

ORIE 5355: Applied Data Science -
Decision-making beyond Prediction
Lecture 3: Survey weighting methods

Nikhil Garg

Announcements

- Homework 1 posted
- Please make sure you have access to EdStem and are receiving announcement notifications
- Office hours set: (all on course calendar) Location TBD
- Contact each other via the homework buddy form
- Wait lists should be clearing soon: contact Student Affairs (not me)

Mean estimation from surveys

The task: estimate mean opinion

- Each person j has an opinion, $Y_j \in \{0, 1\}$
- We want to measure $\bar{y} = E[Y_j]$, the population mean opinion on some issue
- Each person also has covariates, x_j^k
- We also may care about *conditional* means
 $E[Y_j \mid \text{ORIE program}]$

Example:

“Do you like the class so far?”

Options: “yes” and “no”

\bar{y} : What fraction of people like the class so far?

Degree program, whether you like waking up at 9:30, etc

Fraction of people in ORIE who like the class

This problem is everywhere

- What fraction will vote for the Democrat in the next election
- What is the average rating of this product?
- Do people want the city to build a foot bridge to Manhattan?
- Are people happy with this new feature I just deployed?

Naïve method: take the mean

- Get list of people (watched the movie; from phone book)
- Call them, suppose everyone answers and get Y_j from each
- We now have $\{Y_j\}_{j=1}^N$, if called N people Random sample of people in this class
- Simply do, $\hat{y} = \frac{1}{N} \sum_j Y_j$ Average opinion of the sample
- By law of large numbers, if Y_i is independent and identically distributed according to the true population's opinion, then

$$\hat{y} \rightarrow \bar{y} \text{ as } N \rightarrow \infty$$

\bar{y} : Actual opinion of the class

What goes wrong

People don't give "true" opinion

Why?

- You're asking about something sensitive
- "social desirability" – people like making other people happy
- They're getting paid to answer the survey and just want to finish
- You know they other person is also going to rate you

Of course, then you're (likely) not going to succeed

People gave you \tilde{Y}_j , instead of Y_j

$$\hat{y} = \frac{1}{N} \sum_j \tilde{Y}_j$$

You lie because you want a better grade

\hat{y} does not converge to \bar{y} , *unless errors cancel out*

Your sample does not represent your population

- You just posted a poll on Facebook or Twitter, anyone could respond
- You called only landlines, and no one under 50 owns one anymore
- You only asked people to rate a movie after they've seen it
- You can only rate an item if you bought it *and didn't return it*
- Those with certain opinions are more likely not to answer
 - After bad experiences on online platforms
 - “Shy Trump voters” (?)

=> People who answer the poll are different than your population – “differential non-response”

Your sample does not represent your population, in math

- For each person j , let $A_j \in \{0,1\}$ be whether they answered
- You have $\mathbf{Y} = \{(A_j, Y_j)\}_{j=1}^N$, if called N people
Where $Y_j = \emptyset$ if $A_j = 0$ (they did not answer)

- Again, you do

$$\hat{y} = \frac{1}{|\{j \mid A_j = 1\}|} \sum_{j \in \{j \mid A_j = 1\}} Y_j$$

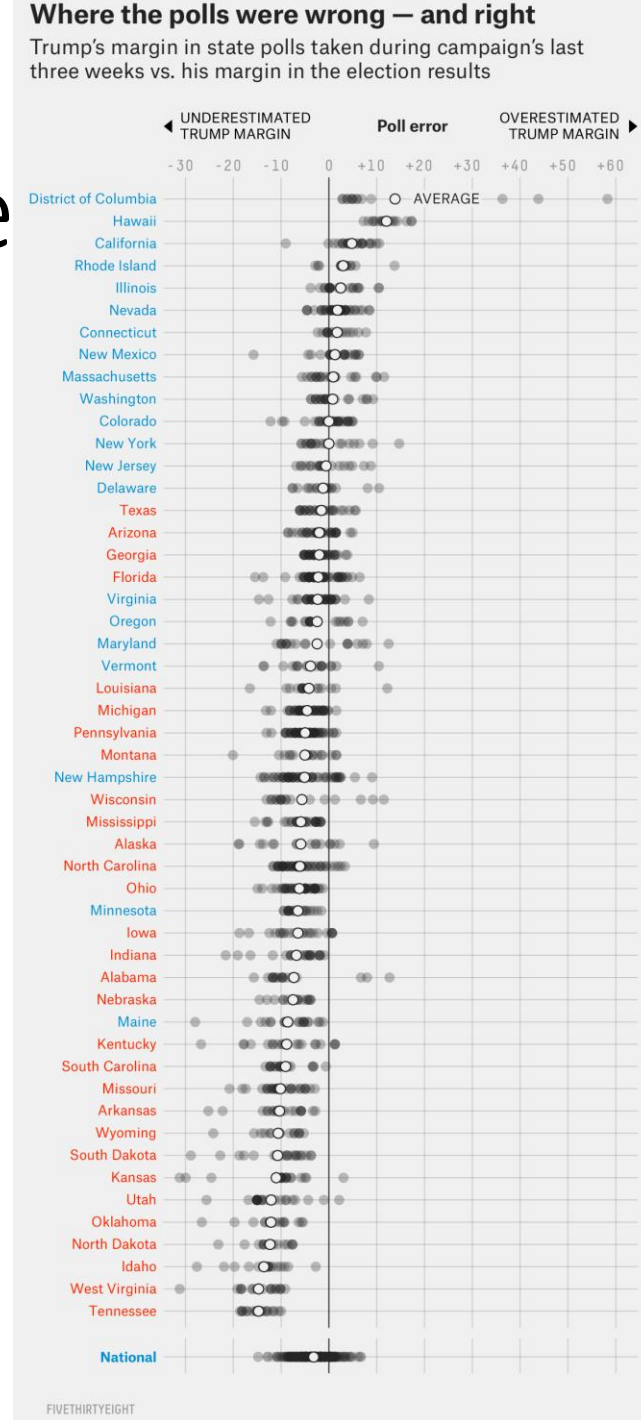
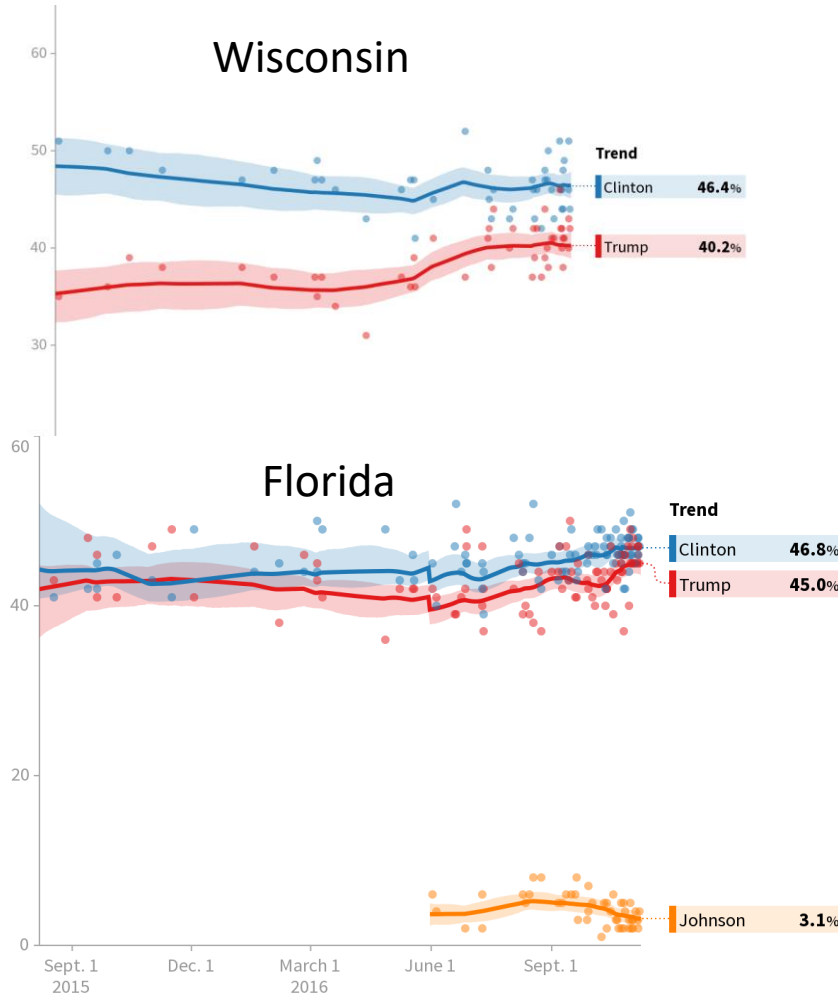
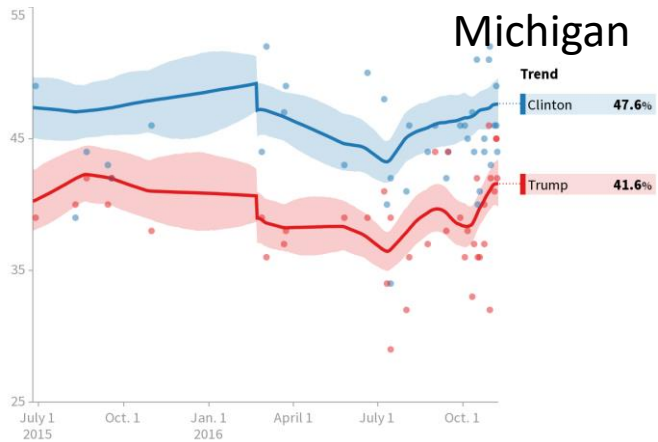
where $\{j \mid A_j = 1\}$ denotes the set of people who answered
and so $|\{j \mid A_j = 1\}|$ is the number of people who answered

\hat{y} does not converge to \bar{y} unless Y_j and A_j are uncorrelated

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

Case study: Polling in US 2016 presidential election

Polls were off (a bit) in the 2016 e

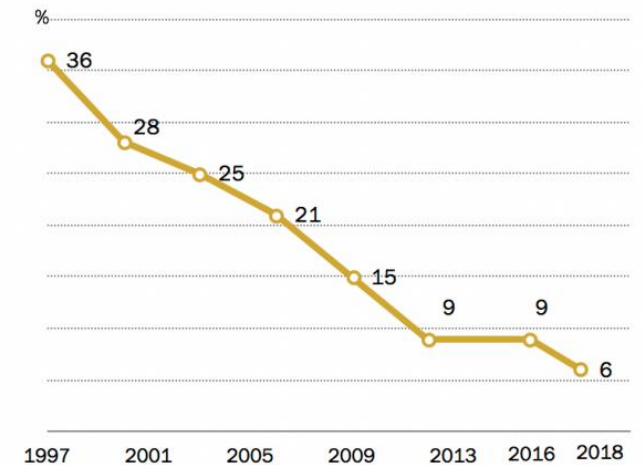


What happened?

- Professional pollsters spend a lot of time on getting opinions right
[We'll cover some of their techniques next time]
- But, polling is an increasingly challenging business
Basically no one answers a phone poll
Modeling opinions/turnout in diverse democracy is hard
“social desirability” → “shy Trump voters” (?)
- In 2016, turns out that less educated voters both:
Were less likely to answer polls
Were more likely to vote Trump

After brief plateau, telephone survey response rates have fallen again

Response rate by year (%)



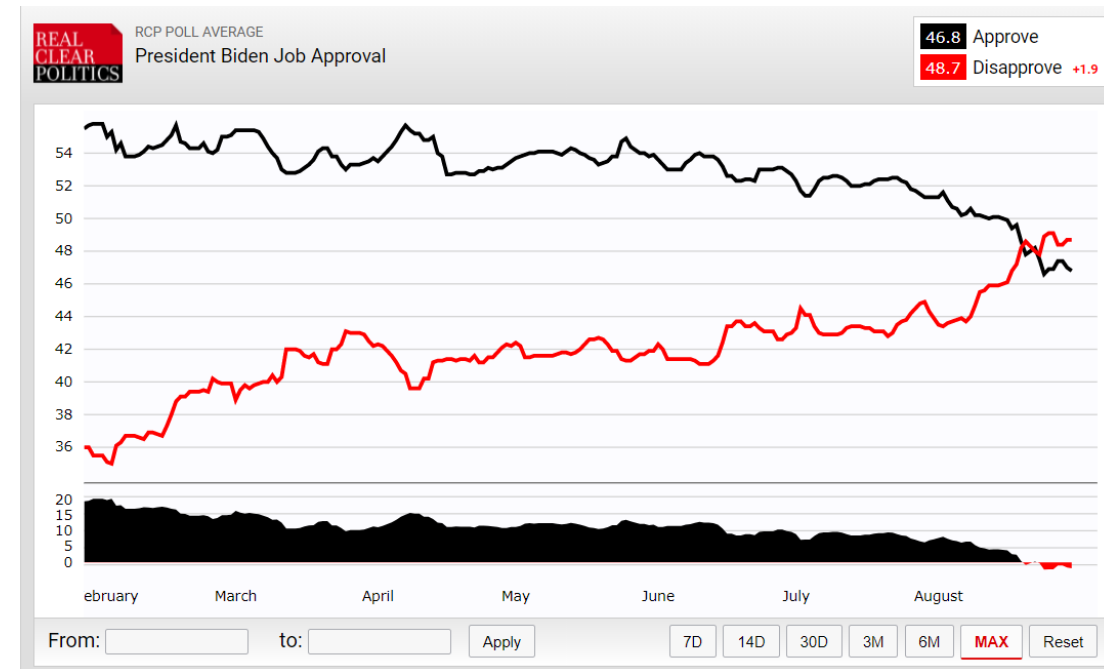
Note: Response rate is AAPOR RR3. Only landlines sampled 1997-2006. Rates are typical for surveys conducted in each year.

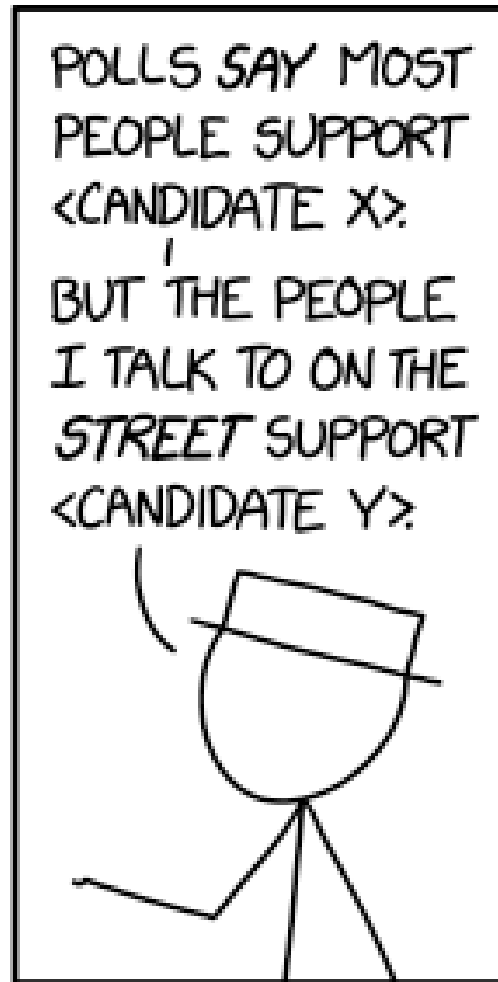
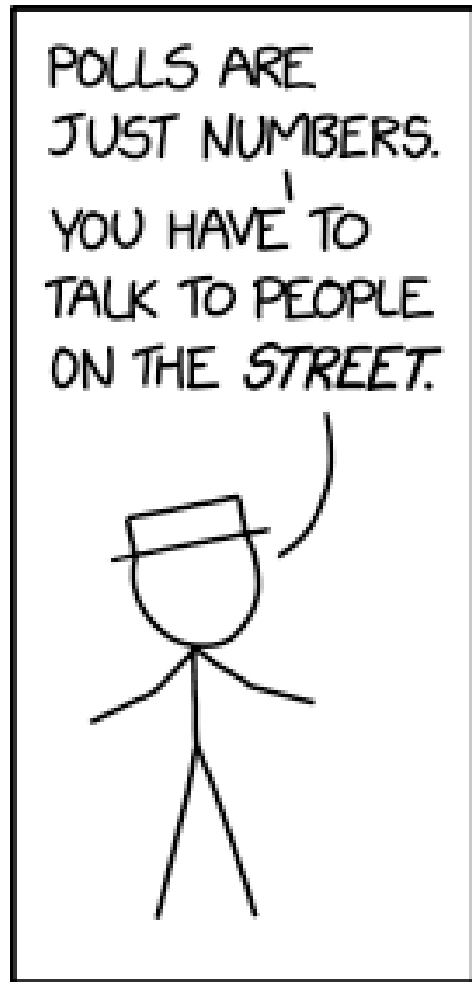
Source: Pew Research Center telephone surveys conducted 1997-2018.

PEW RESEARCH CENTER

Differential non-response is everything

- Differential non-response shows up everywhere you're gathering opinions
- Your training data for whatever model you train faces the same issue!
- Standard "margin of error" calculations do not take this into account
- Differential non-response *over time* often explains "swings" in polls!





Other pollsters complain about declining response rates, but our poll showed that 96% of respondents would be 'somewhat likely' or 'very likely' to agree to answer a series of questions for a survey.

Reminder: the task

- Each person j has an opinion, $Y_j \in \{0, 1\}$
- We want to measure $\bar{y} = E[Y_j]$, the population mean opinion on some issue
- Each person also has covariates, x_j^k
- We also may care about *conditional* means
 $E[Y_j \mid \text{ORIE program}]$

Example:

“Do you like the class so far?”

Options: “yes” and “no”

\bar{y} : What fraction of people like the class so far?

Degree program, whether you like waking up at 9:30, etc

Fraction of people in ORIE who like the class

Challenge 1: people don't give "true" opinion

People gave you \tilde{Y}_j , instead of Y_j

You lie because
you want a better
grade

$$\hat{y} = \frac{1}{N} \sum_j \tilde{Y}_j$$

\hat{y} does not converge to \bar{y} , *unless errors cancel out*

Challenge 2: Sample doesn't represent pop

- For each person j , let $A_j \in \{0,1\}$ be whether they answered
- You have $\mathbf{Y} = \{(A_j, Y_j)\}_{j=1}^N$, if called N people
Where $Y_j = \emptyset$ if $A_j = 0$ (they did not answer)
- Again, you do

$$\hat{y} = \frac{1}{|\{j \mid A_j = 1\}|} \sum_{j \in \{j \mid A_j = 1\}} Y_j$$

where $\{j \mid A_j = 1\}$ denotes the set of people who answered
and so $|\{j \mid A_j = 1\}|$ is the number of people who answered

\hat{y} does not converge to \bar{y} unless Y_j and A_j are uncorrelated

Some people
don't answer
when asked

Plan for rest of the day

Methods for 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 to college
Half the population went to college
Whether MEng or MS degree

- Suppose whether people answer is correlated with education

$$\Pr(A_j = 1) = \begin{cases} 0.1 & \text{if } x_j = 0 \\ 0.4 & \text{if } x_j = 1 \end{cases}$$

Whether you answer is correlated with degree program

- Education also correlated with opinion Y_j in some unknown manner

- We want to measure $\bar{y} = E[Y_j]$, the population mean

- *No other correlations between whether they answer and opinion:*

Opinion Y_j is independent of whether they respond A_j , conditional on x_j

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

New notation

- Number of people *called*:
- Population response rate for group ℓ :
- Population mean response for group ℓ :
- Population fraction for group ℓ :
- Corresponding *sample* values are:

N # of people in class
 A^ℓ Response fraction in degree ℓ
 \bar{y}^ℓ "Likes class" fraction in degree ℓ
 P^ℓ Fraction of class in degree ℓ

(i.e., $N\hat{P}^\ell\hat{A}^\ell = |\{j \mid A_j = 1, x_j \text{ in Group } \ell\}|$)

$\hat{A}^\ell, \hat{y}^\ell, \hat{P}^\ell$
 # who answered in degree ℓ

and so:

$$\bar{y} = \frac{P^0\bar{y}^0 + P^1\bar{y}^1}{P^0 + P^1} = P^0\bar{y}^0 + P^1\bar{y}^1 = 0.5\bar{y}^0 + 0.5\bar{y}^1$$
in example
True mean opinion

$$\hat{y}_{naive} = \frac{\hat{A}^0\hat{P}^0\hat{y}^0 + \hat{A}^1\hat{P}^1\hat{y}^1}{\hat{A}^0\hat{P}^0 + \hat{A}^1\hat{P}^1} \rightarrow \frac{A^0P^0\bar{y}^0 + A^1P^1\bar{y}^1}{A^0P^0 + A^1P^1} = 0.2\bar{y}^0 + 0.8\bar{y}^1$$
Mean poll response

Naïve method in more detail

$$\begin{aligned}\hat{y}_{naive} &= \frac{\left(\sum_{j \in \{j \mid A_j=1, x=0\}} Y_j + \sum_{j \in \{j \mid A_j=1, x=1\}} Y_j\right)}{|\{j \mid A_j = 1, x = 0\}| + |\{j \mid A_j = 1, x = 1\}|} \\ &= \frac{\hat{A}^0 \hat{P}^0 \hat{y}^0 + \hat{A}^1 \hat{P}^1 \hat{y}^1}{\hat{A}^0 \hat{P}^0 + \hat{A}^1 \hat{P}^1} = \frac{(\#(Y_j=1) \text{ from Group 0} + \#(Y_j=1) \text{ from Group 1})}{\text{Total Respondants}} \\ &\rightarrow \frac{P^0 A^0 \bar{y}^0 + P^1 A^1 \bar{y}^1}{P^0 A^0 + P^1 A^1} \neq \bar{y} \text{ unless } A^0 = A^1\end{aligned}$$

$P^0 A^0 / (P^0 A^0 + P^1 A^1)$ 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 homogenous *within* groups

(Exactly the setup we have)

Y_j is independent of A_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 frequency, 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)

$$\hat{y} = \frac{\left(\sum_{j \in \{j \mid A_j=1, x=0\}} Y_j + \sum_{j \in \{j \mid A_j=1, x=1\}} Y_j \right)}{|\{j \mid A_j = 1, x = 0\}| + |\{j \mid A_j = 1, x = 1\}|}$$
$$= \frac{(\#1 \text{ from group 0} + \#1 \text{ from group 1})}{\text{Total Respondants}}$$
$$\rightarrow \frac{N^0 A^0 \bar{y}^0 + N^1 A^1 \bar{y}^1}{N^0 A^0 + N^1 A^1} = \bar{y} \text{ if } \frac{N^0 A^0}{P^0} = \frac{N^1 A^1}{P^1}$$

Calling more in
ratio of non-
response

Now
 $N^\ell \hat{A}^\ell$ instead of
 $N \hat{P}^\ell \hat{A}^\ell$

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 $A^0 = A^1$ (and $N^0 = N^1$)

Stratification in practice

- You often don't know group specific response rates A^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 A_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{y}^ℓ of the true opinion \bar{y}^ℓ

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

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

If $W^\ell = P^\ell$ and $\hat{y}^\ell \rightarrow \bar{y}^\ell$, then $\hat{y} \rightarrow \bar{y}$

Details differ in how to construct estimate \hat{y}^ℓ , how to calculate weight W^ℓ , and what groups ℓ to consider

Naïve Weighting

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

$$\hat{y}^\ell = \frac{\sum_{j \in \{j \mid A_j = 1, x = \ell\}} Y_j}{|\{j \mid A_j = 1, x = \ell\}|}$$

(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 A_j , conditional on group ℓ

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

Problem 1: Estimate $\hat{y}^\ell = \frac{\sum_{j \in \{j \mid A_j=1, x=\ell\}} Y_j}{|\{j \mid A_j=1, x=\ell\}|}$ might be really bad

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

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

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”)
 “ $W^\ell = P^\ell$ 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 “ $W^\ell = abcd$ 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 $\hat{y}^\ell = \frac{\sum_{j \in \{j \mid A_j=1, x=\ell\}} Y_j}{|\{j \mid A_j=1, x=\ell\}|}$ might be really bad

Too few respondents in a group \rightarrow 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
- “Multi-level regression with post-stratification” (MRP)
Main idea: Train a (Bayesian) regression model to get estimate \hat{y}^ℓ for each set of covariates. Then, “post-stratify” by weighting \hat{y}^ℓ by population fraction P^ℓ
For groups with many samples, estimate \hat{y}^ℓ just based on that group; otherwise, based on “neighboring” groups

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!

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 correlates 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 by 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 A_j .
- You don't observe u_j for any individual j
- u is the only unmeasured confounding: A_j is uncorrelated with true opinion Y_j given u_j -- but we don't have u_j
- You have an estimate \hat{y} (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*

$$\hat{y} - \bar{y} \rightarrow (\tilde{P}^1 - P^1) (E[Y_j | u_j = 1] - E[Y_j | u_j = 0])$$

More detail: Notation and Insight

- True population fractions of u : $P^1 = \Pr(u_j = 1)$, $1 - P^1 = \Pr(u_j = 0)$
- Response fractions: $\tilde{P}^\ell = \Pr(u_j = \ell | A_j = 1)$
- $\bar{y} \stackrel{\text{def}}{=} E[Y_j] = P^1 E[Y_j | u_j = 1] + (1 - P^1) E[Y_j | u_j = 0]$
- $\hat{y} \rightarrow E[Y_j | A_j = 1] = \tilde{P}^1 E[Y_j | u_j = 1, A_j = 1] + (1 - \tilde{P}^1) E[Y_j | u_j = 0, A_j = 1]$
- Insight:

$$E[Y_j | u_j = \ell, A_j = 1] = E[Y_j | u_j = \ell]$$

“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

$$\begin{aligned}\bar{y} &= P^1 E[Y_j | u_j = 1] + (1 - P^1) E[Y_j | u_j = 0] \\ \hat{y} &\rightarrow \tilde{P}^1 E[Y_j | u_j = 1] + (1 - \tilde{P}^1) E[Y_j | u_j = 0]\end{aligned}$$

Rearrange:

$$\begin{aligned}\hat{y} &\rightarrow \bar{y} + (\tilde{P}^1 - P^1) E[Y_j | u_j = 1] + (P^1 - \tilde{P}^1) E[Y_j | u_j = 0] \\ &= \bar{y} + (\tilde{P}^1 - P^1) (E[Y_j | u_j = 1] - E[Y_j | u_j = 0])\end{aligned}$$

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

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?