Quantifying the Uncertainty of Imputed Demographic Disparity Estimates

CRIW Pre-Conference, Summer 2023

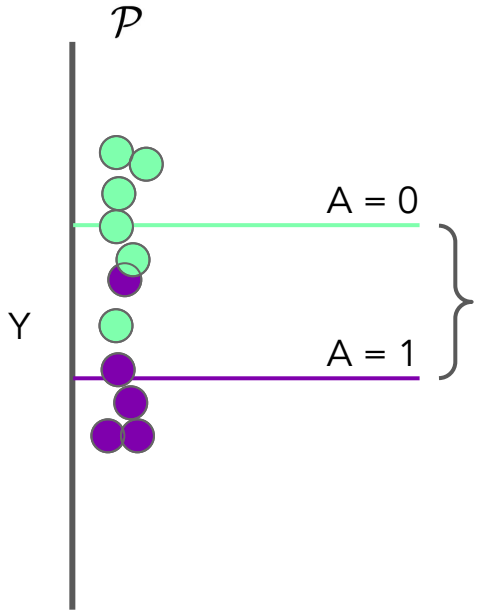
Benji Lu
Yale University
UC Berkeley

Jia Wan
University of Oxford

Jacob Goldin
University of Chicago
NBER

Daniel E. Ho
Stanford University

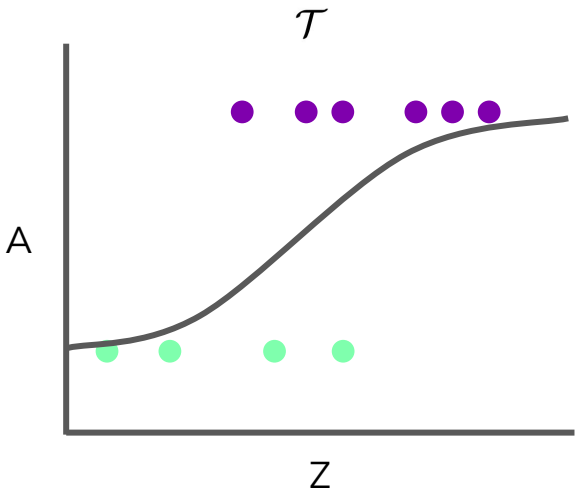
Estimand: Demographic Disparity

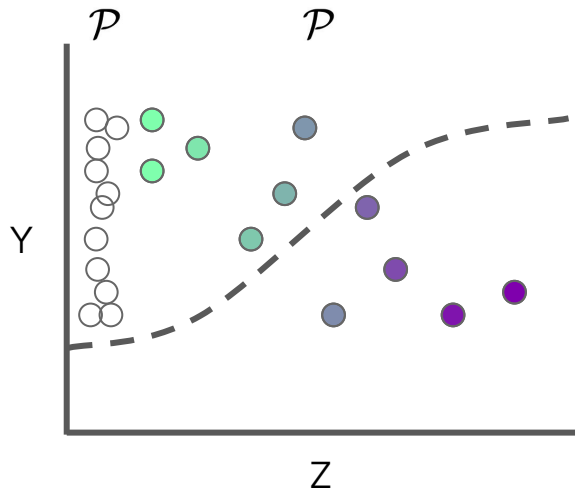


The Probabilistic Estimator

$$\hat{\delta} = \frac{\sum_{i=1}^{n_{\mathcal{P}}} \widehat{\Pr}_{\mathcal{P}}(A_i = 1|Z_i)Y_i}{\sum_{i=1}^{n_{\mathcal{P}}} \widehat{\Pr}_{\mathcal{P}}(A_i = 1|Z_i)} - \frac{\sum_{i=1}^{n_{\mathcal{P}}} \widehat{\Pr}_{\mathcal{P}}(A_i = 0|Z_i)Y_i}{\sum_{i=1}^{n_{\mathcal{P}}} \widehat{\Pr}_{\mathcal{P}}(A_i = 0|Z_i)}$$

Chen et al. (2019)

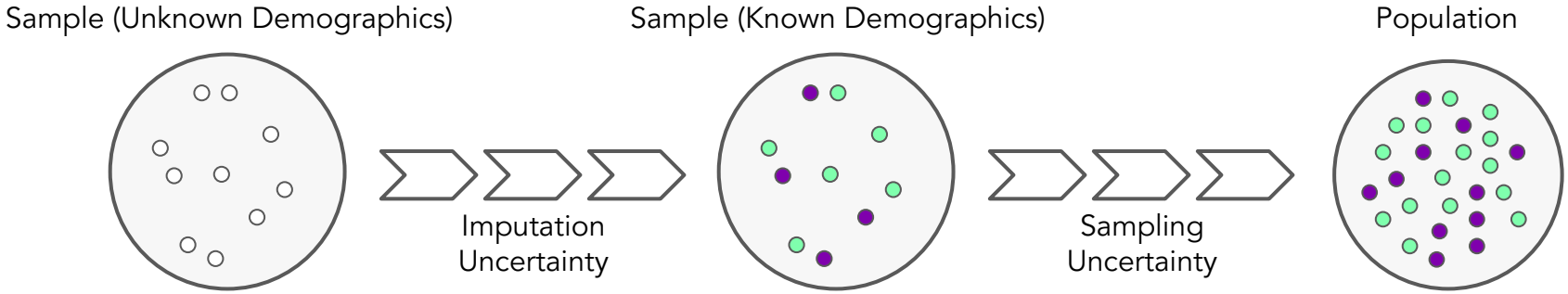




What is the uncertainty of this estimator?

Our Work: Uncertainty Quantification

Prior work treats the demographic imputations as certain, leading to overconfidence in the results.



We propose a dual-bootstrap procedure that incorporates this uncertainty, offering more accurate uncertainty quantification.

The Dual Bootstrap

Algorithm: Dual Bootstrap

Data: Primary set $\mathcal{P} = \{Z_i, Y_i\}_{i=1}^{n_\mathcal{P}}$, training set $\mathcal{T} = \{Z_j, A_j\}_{j=1}^{n_\mathcal{T}}$
Parameters: Model class $\mathcal{F}_A$, number of bootstrap samples B
Result: Confidence interval for the demographic disparity δ

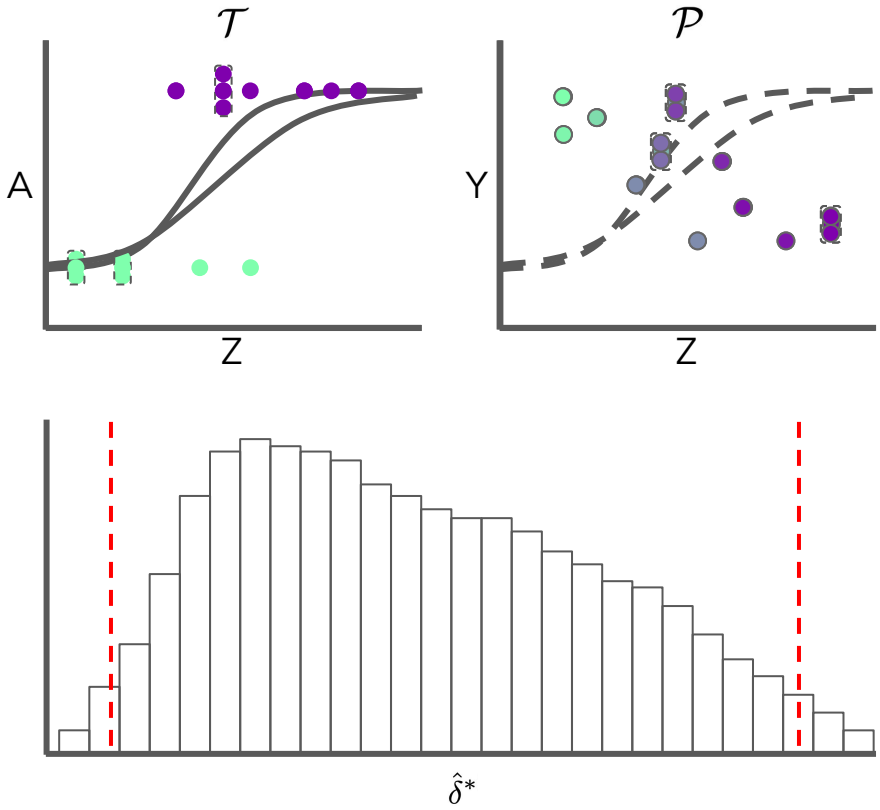
for $b \in \{1, \ldots, B\}$ **do**
  Obtain $\mathcal{T}^{*b}$ by sampling with replacement from $\mathcal{T}$;
  Fit $\widehat{\Pr}_\mathcal{P}^{*b}(A = 1|Z) \in \mathcal{F}_A$ on $\mathcal{T}^{*b}$;
  Obtain $\mathcal{P}^{*b}$ by sampling with replacement from $\mathcal{P}$;
  Compute the bootstrap statistic

$$\hat{\delta}^{*b} = \frac{\sum_{i=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathcal{P}^{*b}(A_i^{*b} = 1|Z_i^{*b}) Y_i^{*b}}{\sum_{i=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathcal{P}^{*b}(A_i^{*b} = 1|Z_i^{*b})} - \frac{\sum_{i=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathcal{P}^{*b}(A_i^{*b} = 0|Z_i^{*b}) Y_i^{*b}}{\sum_{i=1}^{n_\mathcal{P}} \widehat{\Pr}_\mathcal{P}^{*b}(A_i^{*b} = 0|Z_i^{*b})};$$

end
Output the $(1-\alpha)$-level percentile bootstrap confidence interval

$$\left(\hat{\delta}_B^{(\alpha/2)}, \hat{\delta}_B^{(1-\alpha/2)}\right)$$

where $\hat{\delta}_B^{(\alpha)}$ is the empirical α-percentile of the $\hat{\delta}^{*b}$.

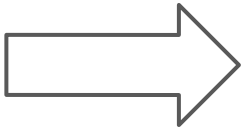


The Dual Bootstrap: BISG

$$\widehat{\Pr}^{*}(\text{Race}|\text{Geography, Surname}) \propto \widehat{\Pr}^{*}(\text{Race}|\text{Surname})\widehat{\Pr}^{*}(\text{Geography}|\text{Race})$$

American Community Survey

Surname	Geography	Race
Brown	10001	Black
Brown	10002	White
Johnson	10001	White
Johnson	10001	Black
Smith	10001	White
Smith	10002	White
Smith	10002	Black
⋮	⋮	⋮



Surname	%White	%Black
Brown	58	36
Johnson	59	35
Smith	71	23

Race	%10001	%10002
Black	0.005	0.016
White	0.006	0.011

Resample the raw American Community Survey data and recompute priors.

Theory

Theorem: Consistency of Bootstrap Confidence Interval

- Fix $\alpha \in [0,1]$, $\lim_{n_{\mathcal{P}} \to \infty} \lim_{n_{\mathcal{T}} \to \infty} \lim_{B \to \infty} \Pr\left(\hat{\delta}_B^{(\alpha/2)} \leq \delta \leq \hat{\delta}_B^{(1-\alpha/2)}\right) \geq 1 - \alpha.$

Assumptions and Regularity Conditions

- Assume $\widehat{\delta}$ is asymptotically unbiased, i.e., $\mathbb{E}[\mathrm{Cov}_{\mathcal{P}}(A, Y|Z)] \to 0$
- The imputation model is correctly specified and consistent
  - As $n_{\mathcal{T}} \to \infty$, $\widehat{\Pr}_{\mathcal{P}}(A = 1|Z)$ converges to a deterministic function of Z that is the real conditional probability
- Standard regularity conditions sufficient for bootstrap consistency
  - $0 < \underline{c} \leq \widehat{\Pr}_{\mathcal{P}}(A = 1|Z) \leq \overline{c} < 1$; Y is bounded

Theory

Proof sketch:

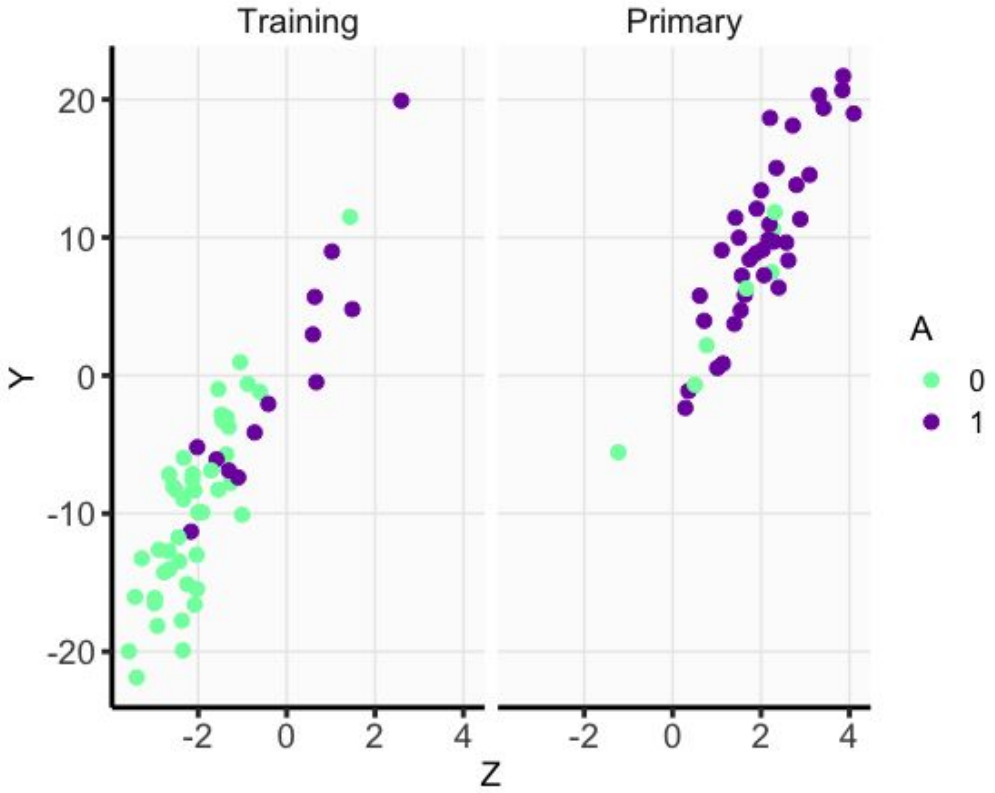
1. $\widehat{\delta}$ is a Z-estimator

$$\sum_{i=1}^{n} \psi\left(\widehat{\Pr}_{\mathcal{P}}(A_i = 1|Z_i), Y_i, \hat{\theta}_1, \hat{\theta}_2, \hat{\delta}\right) = 0$$

$$\text{where} \quad \psi\left(\widehat{\Pr}_{\mathcal{P}}(A_i = 1|Z_i), Y_i, \theta_1, \theta_2, \delta\right) = \begin{pmatrix} \widehat{\Pr}_{\mathcal{P}}(A|Z)(Y - \theta_1) \\ \left[1 - \widehat{\Pr}_{\mathcal{P}}(A|Z)\right](Y - \theta_2) \\ \theta_1 - \theta_2 - \delta \end{pmatrix}$$

2. Known regularity conditions for bootstrap consistency of Z-estimators
    - Master theorem (Kosorok ‘08)

Experiment #1: Regression imputation on simulated data



Primary

Z	~~A~~	Y

Training

Z	A	~~Y~~

	Dual Bootstrap	Single Bootstrap
Coverage (95%)	97%	66%

Experiment #2: BISG on Census data

Primary Dataset: American Family Cohort (AFC)
- Electronic health records of 7.7 million patients from 2010-present

Surname	Geography	1_{Obesity}

Training Dataset: American Community Survey (ACS)
- Census at the block group level

Surname	Geography	Race

$$\delta = \mathbb{E}[1_{\text{Obesity}} | \text{white}] - \mathbb{E}[1_{\text{Obesity}} | \text{non-white}]$$

	Resampling ACS only	Dual Bootstrap (resampling AFC and ACS)	Single Bootstrap (resampling AFC only)
Standard deviation of $\widehat{\delta}$ (S.E.)	0.2 bps (3.8e-8)	6.8 bps (1.5e-6)	6.7 bps (1.5e-6)

* 1 bps = 0.0001

Experiments: Regression-based vs. BISG

Simulated data with regression-based imputation

	Dual Bootstrap	Single Bootstrap
Coverage	0.97	0.66

Census data with BISG imputation

	Dual Bootstrap	Single Bootstrap
Standard deviation of $\widehat{\delta}$	6.8 bps	6.7 bps

Possible reason for the difference in dual bootstrap performance:

- If the training data size is large relative to the model complexity, then imputation uncertainty matters less.

Next Steps

- Different performances of ML/regression-based vs. BISG imputations.
- Choice of imputation on large dataset vs. small survey that measures all relevant features.
- Relaxing assumptions:
    - Model misspecification, sampling representativeness, asymptotic bias…

Thank you!

Questions?