

ORIE 5355: Applied Data Science -  
Decision-making beyond Prediction  
Lecture 4: Weighting + other topics in Data Collection

Nikhil Garg

# Announcements

- HW1 due 9/13
- Quiz 1 will also be that week
- No class on Monday (Labor day weekend)
- Guest lecture next class – please attend in person
  - Dhrumil Mehta, Columbia University (formerly 538)

Questions from last time?

# Plan for today

- Weighting techniques
- 1 slide on quantifying uncertainty
- Other topics
  - Data collection case studies beyond polling  
Ratings + Recommendations
  - Other topics in data collection  
Differential privacy, Bias, Eliciting complex opinions,  
Modeling opinion dynamics
- Module summary + questions

Weighting

# Stratification summary: change who you call

- Suppose you have  $L$  mutually exclusive demographic groups
- There are  $N^\ell$  people in group  $\ell$  in the population you care about
- Each group  $\ell$  has group response rate  $A^\ell$
- Call number of people in each group proportional to  $N^\ell / A^\ell$

This reduces bias if group response rates are different across groups

Always reduces variance caused by sample groups not matching population groups

# 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  $\ell$ , i.e., get an estimate  $\hat{y}^\ell$  of the true opinion  $\bar{y}^\ell$

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

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

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  $\hat{y}^\ell$ , how to calculate weight  $W^\ell$ , and what groups  $\ell$  to consider



# Naïve Weighting

**Step 1:** Use the mean response for each group  $\ell$  separately, i.e.

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

**Step 2:** Weight  $W^\ell$  is our best guess of true population fraction  $P^\ell$  for group  $\ell$

# Complication: How many groups/which ones?

- If group too broad (e.g., group  $\ell$  just gender), then break cardinal rule:  
Need: Opinion  $Y_j$  is independent of whether they respond  $A_j$ , conditional on group  $\ell$

- 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  $P^\ell$

# 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  $\ell$  just based on gender); we give you  $P^\ell$
- Then we define groups based on 2 covariates; we give you  $P^\ell$
- Then we define groups based on 2 covariates and ask you to construct  $P^\ell$  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  $\hat{y}^\ell$  for each set of covariates. Then, “post-stratify” by weighting  $\hat{y}^\ell$  by population fraction  $P^\ell$   
For groups with many samples, estimate  $\hat{y}^\ell$  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

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

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

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*
  - “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


# Case study: Ratings and recommendations



# Overview

- So far, we've talked about explicit opinion collection in polling
- The same challenges apply in other settings
- Some differences
  - Often we don't care about "absolute" opinion but "relative" opinions
  - We care a lot about "heterogeneous" opinions
  - We often have other "implicit" data on people's opinions
- Briefly discuss some of these challenges in context of ratings and recommendations

# Rating systems



**Detailed Seller Ratings** (last 12 months)

Criteria	Average rating
Item as described	★★★★★
Communication	★★★★★
Shipping time	★★★★★
Shipping and handling charges	★★★★★



### Customer Reviews


★★★★★ 4  
4.6 out of 5 stars ▾

5 star	<div style="width: 75%;"><div style="background-color: #FFD700; height: 10px;"></div></div>	75%
4 star	<div style="width: 25%;"><div style="background-color: #FFD700; height: 10px;"></div></div>	25%
3 star	<div style="width: 0%;"><div style="background-color: #FFD700; height: 10px;"></div></div>	0%
2 star	<div style="width: 0%;"><div style="background-color: #FFD700; height: 10px;"></div></div>	0%
1 star	<div style="width: 0%;"><div style="background-color: #FFD700; height: 10px;"></div></div>	0%

[See all verified purchase reviews ▸](#)

Share your thoughts with other customers

Write a customer review



**Private Feedback**  
This feedback will be kept anonymous and never shared directly with the freelancer. [Learn more](#)

Reason for ending contract:  
Please select...

Would you hire this freelancer again, if you had a similar project?  
 Definitely Not  Probably Not  Probably Yes  Definitely Yes

---

**Public Feedback**  
This feedback will be shared on your freelancer's profile only after they've left feedback for you. [Learn more](#)



**Feedback to Freelancer**

- ★★★★★ Skills
- ★★★★★ Quality of Work
- ★★★★★ Availability
- ★★★★★ Adherence to Schedule
- ★★★★★ Communication
- ★★★★★ Cooperation

Total Score: **0.00**


Share your experience with this freelancer to the oDesk community:

[See an example of appropriate feedback](#)

AT&T Park


2nd & King



★★★★★

What did you love about Samantha?


*Comments are anonymously shared with drivers.*



**14 Reviews** ★★★★★

Summary	Accuracy	★★★★★	Location	★★★★★
	Communication	★★★★★	Check In	★★★★★
	Cleanliness	★★★★★	Value	★★★★★

Translate reviews to English



Great location next to République stop. Nice communication from the

# Measurement error: Ratings Inflation

4.68★  
**DRIVER RATING**  
 Unfortunately, your driver rating last week was **below average.**

**UBER lyft**

★ ★ ★ ★ ★

**DON'T FORGET TO RATE 5 STARS**

FACT: WHEN A DRIVER'S RATING FALLS BELOW 4.7 THEY BECOME DEACTIVATED.

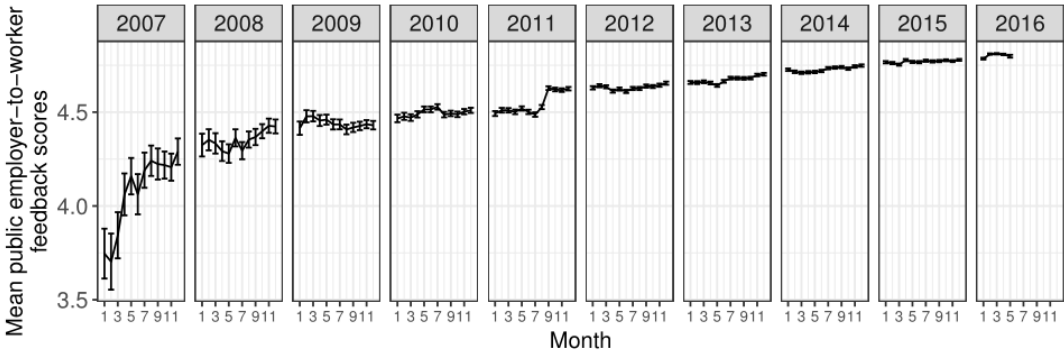
**MORE DRIVERS MEANS LESS SURGES**

OK, could be better

★ ★ ★ ★ ☆

How can Sajid improve?

Figure 2: Monthly average public feedback scores assigned to workers by employers on completed projects.



[Filippas, Horton, Golden 2017]

**When 4.3 Stars Is Average: The Internet's Grade-Inflation Problem**

THE WALL STREET JOURNAL The Wall Street Journal April 5, 2017

**UNDERSTANDING ONLINE STAR RATINGS:**

- ★★★★★ [HAS ONLY ONE REVIEW]
- ★★★★★ EXCELLENT
- ★★★★☆ OK
- ★★★★☆
- ★★★★☆
- ★★★☆☆
- ★★★☆☆
- ★★★☆☆
- ★★★☆☆
- ★★★☆☆ CRAP
- ★★★☆☆
- ★★★☆☆

<https://xkcd.com/1098/>

# Why ratings inflation & what to do about it?

- Many hypotheses for why ratings inflate
  - Explicit pressure from sellers – worry about retaliation
  - Implicit pressure – don't want to hurt people's livelihoods
  - Either misreport, or selection – less likely to report after bad experience
- Inflation is a type of measurement error:
  - The “quality” scale doesn't match well to the “rating” scale
  - Inflation over time – mapping from quality to rating changes over time
  - Why does it matter? We ask you this in the homework
- What to do about it:
  - Try to reduce some of the pressure
  - Weighting to tackle selection: paper in the homework: [Nosko & Tadelis]
  - Change the rating scale: [Garg and Johari]

# Experiment Description

**Status quo:** Clients hire freelancers, rate them at contract end

Form includes a numeric rating from 0 to 10, with avg >8/10

**Challenge:** Can we induce different (non-inflated) ratings by changing the question we ask on the rating form?

## Experiment design

- Add additional question to private portion of the form (6 treatments)  
Randomization at the *client* level
- Observe ratings for 3 months (180k jobs, 60k clients, 80k freelancers)

# Treatment groups

Treatment	Question Phrasing	Answer choices
Numeric	How would you rate this freelancer overall?	0 – 5

# Treatment groups

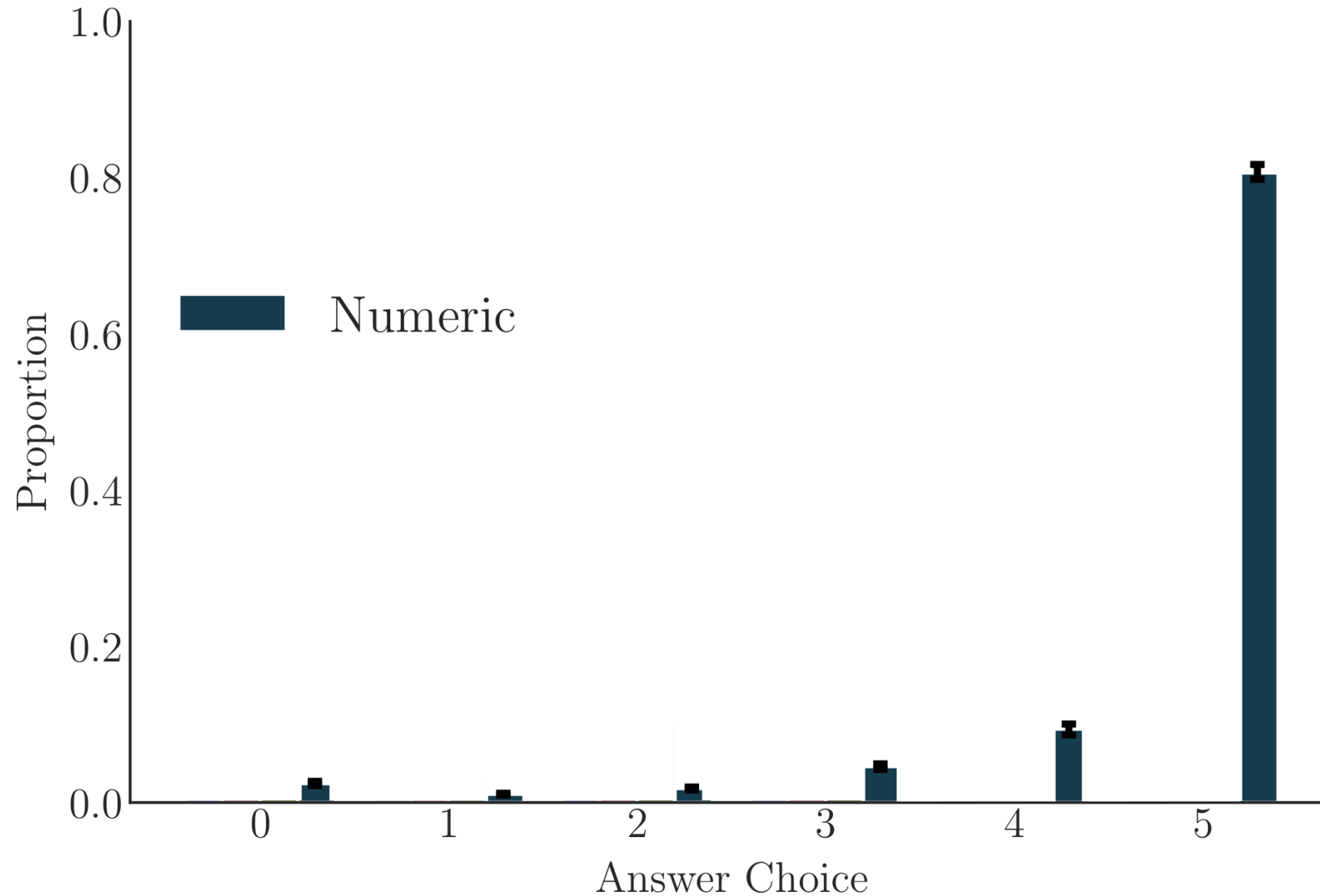
Treatment	Question Phrasing	Answer choices
<b>Numeric</b>	How would you rate this freelancer overall?	0 – 5
<b>Adjectives</b>	How would you rate this freelancer overall?	Terrible Mediocre Good Great Phenomenal Best possible freelancer!

# Treatment groups

Treatment	Question Phrasing	Answer choices
<b>Numeric</b>	How would you rate this freelancer overall?	0 – 5
<b>Adjectives</b>	How would you rate this freelancer overall?	Terrible Mediocre Good Great Phenomenal Best possible freelancer!
<b>Expectations</b>	How did this freelancer compare to your expectations?	Much worse than I expected ... Beyond what I could have expected
<b>Average</b>	How does this freelancer compare to others you have hired?	Worst Freelancer I've Hired Below Average Average Above Average Well Above Average Best Freelancer I've hired
<b>Average, random order</b>		
<b>Average, not affect score</b>		

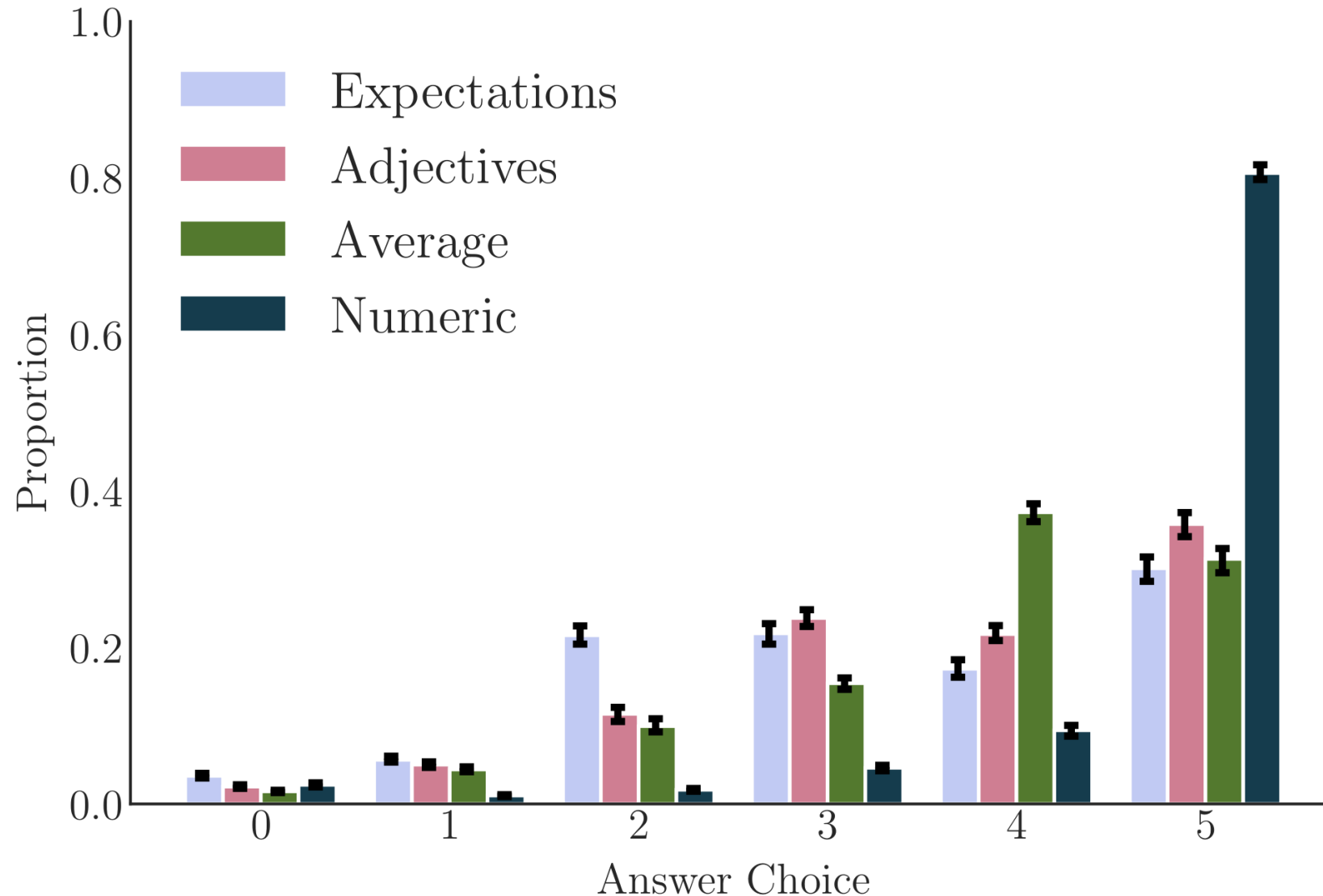


# Result: marginal rating distributions



“Designing Informative  
Rating Systems:  
Evidence from an  
Online Labor Market”  
Nikhil Garg and  
Ramesh Johari

# Result: marginal rating distributions



“Designing Informative Rating Systems: Evidence from an Online Labor Market”  
Nikhil Garg and Ramesh Johari

# Ratings heterogeneity

- There is much ratings “heterogeneity”
  - Different people have different opinions on the same item
  - Different ‘categories’ of items might have different average ratings
- Why does this matter?
  - You want to give each person a personalized “rating” or recommendation
  - You want to compare items across categories
- What to do about it?
  - Personalized recommendations → starting next time
  - “Standardize” ratings across categories
  - Communicate to customers – e.g., “relative” ratings instead of “absolute” ones

# Implicit data collection in recommendations

- You have many implicit signals about people's opinions
  - Do they finish watching the show, or start watching the next episode?
  - Do they keep coming back and buying other things
  - Did they browse other items instead of putting something in their cart?
  - Do they re-hire the same freelancer/work with the same client again?
- These give *different* information than do explicit ratings
  - From a different population of users
  - Often more numerous, but harder to analyze
  - “revealed preference” – might be more predictive of future behavior
- Using such data
  - Train models to predict different future behavior, using various signals
  - Might take away “user agency” – what if they want to change their behavior?

Miscellaneous topics in data and  
data collection

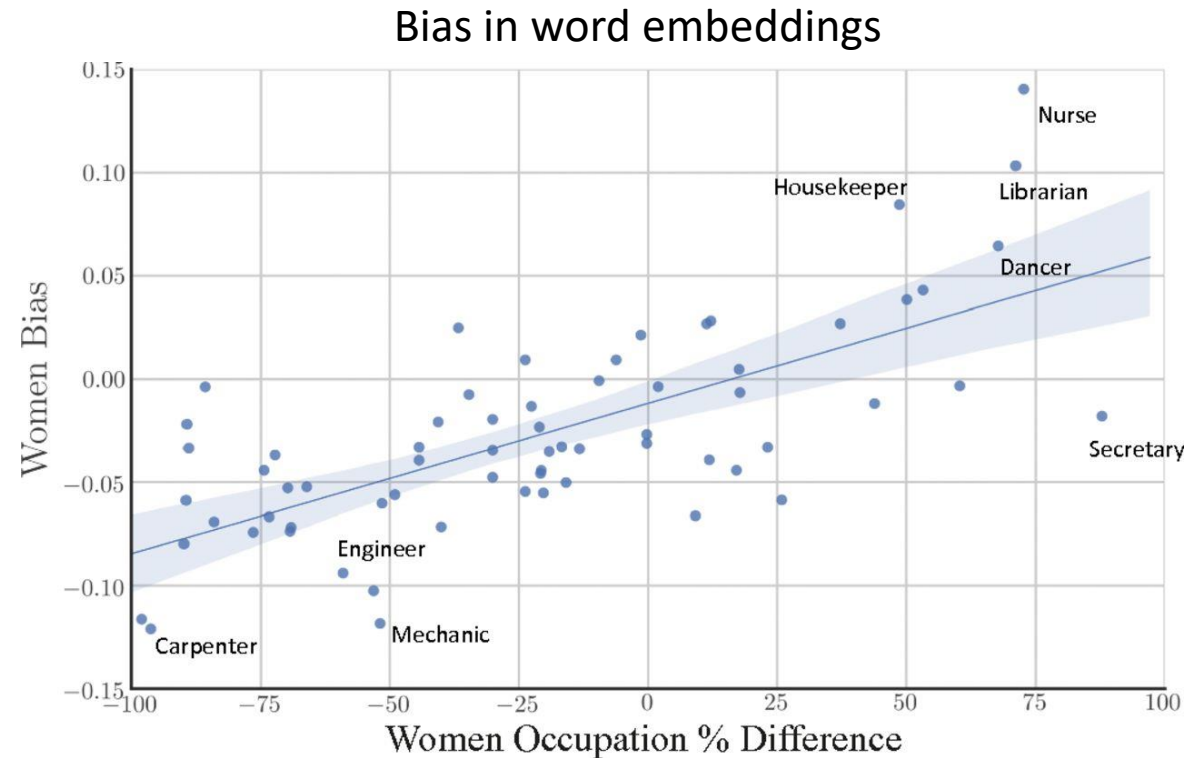


# (Differential) Privacy

- What if you're asking about a sensitive attribute?  
For example, an insurance company wants to estimate the percentage of their policy holders who smoke
- Goal: collect data in a way such that you learn very little about any individual person, but you are accurate across population
- How? Add noise to each response
- Example: Tell each person, "roll a 6-sided dice. If it's 1 or 2, lie about whether you smoke. Otherwise, tell the truth." If fraction  $Y$  people tell you that they smoke, then we know that the truth  $X$  satisfies:  
$$Y = \frac{4}{6}X + \frac{2}{6}(1 - X)$$
- Similar ideas used to collect and share data at Apple and the US Census

# Using biased data

- The world is full of historic inequities
  - Some neighborhoods are over-policed compared to others → data will have more “crimes there”
  - Every possible opinion expressed on forums like Reddit
  - Who succeeded at a university
- Models trained using this data will *reflect* and *amplify* these biases
- Many techniques to audit and mitigate such biases in models



“Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes” by Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou

# Eliciting complex opinions

- So far, we've talked about soliciting “low-dimensional” opinions
  - Binary opinions, or one of a small number of options
- What if we want to solicit opinions on complicated things?
  - How your town should spend \$2M budget across parks, sports teams, art festivals, etc.
  - When should we schedule these five events over 10 time slots?
- You can't ask people to rank every option
- Several standard techniques
  - Participatory budgeting
  - Pairwise comparisons
- More generally, many cool techniques in crowdsourcing



# Data dynamics

- The world is not static
  - Opinions change with external events
  - Your startup is growing and attracting new kinds of customers
  - Weekends are different than weekdays, except on holidays...
- Similar problem as “Problem 1” in survey weighting – if you don’t share data across time, then you don’t have enough data. But if you do share data, then suddenly your dataset differs from what you care about
- Techniques to model opinion dynamics – “smooth” over time
- Some related challenges covered in pricing module

# Module Summary

- Measurement error: The construct you care about is never perfectly captured by the data that you have
- Selection effects/differential non-response happens everywhere you're collecting opinions from people
- You can use stratification and weighting to mitigate selection effects *on known covariates*
- On unknown covariates, quantify uncertainty!

Never take opinion data at face value. Always ask:

- (1) What did I measure, versus what did I care to measure?
- (2) Who answered versus what's the population of interest
- (3) What am I going to *do* with the data, and how does that affect data collection?

Will show up in the rest of the course!

Questions?