinion

$$\hat{\mu}_{naive} = \frac{\hat{O}^0\hat{P}^0\hat{\mu}^0 + \hat{O}^1\hat{P}^1\hat{\mu}^1}{\hat{O}^0\hat{P}^0 + \hat{O}^1\hat{P}^1} \rightarrow \frac{O^0P^0\mu^0 + O^1P^1\mu^1}{O^0P^0 + O^1P^1} = 0.2\,\mu^0 + 0.8\,\mu^1$$

Mean poll response

Naïve method in this notation

$$\hat{\mu}_{naive} = \frac{\left(\sum_{j\in\{j \mid O_j=1, X_j=0\}} Z_j + \sum_{j\in\{j \mid O_j=1, X_j=1\}} Z_j\right)}{\mathrm{n}^0_{\mathrm{obs}} + \mathrm{n}^1_{\mathrm{obs}}}$$

$$= \frac{\mathrm{n}^0_{\mathrm{obs}}\hat{\mu}^0 + \mathrm{n}^1_{\mathrm{obs}}\hat{\mu}^1}{\mathrm{n}^0_{\mathrm{obs}} + \mathrm{n}^1_{\mathrm{obs}}} = \frac{\left(\#(Y_j=1)\text{ from Group 0 } + \#(Y_j=1)\text{from Group 1}\right)}{\text{Total Respondents}}$$

$$\rightarrow \frac{O^0P^0\mu^0+O^1P^1\mu^1}{O^0P^0+O^1P^1} \neq \mu = \frac{P^0\mu^0+P^1\mu^1}{P^0+P^1} \text{ unless } O^0 = O^1$$

$P^0O^0/\,(P^0O^0+P^1O^1)$ is limit fraction of respondents from Group 0

Bias (even with $n \rightarrow \infty$): Limit fraction does not match the population fraction

Variance (with finite n): Sample values do not match limit values

Stratified sampling

Stratification: change who you call

- Suppose you have L mutually exclusive demographic groups:

  A population that is heterogeneous *across* groups

  Relatively homogeneous *within* groups

  (Exactly the setup we have)

Y_j is independent of O_j, conditional on X_j

- Then, instead of calling n completely random people

  Call n^{ℓ} people from group ℓ

  Where n^{ℓ} is determined by how likely each group is to respond

  If MEng students are less likely to respond, call more of them

- Even if each group responds at same rate, this leads to *lower variance* estimates
- With differential response rates, can also correct the *bias* in *mean*

Why does it work?

- With differential response rate: we can “cancel out” the differential response rate by just calling more people from that group

- Even without differential response rates, just differential opinion:

  There are two sources of variance in estimation:

  Which groups are over- and under- sampled due to noise

  What the opinion of each person is

  Stratification mitigates the first source of variance

Why does it work? (Mathematically)

$$\hat{\mu} = \frac{\left(\sum_{j\in\{j \mid O_j=1, X_j=0\}} Z_j + \sum_{j\in\{j \mid O_j=1, X_j=1\}} Z_j\right)}{\mathrm{n}^0_{\mathrm{obs}} + \mathrm{n}^1_{\mathrm{obs}}}$$

Now $n^\ell \hat{O}^\ell$ instead of $\mathrm{n}\hat{P}^\ell \hat{O}^\ell$

$$= \frac{(\text{\#1 from group 0 } + \text{\#1 from group 1})}{\text{Total Respondents}}$$

$$\rightarrow \frac{n^0 O^0 \mu^0 + n^1 O^1 \mu^1}{n^0 O^0 + n^1 O^1} = \mu = \frac{P^0 \mu^0 + P^1 \mu^1}{P^0 + P^1} \text{ if } \frac{n^0 O^0}{P^0} = \frac{n^1 O^1}{P^1}$$

Calling more in ratio of non-response

With stratification, cancel out the bias *because* you simply asked more people from the group with lower response rate

It also reduces variance, even if $O^0 = O^1$ (and $n^0 = n^1$)

Stratification in practice

- You often don't know group specific response rates O^{ℓ}
  - Define groups and then keep sampling until you have enough samples
  - Weighting after sampling (covered next)
- How many groups/what groups do you choose?
  - Our example had a binary covariate we called "education"
  - What about stratifying ethnicity, or intersectional groups (ethnicity x gender)?
  - Why stop there? Why not ethnicity x gender x education x age …?
  - As number of groups increase, number of people in each group goes down
- Remember the rule: create groups such that the response rates is not correlated with what their answer is, *within each group*

Response Y_j *is independent of whether they respond* O_j*, within each group* X_j

Questions?

Weighting

Main idea for weighting

- In stratified sampling, we balanced out the groups according to their population percentage *before* we called people
- With weighting, we try to do the same thing, but *after* we call people and know how many from each group responded
- Why?
  - You might not know response rates per group
  - You might not know a person's demographics until you call them
  - Can run *sensitivity analyses:* "what would the estimate be if this demographic group only composes x% of the population instead of y%?"
- Comes at a cost: doesn't have the same variance reduction properties as does stratified sampling

Main idea, 2 steps:

Step 1: Use the responses to estimate the **mean response for each group** ℓ, i.e., get an estimate $\hat{\mu}^{\ell}$ of the true opinion μ^{ℓ}

Step 2: **Do a weighted average of** $\hat{\mu}^{\ell}$; each group is given weight W^{ℓ}

$$\hat{\mu} = \sum_{\ell} W^{\ell} \hat{\mu}^{\ell}$$

If $W^{\ell} = P^{\ell}$ and $\hat{\mu}^{\ell} \to \mu^{\ell}$, then $\hat{\mu} \to \mu$

Details differ in:

- how to construct estimate $\hat{\mu}^{\ell}$
- how to calculate weight W^{ℓ}
- what groups ℓ to consider

Naïve Weighting

Step 1: Use the mean response for each group ℓ separately, i.e.

$$\hat{\mu}^{\ell} = \frac{\sum_{j \in \{j \mid O_j = 1, X_j = \ell\}} Z_j}{\left|\{j \mid O_j = 1,\ X_j = \ell\}\right|}$$

(calculate the mean response from that group in the survey)

Step 2: Weight W^{ℓ} is our best guess of true population fraction P^{ℓ} for group ℓ

Complication: How many groups/which ones?

- If group too broad (e.g., group ℓ just gender), then break cardinal rule:

  Need: Opinion Y_j *is independent of whether they respond* O_j, *conditional on group* ℓ

- If group is too specific (*ethnicity* x *gender* x *education* x *age*), then:

  Problem 1: Estimate $\hat{\mu}^{\ell} = \frac{\sum_{j \in \{j \mid O_j = 1, X_j = \ell\}} Z_j}{|\{j \mid O_j = 1, X_j = \ell\}|}$ might be really bad

  Too few respondents in a group → high variance (1 person might determine entire average)

  Problem 2: We might not know population fraction P^{ℓ}

This is called a bias-variance tradeoff!

Tackling Problem 2: Population weights

- Suppose very specific group (*ethnicity* x *gender* x *education* x *age*)
- Naïve: try to figure out true population fraction (“joint distribution”)

  “W^{ℓ} = P^{ℓ} *fraction of pop is college educated white women age 35-44”*
- Easier: Use “marginal” distribution for each covariate

  “a fraction of population is women”

  “b fraction of population is college educated”

  “c fraction of population is white”

  “d fraction of population is age 35-44”

  $\Rightarrow$Pretend “W^{ℓ} = *abcd fraction of pop is college educated white women age 35-44”*
- *Not covered -- “raking”: match marginal distribution for each covariate without assuming that marginal distributions make up joint distribution*

The homework

- In the homework, first we define groups just based on a single covariate, for example gender, ethnicity/race, political party, etc.
  - (e.g., group ℓ just based on gender); we give you P^{ℓ}
- Then we define groups based on 2 covariates; we give you P^{ℓ}
- Then we define groups based on 2 covariates and ask you to construct P^{ℓ} based on marginal distributions

Tackling Problem 1: MRP

Problem 1: Estimate $\hat{\mu}^{\ell}$ might be really bad

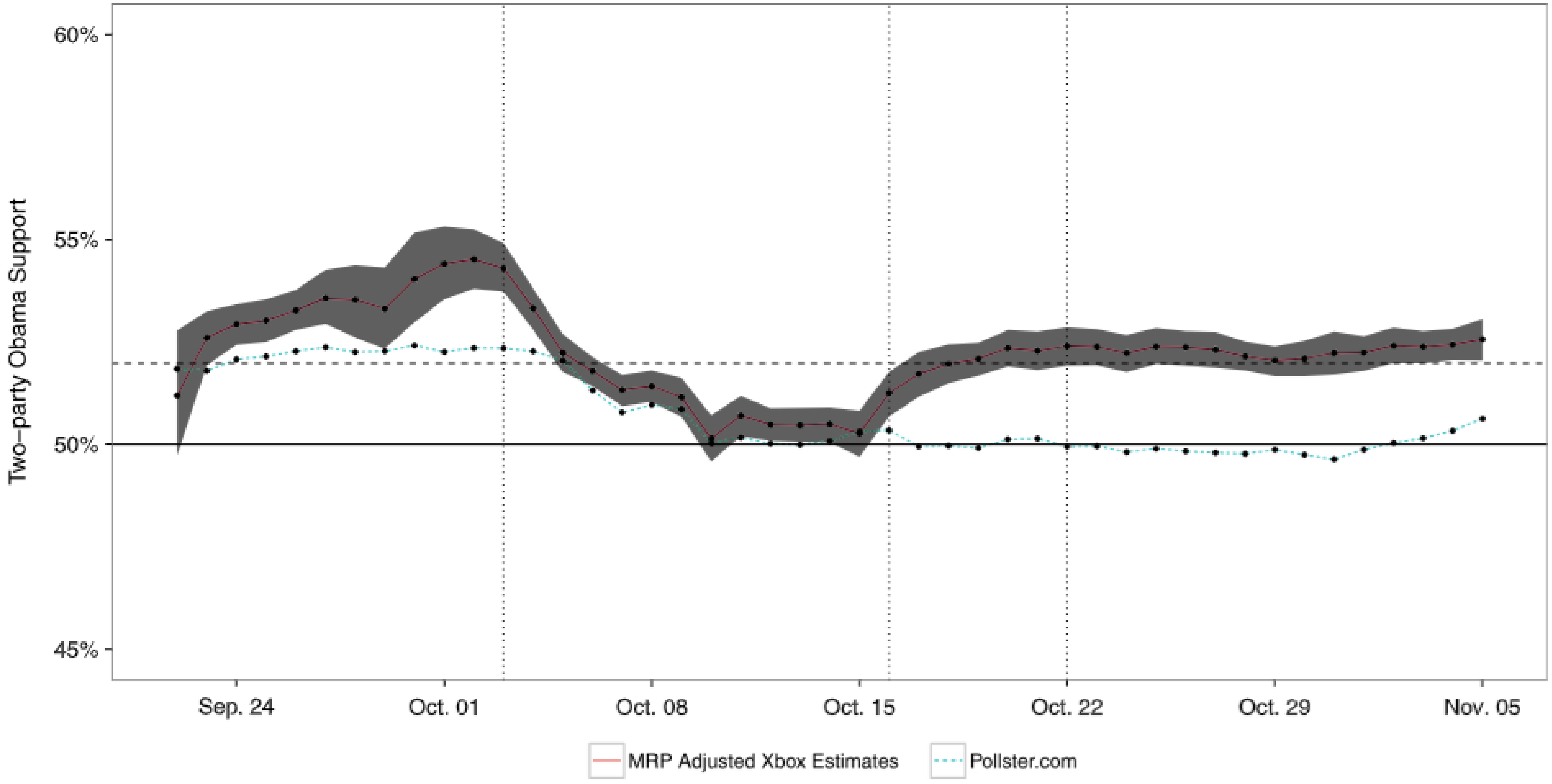
Too few respondents in a group → high variance (1 person might determine entire average)

- Somehow this seems wrong: presumably, the estimate for a group should be very close to that of a “neighboring” group
- “Multilevel regression and poststratification” (MRP)

Main idea: Train a (Bayesian) regression model to get estimate $\hat{\mu}^{\ell}$ for each set of covariates. Then, “post-stratify” by weighting $\hat{\mu}^{\ell}$ by population fraction P^{ℓ}

For groups with many samples, estimate $\hat{\mu}^{\ell}$ just based on that group; otherwise, based on “neighboring” groups

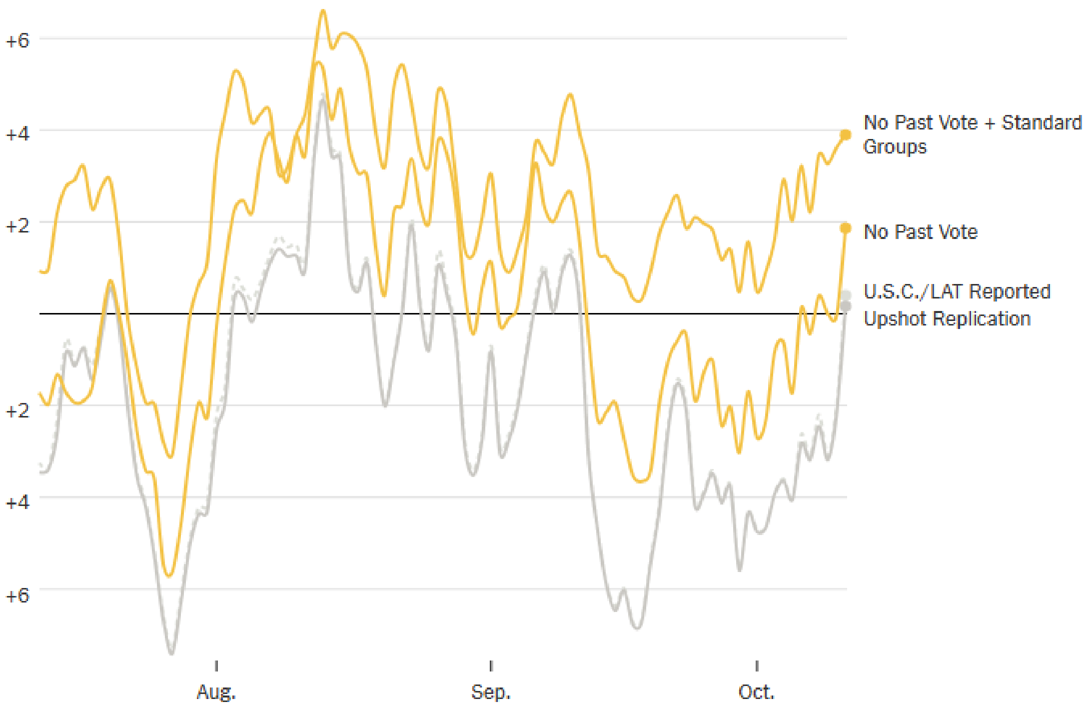
Example: can we use Xbox gamer polls for elections?



Weighting Makes a Difference

About half of the difference between the U.S.C./LAT poll and a typically weighted poll comes from the past vote weight.

How the U.S.C./LAT poll might change using different weighting techniques



Source: Upshot analysis of the U.S.C./Los Angeles Times poll. “Standard” weights include age, race, gender and education, and controlling for changes in past vote.

Parting thoughts on weighting

- Where do the population percentages come from? In political polling, you need to define a universe of "likely voters"
- Methods not covered here: *Inverse Propensity Scoring*, and *Matching*
- Note, can only weight when you observe the covariates for each respondent!
- What if sampling bias is correlated with a feature you don't observe?
  - Next time!

Possible Errors

- Finite sampling error: I got some small number of responses compared to if I had asked a billion people (source of variance)
- Measurement error: Given someone responded, their response doesn't equal their actual opinion
- **Selection effect: true opinion among those who respond differs from population true opinion**
- Construct mismatch: Y isn't actually what I wanted!

Weighting only helps with selection

Parting thoughts

Be purposeful! Does your numeric data capture what you want it to?

Be skeptical! Just because a poll was “random” doesn’t make it good

Unmeasured confounding and quantifying uncertainty

1 slide summary

Challenge

- Stratification and weighting help us when we have covariates that capture the selection bias and different opinions
  - Response rates correlate with education, and we know education level of respondents
- What if we don't have access to these covariates? This is called "unmeasured confounding"

What to do about it

- We can't hope to "correct" for unmeasured confounding
- However, we can *quantify the uncertainty* under *assumptions* on how bad the problem is
  - "If response rates were this different by group, and if this group has this magnitude of different opinion, here's how different my answer would be"

The challenge

- In the last lecture, weighting helped us deal with *measured* selection bias/differential non-response

  Response rates and political opinions both correlate with educational status;

  (1) Education status can be asked for during the poll
  (2) We can roughly guess at voter distribution by education status
  (3) Then use various *weighting* techniques

- What if response rates & opinions depend on a covariate that we don't observe, or that we don't know the population distribution of?
- Very little we can do to recover "point-estimate" of population opinion
- However, we can *quantify the uncertainty* under *assumptions* on how bad the problem is

Setup

- Suppose there is a (binary) covariate u_j that correlates with both the opinion of interest Y_j and whether people respond O_j.
- You don't observe u_j for any individual j
- u_j is the only unmeasured confounding: O_j is uncorrelated with true opinion Y_j given u_j -- but we don't have u_j
- You have an estimate $\hat{\mu}$ (raw average of responses)
- Idea: Make assumptions on "how bad" the unmeasured confounding can get to derive uncertainty regions for your estimate of interest.

How to quantify uncertainty

- If we assume like we did on the last slide: “Conditional on what group the respondent belongs to, their opinion does not correlate with whether they respond”
- Then, you can do some math where your error decomposes into the difference between groups in *whether they respond* and *true opinion differences*

$$\hat{\mu} - \mu \to \left(\tilde{P}^1 - P^1\right)\left(E\left[Y_j \mid u_j = 1\right] - E\left[Y_j \mid u_j = 0\right]\right)$$

More detail: Notation and Insight

- True population fractions of u: $\mathrm{P}^1 = \mathrm{Pr}(u_j = 1), 1 - \mathrm{P}^1 = \mathrm{Pr}(u_j = 0)$
- Response fractions: $\tilde{\mathrm{P}}^{\ell} = \mathrm{Pr}(u_j = \ell \mid O_j = 1)$
- $\mu \stackrel{\text{def}}{=} E[Y_j] = P^1 E[Y_j \mid u_j = 1] + (1 - P^1) E[Y_j \mid u_j = 0]$
- $\hat{\mu} \to E[Y_j \mid O_j = 1] = \tilde{\mathrm{P}}^1 E[Y_j \mid u_j = 1, O_j = 1]$
  $+ (1 - \tilde{P}^1) E[Y_j \mid u_j = 0, O_j = 1]$
- Insight:

$$E[Y_j \mid u_j = \ell, O_j = 1] = E[Y_j \mid u_j = \ell]$$

"Conditional on what group the respondent belongs to, their opinion does not correlate with whether they respond" ← We assumed this on last slide!

More detail: Quantifying uncertainty in math

$$\mu = P^1 E[Y_j \mid u_j = 1] + (1 - P^1) E[Y_j \mid u_j = 0]$$
$$\hat{\mu} \to \tilde{P}^1 E[Y_j \mid u_j = 1] + (1 - \tilde{P}^1) E[Y_j \mid u_j = 0]$$

Rearrange:

$$\hat{\mu} \to \mu + \left(\tilde{P}^1 - P^1\right) E[Y_j \mid u_j = 1] + \left(P^1 - \tilde{P}^1\right) E[Y_j \mid u_j = 0]$$
$$= \mu + \left(\tilde{P}^1 - P^1\right) \left(E[Y_j \mid u_j = 1] - E[Y_j \mid u_j = 0]\right)$$

Then, make assumptions on *whether respond* and *opinion* differences to quantify how far $\hat{\mu}$ can be from μ

If *either* response fractions or opinions between groups are similar, effect of unmeasured confounding is small!

Unmeasured confounding in ML

- In data science, we often care about *causal inference* (later in semester)
  - “What is the causal effect of going to a private high school on college success?”
  - Problem: In the US, private HS attendance correlated with parents’ wealth
- Unmeasured confounding (you might not know parents’ wealth) would mess up your *inference* of the relationship in a regression
- You can also quantify unmeasured confounding and range of effects in such cases

Questions?