A Brief Tutorial on Sample Size Calculations for Fairness Audits

Harvineet Singh[†*]**, Fan Xia**[†]**, Mi-Ok Kim**[†]**, Romain Pirracchio**[†]**, Rumi Chunara**[‡]**, Jean Feng**[†]
[†]University of California, San Francisco; [‡] New York University
[*]`fnu.harvineetsingh@ucsf.edu`

Abstract

In fairness audits, a standard objective is to detect whether a given algorithm performs substantially differently between subgroups. Properly powering the statistical analysis of such audits is crucial for obtaining informative fairness assessments, as it ensures a high probability of detecting unfairness when it exists. However, limited guidance is available on the amount of data necessary for a fairness audit, lacking directly applicable results concerning commonly used fairness metrics. Additionally, the consideration of unequal subgroup sample sizes is also missing. In this tutorial, we address these issues by providing guidance on how to determine the required subgroup sample sizes to maximize the statistical power of hypothesis tests for detecting unfairness. Our findings are applicable to audits of binary classification models and multiple fairness metrics derived as summaries of the confusion matrix. Furthermore, we discuss other aspects of audit study designs that can increase the reliability of audit results.

1 Introduction

Disparities in machine learning models' performance across demographic groups, subsequently referred to as unfairness, have been highlighted by audits in multiple domains including criminal justice [1], face image analysis [2], lending [3], online advertising [4], hiring [5], and care management [6]. Such audits have been influential to rollback products or advocate for improvements [7]. Thus, fairness audits are an important mechanism for ensuring accountability for models.

Frameworks exist that provide guidance on how to define, execute, and report audits of algorithmic systems [8, 9]. However, statistical aspects of fairness audits are largely overlooked except by a few studies [10–13]. Audits produce *estimates* of unfairness based on limited samples collected in the audit study. Therefore, standard considerations for making statistical inferences apply such as representativeness of the study sample to the target population, specification of the hypotheses being tested, and communicating uncertainty or statistical power of the tests. However, previous work often does not report uncertainties in the unfairness estimates (e.g. [1, 2], ostensibly because the evaluation datasets used in such studies are large). Importantly, there is not enough guidance on how to design audits such that they are sufficiently powered to detect model unfairness. Previous guidances on sample sizes (e.g. [14–16]) focus on evaluation metrics such as calibration slope or area under the ROC curve which differ from the commonly-used fairness metrics such as false positive/negative rate disparity, and hence are not directly applicable. Even though guidance exists for measures of differences between groups in the clinical trial [17] and health disparities [18] literature, we are not aware of works that present the formulae directly relevant to fairness metrics.

Sample size calculations are important to plan resources required to collect data since audits may not have existing data to validate the models. This is often the case in external, independent audits where the model developer might not be willing to share data (e.g. [19]). Auditors may require conducting

total n_g	$Y = +$	$Y = -$
$\widehat{Y} = +$	TP_g	FP_g
$\widehat{Y} = -$	FN_g	TN_g

Figure 1: **Confusion matrix and derived performance metrics.** The model predicts $\widehat{Y}$ for the target class Y which is assumed to be either $+$ or $-$. The n_g data points in a group are divided into four parts of the confusion matrix such that $n_g = TP_g + FP_g + FN_g + TN_g$ which respectively correspond to number of true positives, false positives, false negatives, and true negatives. Demographic parity, for instance, is the difference $(TP_1 + FP_1)/n_1 - (TP_2 + FP_2)/n_2$.

surveys, buying data, or querying expensive model services, thus, making data collection expensive. In some cases it is possible to leverage existing open datasets such as images of parliamentarians from across the world to audit facial image processing algorithms [2], data released under legal obligations by lenders [3], or direct sharing by the model developer under strict terms [20]. However, such datasets might not be readily available for domains such as healthcare, or we may want to evaluate the model on a new target population not measured before. Focusing on such cases, we only assume query-access to the model being audited (i.e. given an input feature vector, we can query its prediction), and optionally, the ability to collect small amount of pilot data. We assume that the goal of the audit is to conduct a hypothesis test for whether the model is fair. To this end, we describe the sample size needed such that the probability of detecting unfairness is sufficiently high.

The main contribution of this work is to provide a tutorial on calculating sample sizes for fairness audits. We provide sample size formulae that can be directly used for auditing common fairness measures including, but not limited to, demographic (or statistical) parity [21] and false positive/negative rates disparity [22]. By providing simple-to-use formulae, our hope is to ease the adoption and maximize the effectiveness of fairness audits. As with any sample size calculation, the number of samples should be taken as a rough approximation. Knowing them can help to decide whether existing data is sufficient and to plan resources for the audit.

2 Sample size calculations

Denote the performance of the given model on a target population by M. This could be the accuracy, false positive rate, or different ways of measuring model performance. Suppose the population is partitioned into non-overlapping groups, such as using protected attributes like sex and race/ethnicity (or intersections thereof). Let us denote the performance for subgroup g by M_g. Many unfairness measures are expressed as the difference in performances between two groups, i.e. $U = M_1 - M_2$ (see [23] for examples). We assume that group 1 is the privileged and group 2 is the disadvantaged group, in that we believe U to be non-negative. For more than two groups, one can compare performance with respect to a reference group, say, the privileged group. We restrict to binary classification models for exposition. The calculations can be extended to more than two classes.

An example of an unfairness metric is demographic parity, which is the difference between the fraction of positively predicted points between the two groups. Thus, in this case, M_g is the fraction of individuals from group g predicted to be positive. We will provide sample size calculations for various metrics that summarize the confusion matrix [24] (see Figure 1 and Table 1).

Before performing any sample size calculations, we must first define the sampling design. Since we are interested in measuring group-wise performances, we need to collect enough samples from each group. One way to accomplish this is through a stratified random sampling design where individuals from each group are sampled at random with a fixed probability and independently from other individuals. There are more complex designs that might be more suited to the problem context; here we focus on stratified sampling for its applicability and simplicity. Note that for certain performance measures such as demographic parity, we do not need to collect the true labels Y. In general, we need the group membership, true label, and the features for computing the model predictions.

Defining the hypothesis test. We want to test whether the unfairness exceeds some pre-specified tolerance. That is, the null and the alternative hypotheses are

$$\begin{aligned} \text{H}_0 &: U(M_1, M_2) \leq U_{tol} \\ \text{H}_{alt} &: U(M_1, M_2) > U_{tol}. \end{aligned} \quad (1)$$

Table 1: **Variance of performance metrics in each group.** Denote the fraction of total positives (i.e. prevalence) by $M_P = (TP + FN)/n$ and the fraction of positive predictions are denoted by $M_{PP} = (TP + FP)/n$. Denote fraction of true positives by $M_{TP} = TP/n$, true negatives $M_{TN} = TN/n$, false positives $M_{FP} = FP/n$, and false negative $M_{FN} = FN/n$. Then, true positive rate (or recall, sensitivity) is the ratio $M_{TPR} = M_{TP}/M_P$ and true negative rate (or specificity) is $M_{TNR} = M_{TN}/(1 - M_P)$. We can define accuracy, false positive/negative rates and positive/negative predictive values as functions of TP, TN, FP, FN. All metrics are calculated on the subset of data for the group. So, M_{TPR} for a group g refers to the ratio $M_{TP,g}/(M_{TP,g} + M_{FN,g})$. We omit the g in subscript for conciseness.

Performance metric M_g	Group-wise variance $\sigma^2_{M,g}$
Demographic parity	$M_{PP}(1 - M_{PP})$
True positive rate (TPR) or false negative rate (FNR)	$\frac{1}{M_P}M_{TPR}(1 - M_{TPR})$
True negative rate (TNR) or false positive rate (FPR)	$\frac{1}{1-M_P}M_{TNR}(1 - M_{TNR})$
Precision/Positive predictive value (PPV)	$\frac{1}{M_{PP}}M_{TPR}(1 - M_{TPR})$
Negative predictive value (NPV)	$\frac{1}{1-M_{PP}}M_{TNR}(1 - M_{TNR})$

Here, U_{tol} is the minimum tolerated unfairness which can be set to 0. We want to find the minimal sample size n and allocation $(n_1, n - n_1)$ that achieves the specified power for the test at a given significance level.

The test statistic for conducting this hypothesis test is as follows. Let n denote the number of samples in the study, with $n_1, n - n_1$ samples from the two groups. Given this, we estimate the group performances $\widehat{M_1}$ and $\widehat{M_2}$ from n_1 and $n - n_1$ samples respectively. The estimate of unfairness $\widehat{U}$ is the difference between the two. We use the test statistic as $(\widehat{U} - U_{tol})/\widehat{\sigma}_U$, where $\widehat{\sigma}_U$ is any consistent estimate of the standard deviation of unfairness (discussed later). The null is rejected if the test statistic exceeds some threshold. The threshold is chosen to control the Type I error (also called significance level) at level α (e.g. 0.05) while maximizing power $1 - \beta$ (e.g. 0.8). Recall that Type I error is defined as falsely rejecting the null hypothesis when true (declaring a fair model as unfair) and power is defined as correctly rejecting the null hypothesis when false (declaring an unfair model as so).

The hypothesis test defined in (1) is suitable for external audits since rejecting the null would mean that the model is unfair. U_{tol} is specific to the problem and the performance metric. The auditor might refer to unfairness of the system prior to model deployment (such as standard of care in healthcare settings) to decide currently acceptable unfairness and set it as U_{tol}. We would want high power so as to declare unfairness when present. Internal audits might prefer switching the null and alternative hypotheses to be able to conclude that the model is fair.

Asymptotic behavior of the test statistic. To find the sample size needed to power the hypothesis test, we first find the asymptotic distribution of the test statistic. In particular, using the central limit theorem and the delta method (Theorem 3.7 in DasGupta [25]), the estimates of the subgroup performance measures considered in Figure 1 converge to a Normal distribution with variances shown in Table 1. See Appendix B for details. Consequently, the test statistic converges to a Normal distribution with mean $(U - U_{tol})/\sigma_U$ and variance 1 where $\sigma^2_U = \sigma^2_{M,1}/n_1 + \sigma^2_{M,2}/n_2$. Here, $\sigma^2_{M,g}$ are group-level variances that depend on the performance measure M.

Determining sample size and optimal allocations. The sample size for the test needed to achieve power of $1 - \beta$ while controlling Type I error at level α is calculated assuming unfairness is equal to some $\tau > U_{tol}$. In particular, the minimum sample size to achieve the desired Type I error and power is

$$n = (Z_{1-\alpha/2} + Z_{1-\beta})^2(\sigma^2_{M,1}/p_1 + \sigma^2_{M,2}/(1 - p_1))/(\tau - U_{tol})^2 \quad (2)$$

where Z is the inverse CDF of the standard Normal distribution and p_1 is the fraction of samples (n_1/n) we will allocate to group 1 (e.g. see Wittes [17]). Note that the presumed level of unfairness τ should be selected based on prior knowledge about the model. The smaller the difference between τ and U_{tol}, the more samples we will need.

We can find the optimal allocation p_1 that minimizes the required sample size n by solving the first-order optimality condition for (2) in terms of p_1. The optimal allocation is $p_1 = \sigma_{M,1}/(\sigma_{M,1} + \sigma_{M,2})$,

which is also known as Neyman allocation [26] in survey sampling literature. That is, we should allocate samples in proportion to the variability of the metric in the groups. We note that the allocation is not simply to have equal number of samples per group. How much to over- (or even under-) sample the minority group depends on the variances. Sample size under the optimal allocation is

$$n = ((Z_{1-\alpha/2} + Z_{1-\beta})(\sigma_{M,1} + \sigma_{M,2})/(\tau - U_{tol}))^2. \tag{3}$$

To use these sample size formulae, we need estimates of group performances to estimate the group standard deviations $\sigma_{M,g}$ using the Table 1. These can be approximated from data used to train the model or from previously-conducted audits for the model when available. We can alternatively run a small pilot study to estimate the metrics. Since we will adapt the data collection in the main study by estimates from the pilot study, this may bias our unfairness estimates. This can be avoided by pooling estimates from pilot and main studies as described in Singh and Chunara [18].

We illustrate the sample size calculations on an income prediction model trained on American Community Survey data [27] in Appendix A. We observe that sample sizes widely vary depending on the metric. For example, auditing for disparity in TPR requires around 1.6 times as many samples as for demographic parity, mainly due to the increased variance of the unfairness estimate for the TPR metric.

2.1 Other considerations

Even before doing sample size calculations, it is important to define the target population of the audit. This is typically the population which will be subjected to the model predictions such as loan applicants from a lender's area of operation. Identifying the target population is important to ensure that the study collects a representative sample. There might be a mismatch in the target population and the sample collected due to infeasibility of randomly sampling the target population. In this case, our inferences need to stated in terms of the sampled population.

While collecting data for the audit, there are many statistical biases to keep in mind. The target labels collected during the study may differ from the way target label was defined in the model and/or at deployment. Thus, the unfairness identified in the audit might not correspond to the actual unfairness of the model. Finally, there might be non-compliance or selective responses for subjects in the audit which leads to missing data or selection bias in the unfairness estimates. These and other considerations standard in the clinical study literature (e.g. [17]) should be kept in mind while analysing and interpreting the audit results.

3 Conclusion

Independent audits of machine learning models are challenging in part because the training or validation data are typically not available to the auditors. However, reliable estimates of unfairness are important to create public awareness of model failures and demand updates to the model. We present a method to aid auditors in deciding the number of samples needed to test for unfairness. The method applies to common fairness metrics expressed as differences (or other differentiable functions) of group-wise performance metrics, for example, demographic parity, and false positive/negative rate disparities.

As further work, the method can be extended to include performance metrics such as area under the ROC curve and calibration slope. An important next step is to address practical considerations such as missing data or measurement error in group membership or labels [28]. We may want to audit for a large number of groups, especially when considering intersectional identities. How do we reduce the sample sizes required in this case while sharing information across the groups? Also, if there are other types of audits being conducted (e.g. for model calibration or stability [5]), how does one collect a single dataset so that all the different audits are properly powered? The statistical aspects of the audits that we highlight are complementary to aspects such as independence and integrity of the auditing organization, and efficient communication of the audit results to the policy makers and public. Together these may ensure that audits become a widely-adopted mechanism for keeping model developers accountable for their societal impact.

Acknowledgments and Disclosure of Funding

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. `www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, 2016. Accessed on 2020-05-29. 1

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/buolamwini18a.html`. 1, 2

[3] Emmanuel Martinez and Lauren Kirchner. The secret bias hidden in mortgage-approval algorithms. `https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms`, 2021. Accessed: 2023-03-28. 1, 2

[4] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359301. URL `https://doi.org/10.1145/3359301`. 1

[5] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. An external stability audit framework to test the validity of personality prediction in ai hiring. *Data Min. Knowl. Discov.*, 36(6): 2153–2193, nov 2022. ISSN 1384-5810. doi: 10.1007/s10618-022-00861-0. URL `https://doi.org/10.1007/s10618-022-00861-0`. 1, 4

[6] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. 1

[7] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing revisited: Investigating the impact of publicly naming biased performance results of commercial ai products. *Commun. ACM*, 66(1):101–108, dec 2022. ISSN 0001-0782. doi: 10.1145/3571151. URL `https://doi.org/10.1145/3571151`. 1

[8] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372873. URL `https://doi.org/10.1145/3351095.3372873`. 1

[9] Xiaoxuan Liu, Ben Glocker, Melissa M McCradden, Marzyeh Ghassemi, Alastair K Denniston, and Lauren Oakden-Rayner. The medical algorithmic audit. *The Lancet Digital Health*, 4(5): e384–e397, 2022. ISSN 2589-7500. doi: https://doi.org/10.1016/S2589-7500(22)00003-6. URL `https://www.sciencedirect.com/science/article/pii/S2589750022000036`. 1

[10] Tom Yan and Chicheng Zhang. Active fairness auditing. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24929–24962. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/yan22c.html`. 1

[11] Ben Chugg, Santiago Cortes-Gomez, Bryan Wilder, and Aaditya Ramdas. Auditing fairness by betting, 2023.

[12] John J. Cherian and Emmanuel J. Candès. Statistical inference for fairness auditing, 2023.

[13] Pranav Maneriker, Codi Burley, and Srinivasan Parthasarathy. Online fairness auditing through iterative refinement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1665–1676, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599454. URL `https://doi.org/10.1145/3580305.3599454`. 1

[14] Richard D. Riley, Thomas P. A. Debray, Gary S. Collins, Lucinda Archer, Joie Ensor, Maarten van Smeden, and Kym I. E. Snell. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine*, 40(19):4230–4251, 2021. doi: https://doi.org/10.1002/sim.9025. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9025`. 1

[15] Richard D. Riley, Gary S. Collins, Joie Ensor, Lucinda Archer, Sarah Booth, Sarwar I. Mozumder, Mark J. Rutherford, Maarten van Smeden, Paul C. Lambert, and Kym I. E. Snell. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Statistics in Medicine*, 41(7):1280–1295, 2022. doi: https://doi.org/10.1002/sim.9275. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9275`.

[16] Lucinda Archer, Kym I. E. Snell, Joie Ensor, Mohammed T. Hudda, Gary S. Collins, and Richard D. Riley. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine*, 40(1):133–146, 2021. doi: https://doi.org/10.1002/sim.8766. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8766`. 1

[17] Janet Wittes. Sample size calculations for randomized controlled trials. *Epidemiologic reviews*, 24(1):39–53, 2002. 1, 3, 4

[18] Harvineet Singh and Rumi Chunara. Measures of disparity and their efficient estimation. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 927–938, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604697. URL `https://doi.org/10.1145/3600211.3604697`. 1, 4

[19] Shannon Bond. Nyu researchers were studying disinformation on facebook. the company cut them off, 2021. URL `https://www.npr.org/2021/08/04/1024791053/facebook-boots-nyu-disinformation-researchers-off-its-platform-and-critics-cry-f`. Accessed 26 September 2023. 1

[20] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 666–677, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445928. URL `https://doi.org/10.1145/3442188.3445928`. 2

[21] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009. doi: 10.1109/ICDMW.2009.83. 2

[22] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. 2

[23] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/agarwal18a.html`. 2

[24] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021. 2

[25] A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer New York, 2008. ISBN 9780387759715. URL `https://books.google.com/books?id=sX4_AAAAQBAJ`. 3, 8

[26] J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938. doi: 10.1080/01621459.1938.10503378. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1938.10503378`. 4

[27] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021. 4, 9

[28] Michelle Bao, Angela Zhou, Samantha A Zottola, Brian Brubach, Sarah Desmarais, Aaron Seth Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL `https://openreview.net/forum?id=qeM58whnpXM`. 4

[29] U.S. Census Bureau. American community survey, 2023. URL `https://www.census.gov/programs-surveys/acs/microdata.html`. Accessed 15 March 2023. 7

[30] Wayne A. Fuller. *Probability Sampling from a Finite Universe*. John Wiley & Sons, Ltd, 2009. ISBN 9780470523551. doi: https://doi.org/10.1002/9780470523551. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470523551`. 8

A Example

We apply our method to evaluate fairness of an income prediction model trained on survey data. American Community Survey (ACS) data contains responses on education, housing, health, demography, and many other variables from a representative sample of the US households [29].

Experiment setup. We consider the task of predicting income level of an individual from attributes related to their education, work, and demography. 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 unfairness and sample sizes on the test data. We evaluate disparity in model performance between white and Black or African American population groups. In our processed data, Black or African American population group constitutes 7% of the total data points (2562 out of 38869 points). True positive rate for white population is 0.79 which is 11 percentage points higher than for the Black or African American population at 0.68. Table 2 gives an example of the sample sizes obtained from (3) for unfairness defined by different performance metrics.

Take for instance demographic parity (DP). Rate of positive predictions $M_{PP,1}, M_{PP,2}$ in Black or African American and white population groups on the test set are $0.3478, 0.4404$ respectively. Thus, the unfairness value is 0.093 which we take as τ. Group variances $\sigma_{M,1}^2, \sigma_{M,2}^2$ from Table 1 are $0.227, 0.246$. Plugging these in (3) we get the minimum required sample size to detect τ is

$$n \approx (Z_{1-0.05/2} + Z_{1-0.2})^2 \times (0.227^{1/2} + 0.246^{1/2})^2/0.093^2 \approx 855.$$

Table 2: **Example of sample size calculations for an income prediction model on ACS data.** We report the unfairness (its absolute value) and sample sizes needed to detect that value for different metrics. We evaluate the n at optimal allocation for different unfairness metrics and for a hypothesis test at $\alpha = 0.05, \beta = 0.2, U_{tol} = 0$. We set τ to be the actual unfairness estimate for the metric computed on the test data. Sample sizes vary a lot depending on the metric of interest. The sample size for NPV is 846,671 mainly because we required an extremely precise value of $\tau = 0.003$. Note that the numbers are only meant to illustrate the sample size formulae and are not for direct use.

Performance metric M	Presumed unfairness τ	Sample size n
DP	0.09	855
TPR	0.11	1,379
TNR	0.03	6,540
PPV	0.06	4,364
NPV	0.003	846,671

B   Derivations

Asymptotic distribution of performance metrics. We first note that for many performance metrics, $\widehat{M}_g$ are sample averages or ratios of sample averages. For example, accuracy is an average of correct predictions, $(TP + TN)/n$. This implies that $(\widehat{M}_1, \widehat{M}_2)$ follows a multivariate Normal distribution asymptotically by the central limit theorem. It has mean M_1, M_2 and the variance depends on the particular performance metric M. This follows from Fuller [30, Theorem 1.3.2] for data collected by *simple random sampling* without replacement in each group independently. Throughout we ignore the finite sample correction which multiplies $(1 - n/N)$ to the variance where N is the population size. We assume that the sample size is negligible compared to the population size such that $n/N \to 0$. This is the case while surveying large populations.

For true positive rate (TPR), we need the joint distribution of TP and P which is the following,

$$\sqrt{n} \begin{pmatrix} \widehat{M}_{TP} - M_{TP} \\ \widehat{M}_P - M_P \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} M_{TP}(1 - M_{TP}) & M_{TP}(1 - M_P) \\ M_{TP}(1 - M_P) & M_P(1 - M_P) \end{bmatrix} \right)$$

Similarly, for true negative rate (TPR), we need the joint distribution of TN and N which is the following,

$$\sqrt{n} \begin{pmatrix} \widehat{M}_{TN} - M_{TN} \\ \widehat{M}_N - M_N \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} M_{TN}(1 - M_{TN}) & M_{TN}(1 - M_P) \\ M_{TN}(1 - M_P) & M_P(1 - M_P) \end{bmatrix} \right)$$

Asymptotic distribution of ratios of performance metrics. Now, we require the asymptotic distribution of ratios of TP and P, or TN and N. Distribution of differentiable functions of Normally distributed random variables can be computed by the delta method. Denote some performance metrics of interest arranged in the vector $\mathbf{M}$, for instance, (M_{TP}, M_{TN}, M_P).

Theorem 1 (Delta method e.g. Theorem 3.7 in DasGupta [25]) *Given a sequence of k-dimensional random vectors $\{\widehat{\boldsymbol{M}}_n\}$ such that $\sqrt{n}(\widehat{\boldsymbol{M}}_n - \theta) \xrightarrow{distr.} \mathcal{N}_k(0, \Sigma(\theta))$. Consider a function $d : \mathbb{R}^k \to \mathbb{R}$ where d is once-differentiable at θ and $\nabla d(\theta)$ is the gradient vector at θ. Then, we have*

$$\sqrt{n} \left(d(\widehat{\boldsymbol{M}}_n) - d(\theta) \right) \xrightarrow{distr.} \mathcal{N} \left(0, \nabla d(\theta)^\top \Sigma(\theta) \nabla d(\theta) \right) \tag{4}$$

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

By Theorem 1, the empirical estimate of the performance metric $d(\widehat{M})$ follows a Normal distribution asymptotically with mean $d(\mathbf{M})$ and variance $\nabla d(\mathbf{M})^\top \Sigma \nabla d(\mathbf{M})$.

Variance for estimated TPR which is $\widehat{M}_{TP}/\widehat{M}_P$ is $\frac{1}{nM_P} M_{TPR}(1 - M_{TPR})$. Similarly variance for estimated TNR is $\frac{1}{n(1-T_P)} M_{TNR}(1 - M_{TNR})$.

Asymptotic distribution of unfairness metrics. We need to compute the variance of the unfairness estimate $\widehat{U}$. By definition, unfairness is a difference, $\widehat{M}_1 - \widehat{M}_2$, of statistically independent quantities since we compute the metrics separately on independently-sampled groups. Thus the asymptotic variance of unfairness estimate is the sum of variances, $\sigma_U^2 = \sigma_{M,1}^2/n_1 + \sigma_{M,2}^2/n_2$. We can combine the above to get variance for disparity in TPR as

$$\sigma_U^2 = \frac{1}{M_{P,1}} M_{TPR,1}(1 - M_{TPR,1}) + \frac{1}{M_{P,2}} M_{TPR,2}(1 - M_{TPR,2}).$$

For the hypothesis test, we use the test statistic as $(\widehat{U} - U_{tol})/\widehat{\sigma}_U$. For a consistent estimate $\widehat{\sigma}_U$ converges to σ_U. Then, the test statistic converges to a Normal distribution with mean $(U - U_{tol})/\sigma_U$ and variance 1 by applying the Slutsky's theorem. Now we do sample size calculations for the test.

Sample size calculation. Denote standard deviation under the null and alternative hypotheses by $\sigma_{U,null}, \sigma_{U,alt}$ respectively. We want significance level α which means the following holds for the test statistic.

$$\sup_{u \leq U_{tol}} P\left(\frac{\widehat{U} - U_{tol}}{\sigma_{U,null}} > Z_{1-\alpha/2}\Big| U = u\right) \leq \alpha$$

$$\implies P\left(\frac{\widehat{U} - U_{tol}}{\sigma_{U,null}} > Z_{1-\alpha/2}\Big| U = U_{tol}\right) \leq \alpha$$

The following should hold as well for power $1 - \beta$.

$$\inf_{u \geq \tau} P\left(\frac{\widehat{U} - u}{\sigma_{U,alt}} > Z_{1-\alpha/2}\Big| U = u\right) \geq 1 - \beta$$

$$\implies P\left(\frac{\widehat{U} - U_{tol}}{\sigma_{U,alt}} > Z_{1-\alpha/2}\Big| U = \tau\right) \geq 1 - \beta$$

$$\implies P\left(\frac{\widehat{U} - \tau}{\sigma_{U,alt}} > Z_{1-\alpha/2} - \frac{(\tau - U_{tol})}{\sigma_{U,alt}}\Big| U = \tau\right) \geq 1 - \beta$$

Setting $Z_{1-\alpha/2} - \frac{(\tau - U_{tol})}{\sigma_{U,alt}} \leq -Z_{1-\beta}$, we have by rearranging terms that

$$\sigma_{U,alt}^2 \leq \frac{(\tau - U_{tol})^2}{(Z_{1-\alpha/2} + Z_{1-\beta})^2}$$

$$\iff n \geq \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2(\sigma_{M,1}^2/p_1 + \sigma_{M,2}^2/(1 - p_1))}{(\tau - U_{tol})^2}$$

We can replace p_1 by the optimal allocation $\sigma_{M,1}/(\sigma_{M,1} + \sigma_{M,2})$ to get the minimum sample size.

C Code availability

Code to query ACS data is taken from the `folktables` packages [27]. Scripts to compute the sample sizes will be made available at `https://github.com/harvineet/sample-size-fairness-audits`.