Quantifying the Uncertainty of Demographic Disparity Estimates when Demographic Labels Are Unobserved

Extended Abstract

March 16, 2023

Abstract

Demographic disparities—the simple, unadjusted difference in average outcomes between two demographic groups—are a common starting point for equity assessments. But they can be challenging to obtain when the demographic labels are unobserved. One approach is to impute the missing demographic labels using proxies, then use those imputations to estimate the disparity. Sometimes, a separate dataset is used to train an imputation model relating the proxies to the demographic labels. We offer two contributions within this framework. First, we propose a dual-bootstrap approach for accurate uncertainty quantification of the disparity estimate. Conventional standard errors understate the disparity estimate's uncertainty because they treat the imputations as given and fixed, instead of unknown quantities that must be estimated with error. As a result, they produce improperly narrow confidence intervals and small p-values. By contrast, our dual-bootstrap approach explicitly accounts for imputation error, enabling inference that is valid asymptotically and more accurate in finite samples. Second, we offer practitioners reasonable heuristics to determine at the outset whether imputation is the advisable course of action when demographic labels are missing. Specifically, we help practitioners forecast the variance of the imputation-based estimator so that they can decide whether the expected level of certainty is worth the cost. We apply our methods to a novel, highly detailed dataset of over 7.7 million patients' electronic health records to estimate demographic disparities in health outcomes, which we verify against ground truth obtained by linking the records to U.S. Census demographic information.

1 Introduction

As our ability to collect and analyze data improves, decision systems are becoming more complex, automated, and opaque. Demand is growing to ensure that such systems do not inadvertently produce inequitable outcomes for historically underserved populations. Meeting this demand is of concern across public and private sectors. For example, on his first day in office, President Biden issued Executive Order 13985, which ordered federal agencies to assess the equity of their policies and actions. And in 2022, Meta entered into a settlement agreement with the U.S. Department of Justice to resolve allegations that its algorithm for serving personalized housing advertisements discriminated against users based on characteristics protected by the Fair Housing Act.

The growing demand to root out inequity raises complex and difficult questions about how best to define and measure it. One common starting point is the demographic disparity: the simple, unadjusted difference in average outcomes between two demographic groups. The demographic disparity is far from the final word on the matter. But it has a key methodological advantage: Usually, it is easy to compute.

Not so, however, if the demographic data are missing. And missing they often are, at a frequency that might surprise some observers. Take President Biden's Executive Order 13985, for example. Of the 25 agencies that responded to the order, 21 reported that they lacked the demographic data needed to compute demographic disparities, let alone do a more searching equity assessment. The lack of data presents a serious methodological hurdle—but one that might be surmountable, with research ongoing to appraise the problem and propose solutions. Our paper contributes to that growing body of work.

2 Framework

One solution to the problem of missing demographic data is to impute it based on proxy features, then use those imputations to estimate the demographic disparity. Methods of imputation vary. A common approach is Bayesian Improved Surname and Geocoding (BISG) (Elliott et al., 2009), which naively applies Bayes' rule to estimate a person's race or ethnicity based on Census information about the racial or ethnic makeup of people sharing the person's surname and geographic location. Another approach is to train machine learning models of the relationship between demographic labels and proxy features using data where both are observed, then apply those models to predict demographic labels for data where such labels are unobserved (e.g., Cheng et al., 2023).

With the imputations in hand, researchers can then use them to estimate the disparity between two demographic groups. Here, too, methods vary. One common approach is to compute a weighted difference in mean outcomes between the two demographic groups, where the weights are given by the estimated probabilities of belonging in a given demographic group. Chen et al. (2019) show that this estimator is consistent under an ignorability-type assumption. Another approach is to regress the outcomes on the estimated probabilities of belonging in a given demographic group using ordinary least squares. Elzayn et al. (2023) show that this estimator is consistent under a different ignorability-type assumption and relate it to the probability-weighted estimator of Chen et al. (2019) under certain conditions.

3 Methodological Contributions

We offer two contributions within the above framework for estimating demographic disparities when demographic features are unknown. First, we propose a "dual-bootstrap" procedure that better characterizes the uncertainty of the imputed demographic disparity estimates than traditional methods do, enabling more accurate statistical inference. Second, we offer heuristics to help researchers decide when the cost and effort of implementing an imputation-based solution is worthwhile relative to other possible solutions to the problem of missing demographic data.

3.1 Uncertainty Quantification

We propose a dual-bootstrap procedure that produces more accurate statistical inferences, like confidence intervals and p-values, for imputation-based disparity estimators. To our knowledge, the literature on uncertainty quantification within the imputation-based framework has taken the traditional approach of considering variation induced by random sampling only (see Miratrix et al., 2018). This ignores a key source of uncertainty: the imputations themselves. Treating the imputations as

given and fixed, instead of unknown quantities that must be estimated with error, understates the uncertainty of the final demographic disparity estimate, yielding overconfident inference.

Our dual-bootstrap procedure offers more accurate statistical inference by accounting for both sources of uncertainty. It calls for the analyst to bootstrap not only the data for which it seeks to estimate the demographic disparity, but also the data on which it bases the demographic imputations. The former accounts for sampling uncertainty; the latter accounts for imputation uncertainty. The method is straightforward to implement. And it can flexibly accommodate a wide range of imputation methods without requiring parametric assumptions. We prove that our dual-bootstrap procedure produces asymptotically valid statistical inference under standard regularity conditions. And we show via simulations that it produces more accurate inference than standard approaches do in finite samples.

3.2 Imputation Heuristics

We develop heuristics to guide researchers as they decide whether to use imputation-based solutions to overcome the problem of missing demographic data. Imputation-based solutions can be costly. They often require a second dataset on which to base the demographic imputations. This second dataset must be suitably related to the first, so that the imputations can be ported. The process of identifying, vetting, and acquiring such a dataset can take time and resources, and might require coordination within or across organizations. So too might the process of analyzing the second dataset and validating the imputation method. Moreover, efficacy is not guaranteed; the imputation-based approach might yield minimal insights for a number of reasons, ranging from insufficient sample size to uninformative demographic proxies. Other methods, like developing and administering a survey that collects complete data, might yield better results at equal or lower cost.

Our heuristics help researchers decide whether an imputation-based solution is worth the time and expense, relative to alternative solutions. Specifically, we compare the relative accuracy of two different ways of estimating a demographic disparity: (1) analyzing a large but incomplete dataset that requires imputation of missing demographic data, and (2) analyzing a small but complete dataset without any missing data. We distill the comparison down to a few key metrics: the sample sizes of the two datasets, the informativeness of the demographic proxies, and the variance of the outcome. Given these metrics, our heuristic roughly quantifies for researchers how much more accurate one approach is likely to be compared to the other. With this information in hand, the researcher can then decide which approach is more worth its cost.

4 Application

We apply our dual-bootstrap procedure to estimate demographic disparities in health outcomes using a novel dataset of over 7.7 million patients' electronic health records. Like Cheng et al. (2023), we use machine learning methods to impute patients' race based on the rich information contained in their health records. With those imputations, Cheng et al. (2023) then estimate demographic disparities in health outcomes like asthma and obesity and calculate conventional standard errors. Notably, their confidence intervals do not contain the ground-truth demographic disparities that they are able to directly compute using demographic data observed in a subset of the data. This might be because the observed demographic data is inaccurate, or because the subset is not representative of the broader dataset. But our work suggests that this might also partly be because their conventional standard errors do not account for imputation uncertainty. To rectify

these issues, we (1) link the electronic health records to U.S. Census data to obtain comprehensive and accurate demographic information, and (2) apply our dual-bootstrap procedure instead of using conventional standard errors, with the expectation that it will produce wider confidence intervals that might cover the true disparities.

References

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019). Fairness Under Unawareness: Assessing Disparity when Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 339–348, New York, NY, USA. Association for Computing Machinery.

Cheng, L., Gallegos, I., Ouyang, D., Goldin, J., and Ho, D. E. (2023). How Redundant Are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race Is Unobserved. *Preprint*.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. (2009). Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83.

Elzayn, H., Smith, E., Hertz, T., Ramesh, A., Fisher, R., Ho, D. E., and Goldin, J. (2023). Measuring and Mitigating Racial Disparities in Tax Audits.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth Weighting? How to Think About and Use Weights in Survey Experiments. *Political Analysis*, 26(3):275–291.