Measures of Disparity and their Efficient Estimation

Harvineet Singh
Rumi Chunara
hs3673@nyu.edu
rumi.chunara@nyu.edu
New York University
USA

ABSTRACT

Quantifying disparities among population groups is an important task in public health, economics, and sociology among other social sciences, and increasingly in machine learning. In this work, we study the problem of how to collect data to measure disparities. Prior work provides sample size calculations for a narrow set of disparity metrics, namely those that are differences or ratios of mean outcomes between groups. However, a variety of metrics are used in practice, many of which can be expressed as arbitrary functions of group means. For this general class of metrics, we derive the number of samples to be collected from each group for estimating disparities more precisely for a fixed sample size. Our analysis builds on approximations to the sampling distribution of the estimates which also guides sample size calculations for hypothesis tests asking if there are significant disparities. We apply the methods to a machine learning dataset to evaluate a model's fairness and to two nationwide surveys used for understanding population-level attributes like employment and health. Results show that the methods improve estimation error when groups have different variances. Absent any information on the groups, we find that sampling the groups equally performs well in practice.

KEYWORDS

disparity estimation, fairness metrics, optimal data collection; AI, health, and well-being, Social Sciences

1 INTRODUCTION

Measurement of disparities in outcomes, behaviors, and resources is essential to track progress towards mitigating inequities. For instance, the Healthy People initiative in the United States (US) tracks disparities in a number of health outcomes to guide actions towards achieving health equity [41]. Taking an example from this initiative, the infant death rate among non-Hispanic black mothers was 2.625 times the infant death rate for Asian or Pacific Islander mothers (best performing group) in 2011 [41, Table 5]. Even the *fairness metrics* in the fair machine learning literature are an instance of disparity measured among model outputs for population groups [29]. Given its vast uses, quantification of disparity has been an important object of study in many disciplines [18, 26, 40].

Disparities can be quantified in many ways. The choice of which measure to use (such as absolute difference vs ratio or how to weigh different groups' data) makes normative assertions which influence interpretation of the results [17, 34]. For this reason, we look at a broad class of disparity measures in this work. Examples include the commonly used difference or ratio of mean outcomes for the two groups as well as variance and entropy of the mean outcomes. Selecting an appropriate metric for a given application is an important task [18], however, it is out of scope of our work.

There are many data-related challenges to quantifying disparities. Collecting data randomly from a population may not sufficiently include minority groups. Further measurements may be more variable or noisier for some groups. These challenges of data scarcity and hetereogeneity across groups remain even when evaluating disparities (unfairness) of algorithms [25], especially when considering intersectional group definitions [42]. This motivates groups to be differentially sampled depending on their size and data quality to get precise disparity estimates. However, existing methods for calculating sample sizes do not explicitly address disparity measures [14, 35].

Accordingly, we address the important question of **how to collect data to precisely measure a given disparity measure**, see Figure 1. Precise estimates of disparity can provide the much needed evidence while advocating for inequity-reducing policies or tracking their progress. Precision is desirable particularly when analyzing the trend of disparities over time as estimates with large confidence intervals may hide whether the disparity is increasing, decreasing, or constant. Understanding how to best measure disparity with limited data can also be beneficial to get initial information if disparities exist between two places or groups of people [20].

To increase efficiency of disparity estimates, we focus on a survey design method known as stratified sampling. Here, the population is divided into multiple strata, such as by geography or race/ethnicity-based groups in our context. For example, the health survey called Behavioral Risk Factor Surveillance System stratifies the US population by geographic location [12]. A random sample is collected for each stratum independently and the stratum-specific estimates are combined together to compute the metric of interest. Stratified sampling is typically studied for the problem of estimating the overall population mean, however, the idea is more widely applicable. In fact, it can be shown that optimizing the number of samples to measure from each stratum can result in estimates that always have better or same precision as sampling the population uniformly at random [46]. This optimal sample size allocation is known as Neyman allocation [30]. The key insight of our work is that estimation of the disparity metrics can be similarly optimized by stratified sampling. The literature on estimating average treatment effects (which is also a disparity measure between outcomes in treatment and control groups) show a similar application [16, 24, 37]. We extend this insight to a broad class of disparity metrics.

In summary, our contributions are as follows. We present a method to efficiently estimate a general class of disparity metrics, that include common metrics from fair machine learning [29] and health disparities literature [18], by tuning sample sizes collected

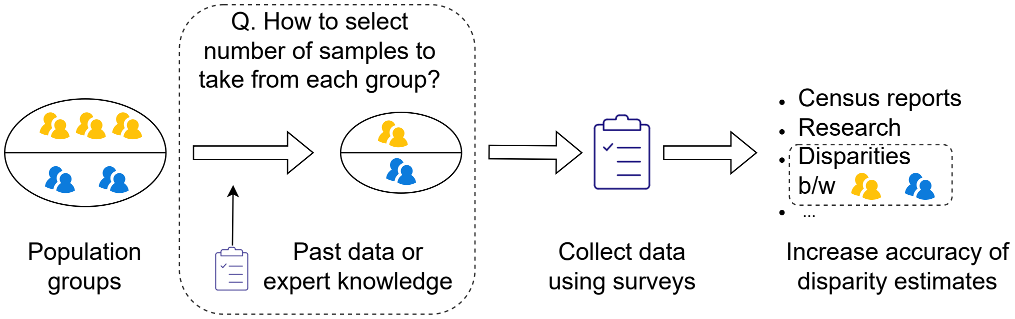

Q. How to select number of samples to take from each group?
Population groups
Past data or expert knowledge
Collect data using surveys
Census reports
Research
Disparities b/w
...
Increase accuracy of disparity estimates

Figure 1: Overview. Surveys such as nationwide census are often used to estimate disparities between population groups. Due to cost concerns, only a small number of units from each group can be sampled and measured. However, such sample size considerations are rarely tailored to the goal of measuring disparities. We study how to allocate a fixed number of samples among the groups to measure the given disparity metric as efficiently as possible.

per group. We apply the method to the problems of hypothesis tests for disparity and evaluating fairness of a prediction model. We demonstrate the method on two real world datasets and highlight scenarios when it improves accuracy of disparity estimates.

2 RELATED WORK

We review work on design of surveys from statistics, fair data collection from machine learning, and measuring disparities from public health and economics.

Survey design. A long line of work in survey statistics is dedicated to the design of surveys which involves, for instance, deciding the populations to study and sample sizes to collect subject to the given sampling resources and the analysis plan [14, Chapter 3]. For stratified sampling designs, the sample size allocations per stratum that minimize the variance of the estimate for a given cost is known as Neyman allocation. Such allocations are well known when the goal is to estimate the population mean outcome [30], average treatment effect [16], and ratio of group means [6]. Note that the average treatment effect is the same as the metric we call difference in means (between treatment and control groups in a randomized experiment). We derive Neyman allocations for a broader class of disparity metrics. The related problem of rare population sampling is typically addressed by sampling disproportionately from strata that have higher prevalence of rare populations [22].

Fair data sampling. Access to representative, high-quality data is important to training and evaluating fair machine learning models. In a survey of machine learning practitioners in industry, Holstein et al. [19] finds that better support for data collection is an unmet need for creating fair models. Approaches exist to guide data collection for *training* fairer models [1, 2, 4, 32, 38] which differs from our goal of *evaluating* fairness. Of note is Rolf et al. [36] which provides optimal sample sizes to collect from population groups to train accurate models building on scaling laws for group-specific losses. Yan and Zhang [44] gives an approach to collect labelled data to evaluate model fairness. However, the work is limited to one notion of fairness, namely demographic parity. Niss et al. [31] studies the problem of testing whether we can construct a *fair dataset* by taking samples from multiple data sources. Fair dataset, here, is defined as the one with the desired fraction of samples from each group. This testing problem is relevant to our work since the sampling ratios computed by our method might not be feasible for the given data source.

Disparity measures. Harper and Lynch [18] presents a comprehensive discussion of measures for quantifying health disparities including issues such as using relative vs absolute measures and weighting the metrics by group size. In economics, several *inequality indices* have been proposed such as Atkinson index and Gini coefficient [3, 26]. Notable is the axiomatic approach to defining a measure in this literature. Starting from axioms such as additive decomposability (the inequality measure being the sum of group-specific inequalities), symmetric, and scale invariance [39]. Speicher et al. [40] considers the Generalized entropy index (defined in Table 1) that satisfies many of such desired properties and applies it to study fairness of predictive models. We take inspiration from Lum et al. [27] which similarly considers a set of fairness metrics defined as functions of group means. It addresses the problem of statistical bias while estimating such metrics via plugging-in group means from the sample. In contrast, we study the problem of reducing variance in estimation. Another closely related work is Friedberg et al. [13] which derives the asymptotic sampling distribution of a newly proposed fairness metric, named deviation from equal representation. We leverage the technique it uses, that is the delta method, for deriving distributions for a broader class of metrics.

3 METHOD

We first describe different disparity metrics that fall under a general class. Then we describe how we estimate them, followed by the sampling method to increase the efficiency of the estimates for each metric.

Notation. We denote outcome variable by the letter y. We use capital Y_i to refer to the mean of the outcome for group $i \in \{1, \cdots, k\}$. We assume that the population consists of k groups which can be overlapping. Empirical estimate of Y_i is denoted by $\widehat{Y_i}$. Number of samples taken from group i is n_i out of the population size of N_i. Total sample size is $n = \sum_{i=1}^{k} n_i$, similarly, population size is $N := \sum_{i=1}^{k} N_i$. Outcome distribution for individuals in group i has a variance of σ_i^2. The function $d()$ denotes the disparity metric.

Disparity metric	$d(Y_1, Y_2; N_1, N_2)$	Efficient sampling ratio for group 1
Difference in means	$Y_2 - Y_1$	$\sigma_1/(\sigma_1 + \sigma_2)$
Between-group variance	$\sum_i (Y_i - \bar{Y})^2$	$\sigma_1/(\sigma_1 + \sigma_2)$
Deviation from equal representation (DER)	$\frac{k}{k-1} \sum_i \left(\frac{Y_i}{\sum_j Y_j} - \frac{1}{k}\right)^2$	$\sigma_1 Y_2/(\sigma_1 Y_2 + \sigma_2 Y_1)$
Ratio of means	Y_2/Y_1	$\sigma_1 Y_2/(\sigma_1 Y_2 + \sigma_2 Y_1)$
Population-Attributable Risk (%)	$(Y_2 - Y_1)/Y_2 \times 100$	$\sigma_1 Y_2/(\sigma_1 Y_2 + \sigma_2 Y_1)$
Mean logarithmic deviation	$\sum_i - \log(Y_i/\bar{Y})$	$\sigma_1 Y_2/(\sigma_1 Y_2 + \sigma_2 Y_1)$
Theil's index	$\sum_i Y_i/\bar{Y} \log(Y_i/\bar{Y})$	$\sigma_1 Y_2/(\sigma_1 Y_2 + \sigma_2 Y_1)$
Generalized entropy index (α)	$1/(\alpha^2 - \alpha) \sum_i ((Y_i/\bar{Y})^\alpha - 1)$	$\sigma_1 Y_2/(\sigma_1 Y_2 + \sigma_2 Y_1)$
Index of disparity	$1/2(Y_i - \bar{Y})/\bar{Y} \times 100$	$\sigma_1 Y_2/(\sigma_1 Y_2 + \sigma_2 Y_1)$
Overall average	$\frac{1}{N} \sum_{i=1}^k N_i Y_i$	$\sigma_1 N_1/(\sigma_1 N_1 + \sigma_2 N_2)$

Table 1: Examples of disparity metrics and their efficient sampling proportions for stratified random sampling. We consider only $k = 2$ groups. The i^{th} group's mean outcome is denoted by Y_i, group size by N_i, standard deviation by σ_i, and $N = \sum_i N_i$ is the total size of population.

3.1 Defining disparity

Broadly speaking, a disparity is some measure of discrepancy between outcomes for two or more population groups. A popular way of comparing the groups is by comparing their mean outcomes, for example, by taking the difference or ratio of the group means.

Given an outcome variable, we define a class of disparity metrics as metrics that are expressed as an arbitrary function of group-wise means of the outcome. Formally, if the vector of group-wise means is $\mathbf{Y} := (Y_1, Y_2, \dots, Y_k)$ for the k groups. Then, we consider a disparity metric of the form $d(\mathbf{Y})$ where d is a function with k inputs. Later, we will require this function to be once-differentiable for our method.

$$\text{Disparity} := \overbrace{d}^{\text{any function}} (Y_1, Y_2, \dots, \overbrace{Y_k}^{\text{group } k\text{'s mean}})$$

We can assign weights to each group reflecting their size or importance while computing the disparity from the group-wise means. Thus, $\mathbf{Y}$ can be defined as $(w_1 Y_1, w_2 Y_2, \dots, w_k Y_k)$. For simplicity, we will write the unweighted outcomes. Disparity metrics include the difference or ratio of group averages or a more involved transformation such as in the metric named deviation from equal representation (DER) [13].

Example 3.1 (Difference in means). $d_{DIFF}(\boldsymbol{Y}) := Y_2 - Y_1$.

Example 3.2 (DER). $d_{DER}(\boldsymbol{Y}) := \frac{k}{k-1} \sum_i \left(\frac{Y_i}{\sum_j Y_j} - \frac{1}{k}\right)^2$.

We can observe the stark contrast between the above two metrics which makes them suitable for different applications. Difference in means depends on the magnitude of the outcomes which is preferable when the absolute value of the disparity matters. On the other hand, DER is scale-invariant as it depends on the relative ratio of outcomes alone. Take for example findings from the UN Women 2018 report [43] – "Compared to men, women do three times the amount of unpaid care and domestic work within families. Gender differences [in prevalence of food security] are greater than 3 percentage points and biased against women in nearly a quarter of the 141 countries sampled and against men in seven countries." The first measure is relative while the second one is absolute. Table 1 gives more examples of commonly-used metrics which can be expressed as $d(\mathbf{Y})$. For instance, Population-Attributable Risk (%) is defined as the percentage reduction in the disease risk for a group if everyone had the disease risk of the reference group [28]. It is computed as $(Y_2 - Y_1)/Y_2 \times 100$ where the mean Y_i is the disease risk, that is the proportion of affected individuals in group i.

Estimating disparity. Given a dataset containing outcomes measured for multiple individuals belonging to each group, a natural way to estimate $d(\mathbf{Y})$ is to estimate the disparity using the group averages $d(\widehat{\mathbf{Y}})$. Here each $\widehat{Y_i} = 1/n_i \sum_j y_{i,j}$ is the sample average of the outcomes $y_{i,j}$ collected for the group i across the n_i samples from the group. In summary, we run stratified sampling. We collect samples from the population after stratifying it based on group membership and compute disparity using the averages from each stratum.

We remark that estimating $d(\mathbf{Y})$ as $d(\widehat{\mathbf{Y}})$, while intuitive, is not guaranteed to give an unbiased estimate. The uncertainty in sample averages for each group need not 'cancel out'. In fact, Lum et al. [27] describes the bias of $d(\widehat{\mathbf{Y}})$ and proposes a debiased estimator for one of the disparity metrics (between-group variance). However, as we show in the next section, $d(\widehat{\mathbf{Y}})$ does have desirable asymptotic behavior.

3.2 Computing asymptotic distribution by the delta method

We analyze the large sample behavior of the estimated disparity metric to use it further in increasing the efficiency of the estimate. Here, we are largely inspired by the analysis of a particular disparity metric DER developed in Friedberg et al. [13] which finds the asymptotic distribution of the estimated DER by the delta method. Delta method uses a first-order Taylor expansion of the function $d(\widehat{\mathbf{Y}})$ around $d(\mathbf{Y})$ to characterize its distribution in the limit. We first recall the multivariate form of the delta method.

Theorem 1 (Delta method e.g. Theorem 3.7 in DasGupta [10]). *Given a sequence of k-dimensional random vectors* $\{\mathbf{Y}_n\}$ *such* *that* $\sqrt{n}(\boldsymbol{Y}_n - \theta) \rightarrow \mathcal{N}_k(0, \Sigma(\theta))$*. Consider a function* $d : \mathbb{R}^k \rightarrow \mathbb{R}$ *where* d *is once-differentiable at* θ *and* $\nabla d(\theta)$ *is the gradient vector at* θ*. Then, we have*

$$\sqrt{n}\left(d(\widehat{\boldsymbol{Y}_n}) - d(\theta)\right) \xrightarrow{distr.} \mathcal{N}\left(0, \nabla d(\theta)^\top \Sigma(\theta) \nabla d(\theta)\right)$$

provided $d(\theta)^\top \Sigma(\theta) \nabla d(\theta)$ *is positive.*

We first note that sample averages $\widehat{\mathbf{Y}}$ follow a multivariate Normal distribution asymptotically by the central limit theorem. It has mean $\mathbf{Y}$ and variance $\Sigma := diag(\sigma_1^2/n_1, \sigma_2^2/n_2, \ldots, \sigma_k^2/n_k)$ which is a diagonal matrix with k elements where σ_i^2 is the variance of the random variable Y_i. This follows from Fuller [14, Theorem 1.3.2] since, in stratified sampling, we perform *simple random sampling* without replacement in each stratum independently. Throughout we ignore the finite sample correction which multiplies $(1 - n/N)$ to the variance where N is the population size. For simplicity of the formulae, we assume that the sample size is negligible compared to the population size such that $n/N \rightarrow 0$. This is the case while surveying large populations for instance in a census.

Given $\widehat{\mathbf{Y}}$ is Normally distributed in the limit, we can apply Theorem 1 to the empirical estimate of disparity $d(\widehat{\mathbf{Y}})$. Asymptotically $d(\widehat{\mathbf{Y}})$ follows a Normal distribution with mean $d(\mathbf{Y})$ and variance $\nabla d(\mathbf{Y})^\top \Sigma d(\mathbf{Y})$ which we will denote by σ_d^2. We provide variances of two of the metrics.

Example 3.3 (Difference in means e.g. [5]).

$$\sigma_{DIFF}^2 := \frac{1}{n}\left(\frac{\sigma_1^2}{p_1} + \frac{\sigma_2^2}{p_2}\right).$$

Example 3.4 (DER with $k = 2$ e.g. [13]).

$$\sigma_{DER}^2 := \frac{16}{n}\frac{(Y_2 - Y_1)^2}{(Y_1 + Y_2)^6}\left(\frac{Y_2^2\sigma_1^2}{p_1} + \frac{Y_1^2\sigma_2^2}{p_2}\right).$$

More examples are given in Table 5 in Appendix B.

3.3 Computing efficient sampling proportions

Our goal is to estimate the disparities efficiently that is with low error for a fixed sample size. From the asymptotic distribution of the estimated disparity in Theorem 1, we observe that the estimate is centered at the true value asymptotically. So, one way to improve its efficiency is to reduce the variance $d(\theta)^\top \Sigma(\theta) \nabla d(\theta)$. We will find the proportion of the samples to be taken from each group that minimize the variance. We term this as *Neyman allocation* for estimating disparities as this extends the efficient allocation for estimating population mean which has the same name [30].

Take for example the difference metric $d(Y_1, Y_2) := Y_2 - Y_1$ computed from a sample of size n containing p_1, p_2 proportions from the two groups where $p_1 + p_2 = 1$. Estimated metric value is $\widehat{Y}_2 - \widehat{Y}_1$. The variance of its asymptotic distribution is $\sigma_d^2(p_1, p_2, n) := \frac{1}{n}(\frac{\sigma_2^2}{p_2} + \frac{\sigma_1^2}{p_1})$ by the delta method. Different sample sizes for each group will result in different variances. To increase the efficiency of the estimate, we can find the proportions that minimize the variance. For a fixed n, the only variable in the function σ_d^2 is p. It is minimized when $p_1^* = \sigma_1/(\sigma_1 + \sigma_2)$ and $p_2^* = 1 - p_1^*$ which can be obtained by solving the first-order condition $\frac{d}{dp_1}\sigma_d^2(p_1, 1 - p_1, n) = 0$. Similarly we can find variance-minimizing proportions for any disparity metric of the form $d(\widehat{\mathbf{Y}})$. We report these efficient sampling proportions in Table 1.

3.4 Practical implementation using a pilot study

We immediately notice that in some cases the efficient sampling proportions in Table 1 depend on the true group means Y_i or standard deviation σ_i, which are the quantities that we seek to estimate in the first place. To circumvent this problem, we estimate the means and standard deviations from a small pilot study. The pilot can be conducted by any randomized sampling procedure as long as we obtain accurate estimates of the group means and variances. This ensures that the estimated sampling proportions for the main study are close to the efficient ones (see Cai and Rafi [9] for a detailed analysis). We choose to sample each individual uniformly at random irrespective of their group. After the pilot study, we compute an estimate of the efficient sampling proportions and then use these to sample groups differentially in the main study. This dependence between the pilot and the main study data means that we can not simply pool them and compute $\widehat{\mathbf{Y}}$ since the samples are not independently sampled as required by the canonical Central Limit Theorem. We instead compute estimates of disparity separately for the pilot and main study and then average them as done previously in adaptive data collection work [5, 45]. Suppose the sample sizes and group averages in the pilot and main study are $(n_{\text{pilot}}, \widehat{\mathbf{Y}}_{\text{pilot}})$ and $(n_{\text{main}}, \widehat{\mathbf{Y}}_{\text{main}})$. Then the estimated sample mean is

$$\widehat{\mathbf{Y}}_{\text{aggregate}} = \frac{1}{n_{\text{pilot}} + n_{\text{main}}}\left(n_{\text{pilot}} \times \widehat{\mathbf{Y}}_{\text{pilot}} + n_{\text{main}} \times \widehat{\mathbf{Y}}_{\text{main}}\right).$$

Disparity is computed as $d(\widehat{\mathbf{Y}}_{\text{aggregate}})$ as earlier.

In summary the overall method is that we run stratified sampling where strata are defined at two levels, by the batch (either pilot or main study) and within each batch by the group membership. The main study is optimized based on estimates obtained in the pilot study. We compute the disparity using the averages from each strata.

4 APPLICATIONS

We apply the asymptotic distribution of disparity estimates to the problem of determining sample sizes for inference on disparities and evaluating fairness of prediction models.

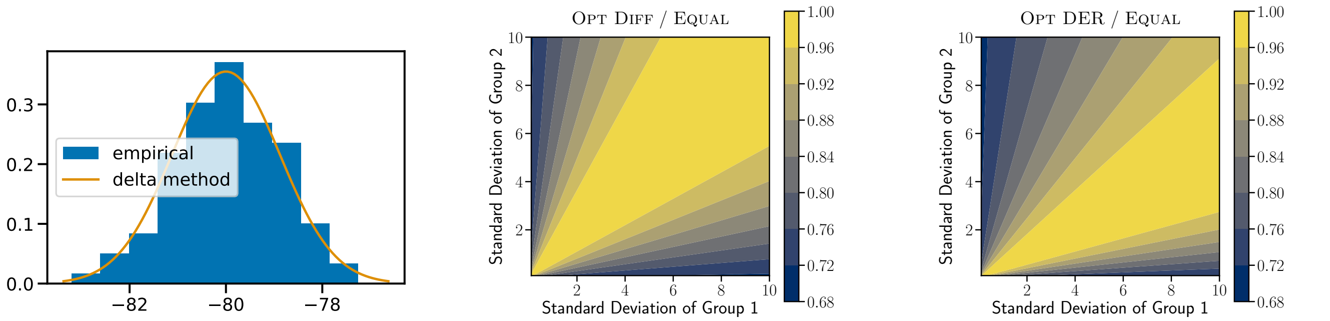
4.1 Determining sample sizes

The normal approximation for the disparity estimates allows determining total number of samples needed for different statistical tasks following standard calculations [15, Chapter 20].

Sample size for a desired precision. Consider a disparity estimate $d(\widehat{\mathbf{Y}})$ with the asymptotic variance of $\sigma_d^2(\mathbf{p}, n)$ given by Theorem 1. If we want a standard error of *se* for the estimate, then the sample size can be computed by solving for n in the equation $se = \sigma_d(\mathbf{p}, n)$ [15, Section 20.3]. For difference in means we get,

$$n = (\sigma_1^2/p_1 + \sigma_2^2/p_2)/se^2.$$

We can further use the efficient allocations to the groups $p_1^* = \sigma_1/(\sigma_1 + \sigma_2)$ and $p_2^* = 1 - p_1^*$ to minimize the sample size. As done earlier in Section 3.4, estimates for the standard deviations σ_1 and σ_2 can be computed from a pilot study or guessed based on expert



(a) Approx. to asymptotic distribution **(b) Relative efficiency for Difference** **(c) Relative efficiency for DER ($Y_1 = 2, Y_1 = 1$)**

Figure 2: Synthetic data. (a) Comparing the empirical sampling distribution of the difference metric with the one given by the delta method. (b,c) Relative efficiency as a function of the standard deviations of groups. Plots show settings where we can expect large improvements in efficiency. That is, regions with values considerably less than 1 like when standard deviations differ between groups in off-diagonal regions of b and c. Since variance of DER depends on both standard deviation and means, we see that plot c is not symmetric along the diagonal as the mean of group 1 is twice that of group 2.

knowledge. Note that we only need to know the ratio of standard deviations to compute the sampling proportions which might be easier to specify.

Sample size for a hypothesis test. Our goal can be to test whether the disparity is significantly high, taken to be a pre-specified value of δ_1, different from a low disparity value of δ_0, such as 0. That is, the null and the alternative hypotheses are $\mathrm{H}_0 : d(\mathbf{Y}) = \delta_0$ and $\mathrm{H}_{\mathrm{alt}} : d(\mathbf{Y}) = \delta_1$. The sample size for the test at significance level α and power $1 - \beta$ can be computed as

$$n = (Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma_d^2 / (\delta_1 - \delta_0)^2,$$

where Z is the inverse CDF of the standard Normal distribution.

4.2 Fairness evaluation of a trained model

Say we are given a trained model and we want to collect data to check the model for fairness violations. A loss function is defined for each data point as $\ell(z, \hat{z}(x))$ where z is the target and $\hat{z}(x)$ is the model prediction for features x. For example, this can be the 0-1 loss $\mathbf{1}[z \neq \hat{z}(x)]$. Then the fairness measure is defined as disparity in the average losses for each group, $Y_i = \mathbb{E}_{(z,x)\sim P_i}[\ell(z, \hat{z}(x))]$. Here P_i is the distribution of target and features for group i. This means that we want to measure $d(\mathbf{Y})$ where each Y_i is the average loss for group i. We can find sample sizes to compute the fairness measure to a desired precision or for a hypothesis test, as done above. This requires that we have the variance of losses for each group $\sigma_i^2 = \mathrm{Var}_{(z,x)\sim P_i}(\ell(z, \hat{z}(x)))$. Per-group variances and means can be estimated from a pilot study as done in Section 3.4.

5 EMPIRICAL STUDY

Through the experiments, we aim to address the following,

Q1. Does the delta method give an accurate approximation? (Figure 2a)

Q2. Does the pilot and main study setup lead to unbiased disparity estimates? (Figure 3)

Q3. Does the use of pilot data affect efficiency of estimates? (Table 3)

Q4. When can we expect our optimal allocation to have large increase in precision? (Figure 2b,c)

Q5. How well does the method do in practice for different metrics? (Tables 2, 4, Figure 4)

We answer these questions using synthetic data and two survey datasets for two tasks, namely, measuring outcome disparities, and evaluation of model fairness.

5.1 Measuring disparities in outcomes

We evaluate our method on estimating disparities in outcomes using a synthetic dataset and two large surveys, ACS and BRFSS. US Census Bureau releases the American Community Survey (ACS) data yearly which contains responses on education, housing, health, demography, and many other variables from a representative sample of the US households [7]. Behavioral Risk Factor Surveillance System (BRFSS) is a nationwide US survey of health-related risk behaviors, chronic health conditions, and use of preventive services [12].[1]

Evaluation setup and baselines. We simulate a survey using the pilot and main study setup in Section 3.4 for different ways of selecting the number of samples per group. We compare the proposed method with two baselines, Equal: equal representation which takes equal number of samples for the two groups, and Uniform: Uniform sampling which samples each individual with the same probability irrespective of their group (thus sampling proportional to the population proportions). Proposed method uses the optimal sampling proportions given in Table 1. For example, for the difference in means metric, we sample in proportion to standard deviations in the pilot data, and name this Opt Diff. Similarly, Opt DER refers to the proposed method for the DER metric. We compute standard deviations as the square root of the sample variance as $\widehat{\sigma_i^2} = 1/n_i \sum_{j=1}^{n_i} (y_{i,j} - \widehat{Y}_i)^2$. Note that this is computed on the pilot sample.

Evaluation criterion. For each method, we compute the root mean squared error (RMSE) between the estimate and the *ground*

[1] https://www.cdc.gov/brfss/index.html

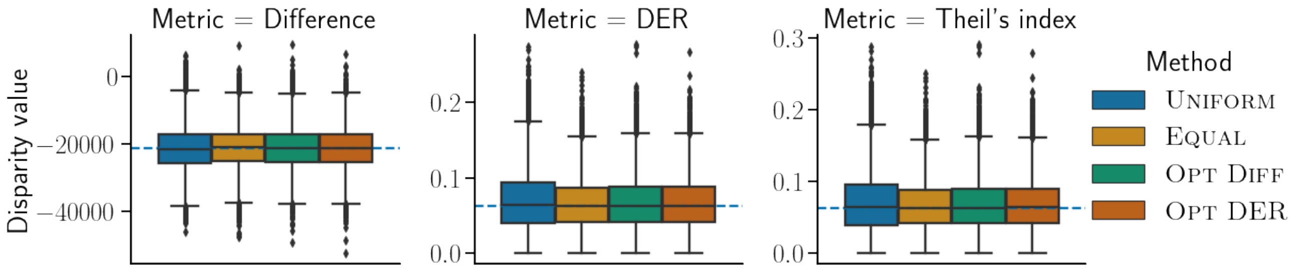


Figure 3: Income disparity from ACS data. Estimates of three disparity metrics obtained from the pilot-main study setup repeated 10,000 times. Dotted line represents the ground truth disparity. We observe that the estimates are unbiased.

truth disparity value. Lower value is better. That is, we report $(\mathbb{E}[(d(\widehat{Y}) - d(Y))^2])^{1/2}$ where the expectation is taken over different draws of the sample from a given population. To estimate this expectation, we repeat the simulations $10,000$ times and report the average squared error. Ground truth disparity $d(Y)$ is taken to be the disparity computed with the whole population's data. We set population size as $N = 100,000$ or all available survey data, whichever is smaller.

Relative efficiency. We can preemptively check how much improvement we can expect in the best case from the efficient allocation by comparing its asymptotic variance with that of other allocations. We define relative efficiency from using the efficient sampling proportions as the ratio of the asymptotic variance for the efficient and equal sampling proportions, similar to Blackwell et al. [5]. Given the asymptotic variance for the disparity metric d is written as $\sigma_d^2(p_1, p_2)$, we compute the relative efficiency as follows,

$$\text{Relative efficiency} := \frac{\sigma_d^2(p_1^*, p_2^*)}{\sigma_d^2(1/2, 1/2)} \leq 1. \tag{1}$$

A low value suggests better precision (lower variance) from sampling by efficient proportions. The value represents the fraction of data points that can be saved from sampling while keeping the same variance as equal sampling. In the experiments, we will instead report the ratio of mean squared error in estimates by efficient and equal sampling as it combines both bias and variance.

	Difference in means		DER	
Method	RMSE	Rel. eff. ↓ (x Equal)	RMSE	Rel. eff. ↓ (x Equal)
Equal	2.94	1.00	0.0017	1.00
Uniform	2.93	1.00	0.0017	1.00
Opt Diff	**2.53**	**0.74**	**0.0016**	**0.81**
Opt DER	2.57	0.77	**0.0016**	**0.82**

Table 2: Error for synthetic data. Root mean squared error for estimates of two disparity metrics improves by sampling using optimal allocation. For comparing the scale of RMSE, the true value of difference in means is -80 and DER is 0.0278. Relative efficiency is defined as ratio of MSEs of optimal and equal sampling proportions.

Results on synthetic data. Data is generated for two groups both with Normal-distributed outcomes where one group's outcome is noisier. For groups $\{1, 2\}$, we generate data as $y_1 \sim \text{Normal}(200, 50)$ and $y_2 \sim \text{Normal}(280, 10)$. Groups are equally represented in the population. The true difference in means is -80 and DER is 0.0278. To test the approximate sampling distribution given by the delta method, we draw 100 populations each of size 10,000 and plot the empirical distribution of the difference in means metric in Figure 2a. We observe that the empirical distribution is close to the one given by the delta method. This supports our use of variance estimates from the delta method for computing the sample sizes.

Table 2 shows the error in disparity estimates for different sampling methods. Out of the total $N = 100,000$ size population, we sample 100 outcomes in the pilot and 500 in the main study. We observe that sampling by the optimal sampling proportions decrease the sample size requirement by a factor of 0.74 for the difference metric and by 0.82 for the DER metric as compared to the error for Equal.

	Difference in means		DER	
Method	RMSE	Rel. eff. ↓ (x Equal)	RMSE	Rel. eff. ↓ (x Equal)
Equal	6326.83	1.00	**0.0342**	**1.00**
Uniform	6619.50	1.09	0.0412	1.46
Opt Diff (w. Pilot)	**6224.65**	**0.97**	0.0363	1.13
Opt DER (w. Pilot)	6246.04	0.97	0.0346	1.02
Opt Diff (Oracle)	6123.42	0.94	0.0349	1.04
Opt DER (Oracle)	6355.95	1.01	0.0343	1.01

Table 3: Error for ACS data. Root mean squared error for estimates of two disparity metrics. Relative efficiency is defined as ratio of MSEs of optimal and equal sampling proportions. For the difference metric, we observe that Opt Diff reduces error. Oracle refers to the proposed methods that use the true standard deviation and mean instead of their approximations from the pilot data. The error for Pilot and Oracle is comparable. Thus, we do not lose efficiency by much by using the approximate sampling proportions. For DER, the errors for Opt DER (both Pilot and Oracle) are similar to Equal.

Results on ACS survey. We query ACS data on annual income and race variable from the 2018 survey using the package `folktables` by Ding et al. [11].[2]. Our goal is to estimate income disparities between white and black or African American population groups. As a convention, we take Y_2 as the outcome for black or African American group when computing disparity metrics (such as $Y_2 - Y_1$). Figure 3 shows the disparity estimates obtained from a pilot of 200 samples and a main study of 500 samples. We observe that the mean of the estimates obtained by repeatedly sampling from the population are close to the true mean, showing unbiasedness for all the methods. Table 3 shows the RMSE of the estimates. We observe that optimal sampling (for difference and DER metrics) has similar error to Equal. This is because the group-wise standard deviations and means are such that the sampling proportions for Opt DER converge to that of Equal (0.53 and 0.5 respectively). For Opt Diff, sampling proportion (0.65) differs from Equal. However the relative efficiency computed as (1) is 0.96. So we expect it to perform similar to Equal.

Results on BRFSS survey. We query BRFSS data from 2014 on the race variable and the age at which respondents were diagnosed with diabetes.[3] Our goal is to estimate disparities in diagnosis age between white and black or African American population groups. We plot the reduction in estimation error achieved by the proposed method as compared to Equal in Figure 4. Both Opt DER and Equal perform similarly for different sample sizes. As in the ACS data, the standard deviations across groups are similar which explains the similar sampling proportions for the two methods. Figure 6 in Appendix C shows that the disparity estimates from BRFSS data are unbiased.

5.2 Measuring fairness of a trained model

Evaluation setup. We consider the task of predicting income level of an individual from attributes related to their education, work, and demography recorded in the ACS data. Prediction target is binary (high vs low income binarized at the income threshold of 50,000 USD). We randomly split the dataset into train (70%) and test (30%). We train a gradient boosting classification model on the train data and compute the fairness metric on test data. We evaluate fairness with respect to disparity in true positive rates, quantified using difference in means and DER metrics. A survey is conducted on the test data as done in Section 5.1 where we take 500 samples in the pilot and 2000 in the main study. We use the same evaluation metric as earlier, that is, RMSE in the fairness estimates.

Results. Table 4 reports the reduction in error for two fairness metrics. We again see that the proposed methods improve on Uniform but perform similarly to Equal. Figure 8 in Appendix D shows that the proposed approach gives unbiased disparity estimates.

In summary, the empirical study shows that the effectiveness of the proposed sampling proportions in providing unbiased estimates of different disparity metrics and reducing the error in estimating them from finite samples. We observe that for the real survey datasets equal allocation of samples to the groups is a good heuristic to get low error. This happens because the groups have similar variance of outcomes. Similar observations on the success of equal allocation have been made in previous studies [9, 37]. However, the proposed methods can help in reducing error in the heteroskedastic case as seen in the synthetic data experiments.

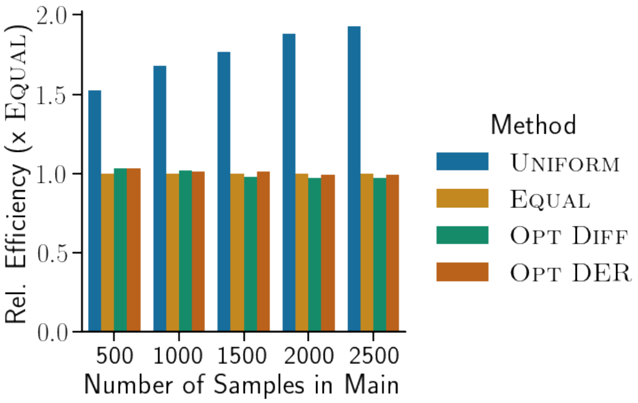


Figure 4: BRFSS data. Relative efficiency, that is reduction in mean squared error relative to Equal, in estimating the DER metric for the outcome: age of diabetes diagnosis. Proposed methods improve upon Uniform for different sample sizes (with a constant pilot sample size of 500). Opt DER and Equal have similar errors (efficiency is close 1).

Method	Difference in means RMSE	Difference in means Rel. eff. ↓ (x Equal)	DER RMSE	DER Rel. eff. ↓ (x Equal)
Equal	0.02	1.00	0.0030	1.00
Uniform	0.03	2.69	0.0053	3.17
Opt Diff	**0.02**	**0.99**	0.0030	0.98
Opt DER	**0.02**	**0.98**	**0.0029**	**0.96**

Table 4: Model fairness evaluation on ACS data. Root mean squared error in estimating fairness of a model for predicting income in ACS data. Fairness is defined as disparity in true positive rates across racial groups. Difference in means for the model is -0.16 biased against black or African American population, DER is 0.0127. For both ways of quantifying fairness, we observe that the proposed sampling methods (Opt Diff, Opt DER) have lower error than Uniform sampling. However, the improvement is similar to using Equal proportions (efficiency is close to 1 for both methods).

6 DISCUSSION

We present an approach to collect data efficiently for the goal of measuring disparities across population groups. For a broad class of disparity measures, defined as arbitrary functions of group-level outcome averages, we propose a sampling approach that maximizes the precision of the disparity estimates. This is achieved by tuning the number of samples taken from each group such as the variance of the asymptotic sampling distribution of the estimates is minimum. The case studies on measuring health outcome disparities from

[2] https://github.com/socialfoundations/folktables
[3] Available at https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system.

survey datasets show the efficacy of the approach. The approach can also be used to evaluate fairness of any given learned model.

A limitation of the work is the narrow focus on disparities as any differences in outcomes without considering the causes of the difference such as social inequities [23]. For a more nuanced analysis of disparities, we may want to look at the differences that remain after adjusting for known risk factors (such as [21]). We ignore these more-involved statistical quantities to only consider differences without adjusting for any features. Further, we may prefer defining disparities using summary statistics other than group averages such as difference in median earnings between women and men as done while calculating gender pay gap. The delta method can still be used with median (or other quantiles) using the corresponding central limit theorem to get the asymptotic sampling distribution [14, Theorem 1.3.10]. Efficient estimation for such measures is an interesting research direction. Another limitation is the need to use up a part of the limited sample size to collect data for the pilot study. Instead we can use sequential sampling methods, such as [8]. We instantiate the approach only for disparities between two groups and leave the derivation of formulae for more groups as further work. Finally, a major assumption we take is that data is always measurable when requested and have no bias. That is, we assume there is no missingness in the outcomes or any systematic errors in the measurements for certain groups. Nonetheless, we hope that our work sheds light on the important problem of disparity estimation and motivates the development of approaches to collect better data in more challenging cases.

REFERENCES

[1] Jacob D Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. 2022. Active Sampling for Min-Max Fairness. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 53–65. https://proceedings.mlr.press/v162/abernethy22a.html

[2] Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. 2022. Fair active learning. *Expert Systems with Applications* 199 (2022), 116981. https://doi.org/10.1016/j.eswa.2022.116981

[3] Anthony B Atkinson. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 3 (1970), 244–263. https://doi.org/10.1016/0022-0531(70)90039-6

[4] Michiel A. Bakker, Duy Patrick Tu, Krishna P. Gummadi, Alex Sandy Pentland, Kush R. Varshney, and Adrian Weller. 2021. Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 346–356. https://doi.org/10.1145/3461702.3462575

[5] Matthew Blackwell, Nicole E. Pashley, and Dominic Valentino. 2022. Batch Adaptive Designs to Improve Efficiency in Social Science Experiments. https://www.mattblackwell.org/files/papers/batch_adaptive.pdf. Accessed 8 November 2022.

[6] Erica Brittain and James J. Schlesselman. 1982. Optimal Allocation for the Comparison of Proportions. *Biometrics* 38, 4 (1982), 1003–1009. http://www.jstor.org/stable/2529880

[7] U.S. Census Bureau. 2023. American Community Survey. https://www.census.gov/programs-surveys/acs/microdata.html Accessed 15 March 2023.

[8] Mark A. Burgess and Archie C. Chapman. 2021. Approximating the Shapley Value Using Stratified Empirical Bernstein Sampling. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 73–81. https://doi.org/10.24963/ijcai.2021/11 Main Track.

[9] Yong Cai and Ahnaf Rafi. 2022. On the Performance of the Neyman Allocation with Small Pilots. https://doi.org/10.48550/ARXIV.2206.04643

[10] A. DasGupta. 2008. *Asymptotic Theory of Statistics and Probability*. Springer New York. https://books.google.com/books?id=sX4_AAAAQBAJ

[11] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems* 34 (2021).

[12] Centers for Disease Control and Prevention. 2013. Behavioral Risk Factor Surveillance System. https://www.cdc.gov/brfss/data_documentation/pdf/UserguideJune2013.pdf Accessed 15 March 2023.

[13] Rina Friedberg, Stuart Ambler, and Guillaume Saint-Jacques. 2022. Representation-Aware Experimentation: Group Inequality Analysis for A/B Testing and Alerting. https://doi.org/10.48550/ARXIV.2204.12011

[14] Wayne A. Fuller. 2009. *Probability Sampling from a Finite Universe*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470523551 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470523551

[15] Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

[16] Jinyong Hahn, Keisuke Hirano, and Dean Karlan. 2011. Adaptive Experimental Design Using the Propensity Score. *Journal of Business & Economic Statistics* 29, 1 (2011), 96–108. https://doi.org/10.1198/jbes.2009.08161 arXiv:https://doi.org/10.1198/jbes.2009.08161

[17] Sam Harper, Nicholas B King, Stephen C Meersman, Marsha E Reichman, Nancy Breen, and John Lynch. 2010. Implicit Value Judgments in the Measurement of Health Inequalities. *The Milbank Quarterly* 88, 1 (2010), 4–29. https://doi.org/10.1111/j.1468-0009.2010.00587.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0009.2010.00587.x

[18] Sam Harper and John Lynch. 2010. Methods for measuring cancer disparities: using data relevant to healthy people 2010 cancer-related objectives. (2010).

[19] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[20] Institute of Medicine and National Academies of Sciences, Engineering, and Medicine. 2016. *Metrics That Matter for Population Health Action: Workshop Summary*. The National Academies Press, Washington, DC. https://doi.org/10.17226/21899

[21] John W. Jackson. 2021. Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework. *Epidemiology* 32, 2 (1 March 2021), 282–290. https://doi.org/10.1097/EDE.0000000000001319 Funding Information: Submitted September 21, 2019; accepted December 7, 2020 From the aDepartment of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; bDepartment of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; cDepartment of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; dJohns Hopkins Center for Health Equity, Baltimore, MD; and eJohns Hopkins Center for Health Disparities Solutions, Baltimore, MD. This research was supported by a grant from the National Heart Lung and Blood Institute (K01HL145320). Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). Correspondence: John W. Jackson, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 614 N. Broadway Room E-6543, Baltimore, MD 21205. E-mail: john.jackson@jhu.edu. Publisher Copyright: © 2021 Lippincott Williams and Wilkins. All rights reserved..

[22] Graham Kalton and Dallas W. Anderson. 1986. Sampling Rare Populations. *Journal of the Royal Statistical Society. Series A (General)* 149, 1 (1986), 65–82. http://www.jstor.org/stable/2981886

[23] Nancy Krieger. 2005. Defining and investigating social disparities in cancer: critical issues. *Cancer Causes & Control* 16 (2005), 5–14. https://link.springer.com/article/10.1007/s10552-004-1251-5

[24] Dejian Lai, Kuang-Chao Chang, Mohammad H Rahbar, and Lemuel A Moye. 2013. Optimal Allocation of Sample Sizes to Multicenter Clinical Trials. *Journal of biopharmaceutical statistics* 23, 4 (2013), 818–828.

[25] Nianyun Li, Naman Goel, and Elliott Ash. 2022. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 396–410. https://doi.org/10.1145/3514094.3534147

[26] Julie A Litchfield. 1999. Inequality: Methods and tools. *World Bank* 4 (1999).

[27] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-Biasing "Bias" Measurement. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 379–389. https://doi.org/10.1145/3531146.3533105

[28] Johan P Mackenbach and Anton E Kunst. 1997. Measuring the magnitude of socio-economic inequalities in health: An overview of available measures illustrated with two examples from Europe. *Social Science & Medicine* 44, 6 (1997), 757–771. https://doi.org/10.1016/S0277-9536(96)00073-1 Health Inequalities in Modern Societies and Beyond.

[29] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[30] J. Neyman. 1938. Contribution to the Theory of Sampling Human Populations. *J. Amer. Statist. Assoc.* 33, 201 (1938), 101–116. https://doi.org/10.1080/01621459.1938.10503378 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1938.10503378

[31] Laura Niss, Yuekai Sun, and Ambuj Tewari. 2022. Achieving Representative Data via Convex Hull Feasibility Sampling Algorithms. https://doi.org/10.48550/ARXIV.2204.06664

[32] Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex 'Sandy' Pentland. 2019. Active Fairness in Algorithmic Decision Making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 77–83. https://doi.org/10.1145/3306618.3314277

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[34] Ana Penman-Aguilar, Makram Talih, David Huang, Ramal Moonesinghe, Karen Bouye, and Gloria Beckles. 2016. Measurement of health disparities, health inequities, and social determinants of health to support the advancement of health equity. *Journal of public health management and practice: JPHMP* 22, Suppl 1 (2016), S33.

[35] Richard D. Riley, Thomas P. A. Debray, Gary S. Collins, Lucinda Archer, Joie Ensor, Maarten van Smeden, and Kym I. E. Snell. 2021. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine* 40, 19 (2021), 4230–4251. https://doi.org/10.1002/sim.9025 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9025

[36] Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. 2021. Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 9040–9051. https://proceedings.mlr.press/v139/rolf21a.html

[37] Evan T. R. Rosenman and Art B. Owen. 2021. Designing experiments informed by observational studies. *Journal of Causal Inference* 9, 1 (2021), 147–171. https://doi.org/doi:10.1515/jci-2021-0010

[38] Amr Sharaf, Hal Daume III, and Renkun Ni. 2022. Promoting Fairness in Learned Models by Learning to Active Learn under Parity Constraints. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2149–2156. https://doi.org/10.1145/3531146.3534632

[39] A. F. Shorrocks. 1980. The Class of Additively Decomposable Inequality Measures. *Econometrica* 48, 3 (1980), 613–625. http://www.jstor.org/stable/1913126

[40] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2239–2248. https://doi.org/10.1145/3219819.3220046

[41] Makram Talih and David T. Huang. 2016. Measuring progress toward target attainment and the elimination of health disparities in Healthy People 2020. *Healthy People Statistical Notes, no 27* (2016). https://www.cdc.gov/nchs/data/statnt/statnt27.pdf

[42] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 336–349. https://doi.org/10.1145/3531146.3533101

[43] UN Women. 2018. Turning promises into action: Gender equality in the 2030 Agenda for Sustainable Development. https://www.unwomen.org/en/digital-library/publications/2018/2/gender-equality-in-the-2030-agenda-for-sustainable-development-2018 Accessed 6 March 2023.

[44] Tom Yan and Chicheng Zhang. 2022. Active fairness auditing. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 24929–24962. https://proceedings.mlr.press/v162/yan22c.html

[45] Kelly Zhang, Lucas Janson, and Susan Murphy. 2020. Inference for Batched Bandits. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9818–9829. https://proceedings.neurips.cc/paper/2020/file/6fd86e0ad726b778e37cf270fa0247d7-Paper.pdf

[46] Konstantin M. Zuev. 2013. Lecture 20-21 Math 408: Mathematical Statistics. https://www.its.caltech.edu/~zuev/teaching/2013Spring/Math408-Lecture-20-21.pdf Accessed 6 March 2023.

A CODE AVAILABILITY

Code to replicate all the experiments is available at the anonymous GitHub repo at https://anonymous.4open.science/r/disparity-variation-705A/.

B VARIANCE FORMULAE FOR DIFFERENT DISPARITY METRICS

Disparity metric	$d(Y_1, Y_2; N_1, N_2)$	Variance σ_d^2
Difference in means	$Y_2 - Y_1$	$\frac{1}{n}\left(\frac{\sigma_1^2}{p_1} + \frac{\sigma_2^2}{p_2}\right)$
Between-group variance	$\sum_i (Y_i - \bar{Y})^2$	$\frac{1}{n}\left(\frac{\sigma_1^2(Y_1-Y_2)^2}{p_1} + \frac{\sigma_2^2(-Y_1+Y_2)^2}{1-p_1}\right)$
Deviation from equal representation (DER)	$\frac{k}{k-1}\sum_i \left(\frac{Y_i}{\sum_j Y_j} - \frac{1}{k}\right)^2$	$\frac{16}{n}\frac{(Y_2-Y_1)^2}{(Y_1+Y_2)^6}\left(\frac{Y_2^2\sigma_1^2}{p_1} + \frac{Y_1^2\sigma_2^2}{1-p_1}\right)$
Ratio of means	Y_2/Y_1	$\frac{1}{n}\left(\frac{\sigma_2^2}{Y_1^2\cdot(1-p_1)} + \frac{Y_2^2\sigma_1^2}{Y_1^4 p_1}\right)$
Population-Attributable Risk (%)	$(Y_2 - Y_1)/Y_2 \times 100$	$\frac{100^2}{n}\left(\frac{\sigma_2^2}{Y_1^2\cdot(1-p_1)} + \frac{Y_2^2\sigma_1^2}{Y_1^4 p_1}\right)$
Overall average	$\frac{1}{N}\sum_{i=1}^k N_i Y_i$	$\frac{1}{n\cdot N^2}\left(\frac{N_1^2\sigma_1^2}{p_1} + \frac{N_2^2\sigma_2^2}{1-p_1}\right)$

Table 5: Examples of disparity metrics and the variance of their asymptotic sampling distributions for stratified sampling. We consider only $k = 2$ groups. The i^{th} group's mean outcome is denoted by Y_i, group size by N_i, standard deviation by σ_i, and $N = \sum_i N_i$ is total size of population.

Table 5 has the formulae for the variance of asymptotic sampling distributions of a few of the disparity metrics from Table 1. These are derived using the delta method and the automated symbolic differentiation package named sympy in Python. The formulae for metrics such as Theil's index and Generalized entropy index are not included since they are long. They can be calculated using the included code.

C ADDITIONAL RESULTS FOR OUTCOME DISPARITY

C.1 Dataset details

We use data from the New York state for ACS and from all available states for BRFSS. In both cases, we use the sample weights provided in the survey to compute weighted means and standard deviations. The outcome and race variables in ACS data are named PINCP and RAC1P, and in BRFSS are named DIABAGE2 and _RACE.

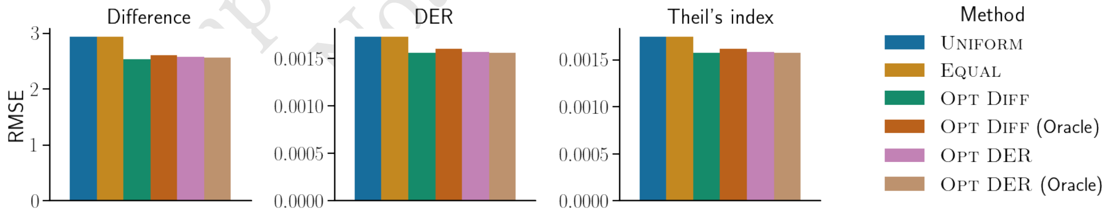


Figure 5: Synthetic data. Error in estimates of three disparity metrics improves by sampling using optimal sampling allocations. Estimated allocations based on pilot data achieve similar error to using the true allocations (Oracle). Pilot study has 100 samples and main study has 500 samples.

D ADDITIONAL RESULTS FOR EVALUATING MODEL FAIRNESS

D.1 Data and model details

We use data from the New York state for predicting annual income from the features that are a part of the ACSIncome data source in the folktables package [11].[4] We do not use sample weights recorded in the ACS data for this experiment, thus, the fairness findings are limited to the population included in the survey. We binarize the annual income as high (vs low) income if it is greater than 50,000 USD.

[4] https://github.com/socialfoundations/folktables

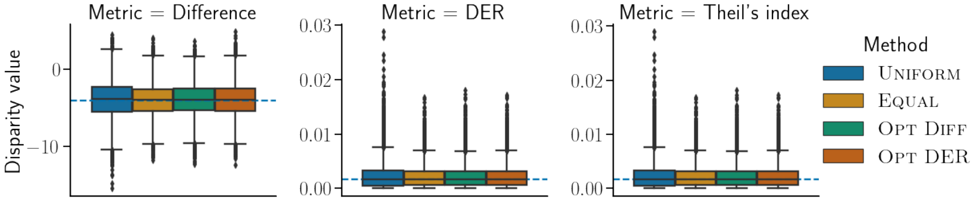


Figure 6: Disparity in age of diabetes diagnosis from BRFSS data. Estimates of the three disparity metrics obtained in pilot and main study setup are unbiased. Both pilot and main study have 500 samples.

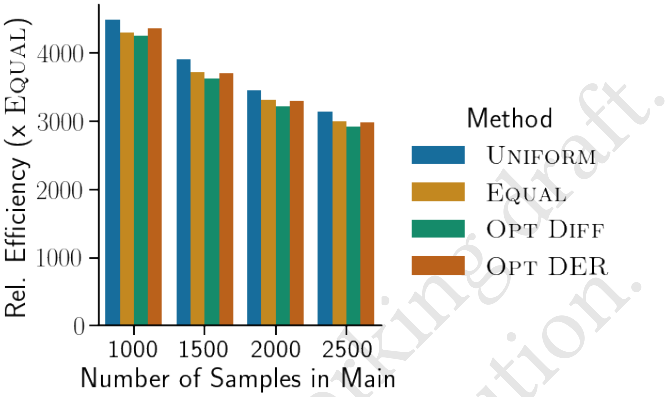


Figure 7: ACS data. Relative efficiency, that is reduction in mean squared error compared to Equal, in estimating the Difference metric for income outcome. Opt Diff improves efficiency for different number of samples in the main study (with a constant sample size of 500 in the pilot).

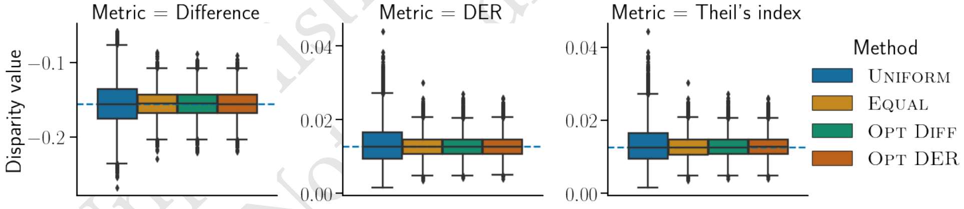
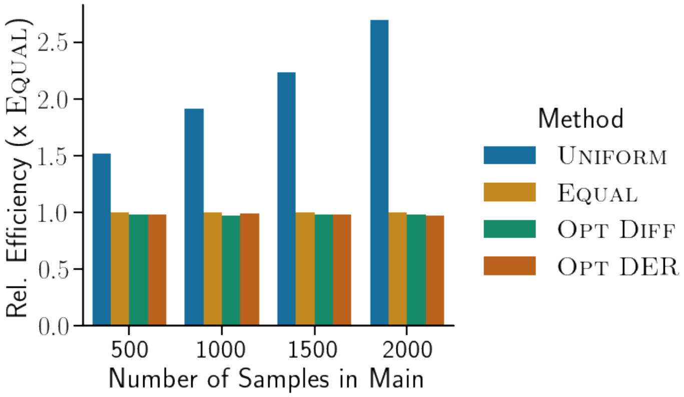


Figure 8: Fairness evaluation of an income prediction model on ACS data. Estimates of the three disparity metrics obtained in the pilot and main study setup are unbiased. Pilot has 500 samples and main has 2000 samples.

We train a gradient boosting classifier with decision trees as weak learners and default hyperparameters for the model class named `GradientBoostingClassifier` in the `scikit-learn` package [33].

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009



(a) Relative efficiency in Difference metric

Figure 9: Relative efficiency of fairness evaluation on ACS data. Relative efficiency in fairness evaluation of a trained model in terms of Difference in true positive rates across white and black or African American groups. Opt Diff improves efficiency for different number of samples in the main study (with a constant pilot sample size of 250).