Fairness Issues and Mitigations in (Private) Socio-demographic Data Processes

Joonhyuk Ko†, Juba Ziani‡, Saswat Das†, Matt Williams*, and Ferdinando Fioretto†

†University of Virginia, ‡Georgia Institute of Technology, *RTI International

University of Virginia

Introduction

Statistical agencies rely on sampling techniques to collect socio-demographic data crucial for policy-making and resource allocation. We show that surveys of important societal relevance introduce sampling errors that unevenly impact group-level estimates, thereby compromising fairness in downstream decisions. Additionally, we show that the privacy-preserving methods used to allocate surveys may improve fairness.

Methodology

The accuracy of estimate $\hat{\theta}_i \coloneqq \hat{\theta}_i(n_i)$, via sample size of n_i, is evaluated through their error and variance: $\mathrm{Err}(\hat{\theta}_i) = |\hat{\theta}_i - \theta_i|$ and $\mathrm{Var}(\hat{\theta}_i) = \mathbb{E}[\hat{\theta}_i^2] - (\mathbb{E}[\hat{\theta}_i])^2$. *Unfairness* is quantified by the maximum discrepancy in estimator's variance between any two groups,

$$\xi_{\mathrm{Var}} = \max_{i,j \in G} |\mathrm{Var}(\hat{\theta}_i) - \mathrm{Var}(\hat{\theta}_j)|.$$

Large surveys like the American Community Survey (ACS) use a two-phase data collection: the first phase involves internet or phone interviews, and the second phase involves in-person *door-to-door* interviews [1]. The following program allocates surveys to each subgroup, ensuring group-level accuracy meets a specified threshold α, while minimizing the total survey cost:

$$\underset{p,z}{\text{minimize}} \; \underbrace{c_1 \left(\sum_{i \in [G]} p_i N_i \right)}_{\text{1st phase cost}} + \underbrace{c_2 \left(\sum_{r \in R} z_r \right)}_{\text{2nd phase cost}} \tag{1a}$$

$$\texttt{s.t.} \quad n_i = \underbrace{p_i N_i \left(1 - F_i^1\right)}_{\text{1st phase samples}} + \underbrace{\sum_{r \in R} z_r g^r N_i^r \left(1 - F_i^2\right)}_{\text{2nd phase samples}} \; \forall i \in [G] \tag{1b}$$

$$\Pr(|\mathrm{Err}(\hat{\theta}_i(n_i))| > \gamma_i) \leq \frac{\sigma^2(\hat{\theta}_i)}{\gamma_i^2} \leq \alpha, \quad \forall i \in [G], \tag{1c}$$

$$0 \leq p_i \leq 1 \;\; \forall i \in [N], \;\; z_r \in \{0,1\} \;\; \forall r \in R. \tag{1d}$$

where c_1 and c_2 are costs of phase 1 and 2, N_i^r is the population size of group i in region r, and F_i^1 and F_i^2 are failure rates for phase 1 and 2. The feasible sampling rate in region r is $g^r \in [0,1]$. The decision variable p_i represents the fraction of group i contacted in phase 1, and the binary variable z_r indicates if region r is selected for phase 2.

A key *challenge* with solving Program (1) is Constraint (1c), which involves probability estimation. This was addressed using Chebyshev's inequality, with the variance of the estimator $\sigma^2(\hat{\theta}_i)$ estimated empirically using prior data, as shown in Figure 1.

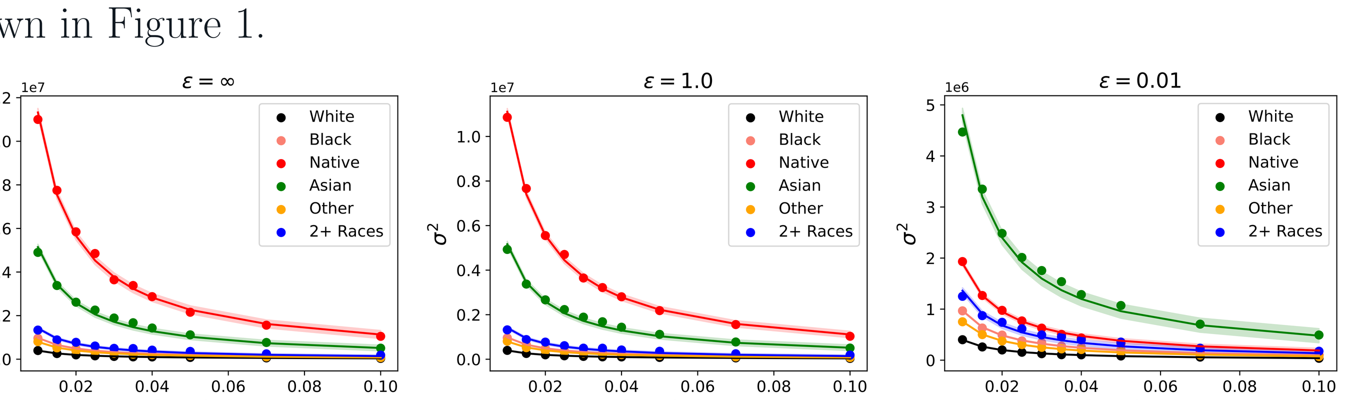
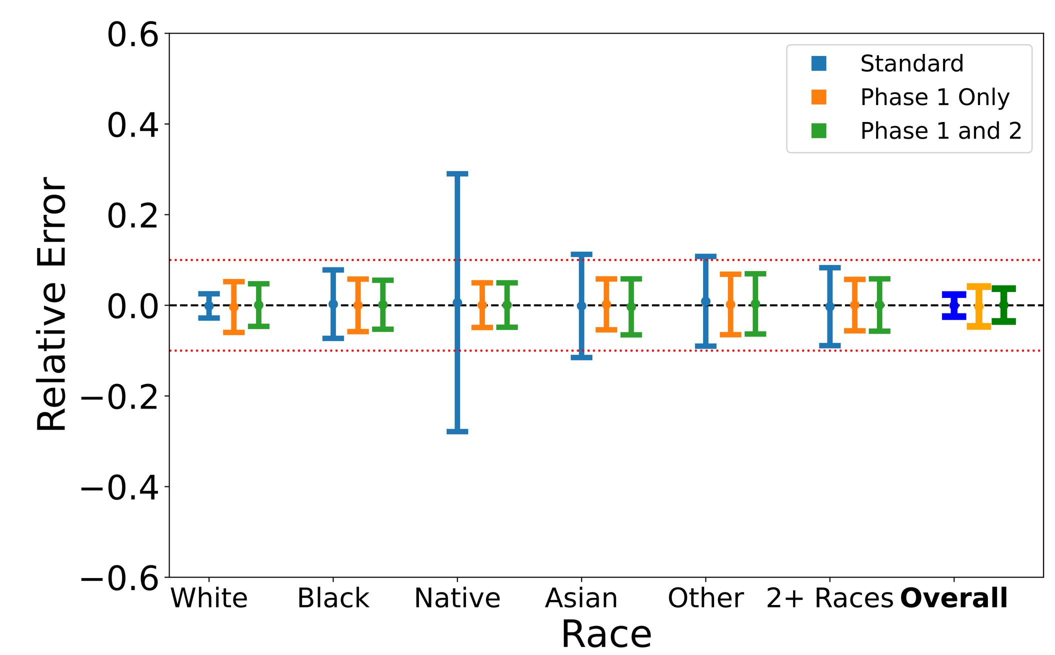


Fig. 1: Estimating the variance of mean *income* in Connecticut with different privacy budget ε.

Results

Optimized sampling process: Errors and Fairness (Fig. 2)

- **Standard Allocation:** Lowest error variance for overall but disproportionately affects minorities, leading to higher variance for these groups and failing to meet confidence constraints.
- **Phase 1 Only:** More uniform error variance across subgroups using the same survey cost. Surveys are allocated more equally, reducing variance for minorities and ensuring all groups meet confidence thresholds.
- **Phase 1 and 2:** Higher success rate in phase 2 at a higher cost per survey but lower overall cost (86% of Phase 1 Only). Uses simple random sampling in selected regions, prioritizing high-density areas of targeted populations. Slightly reduced performance and fairness compared to Phase 1 Only but meets confidence constraints for all groups.



Method	Sampling Rate	Survey Cost	ξ_{Var}	# of Violation
Standard	0.964%	100%	0.260	3/6
Phase 1 Only	0.964%	100%	**0.023**	**0/6**
Phase 1 and 2	**0.770%**	**86%**	0.030	**0/6**

Fig. 2: Relative group errors from estimating mean income in Connecticut

Differential Privacy (DP) [2] is a rigorous privacy notion that characterizes the amount of information of an individual's data being disclosed in a computation. Formally, a randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{R}$ with domain $\mathcal{X}$ and range $\mathcal{R}$ satisfies *ϵ-differential privacy* if for any output $O \subseteq \mathcal{R}$ and datasets $x, x' \in \mathcal{X}$ differing by at most one entry (written $x \sim x'$),

$$\Pr[\mathcal{M}(x) \in O] \leq e^{\epsilon} \Pr[\mathcal{M}(x') \in O] \tag{2}$$

In the context of this paper, the Laplace noise was added to the count N_i^r for every $i \in [G]$ and $r \in R$ to achieve *ϵ-differential privacy*. The noisy counts $\tilde{N}_i^r$ are then post-processed to ensure non-negativity, following the approach used by the U.S. Census [3]:

$$\tilde{N}_i^r = \max\left(0, N_i^r + Lap(\Delta x/\varepsilon)\right). \tag{3}$$

This non-negativity constraint introduces a positive bias, particularly affecting minority groups (Corollary 1), leading to an overestimation of their population as shown in Fig. 3.

Corollary 1. *The bias of the aggregated counts for each subgroup on the state level is*

$$\mathcal{B}(\tilde{N}_i) = \mathbb{E}\left[\tilde{N}_i\right] - N_i = \sum_{r \in [R]} \frac{\Delta x}{2\varepsilon} \exp\left(\frac{-N_i^r \varepsilon}{\Delta x}\right) > 0.$$

ε \ Race	White	Black	Native	Asian	Other	2+ Races	Total
∞	2,039,731	315,568	**7,571**	143,584	215,150	295,844	3,017,448
10	2,039,355	315,182	**7,524**	143,226	214,766	295,482	3,015,535
1	2,039,298	315,199	**7,699**	143,260	214,736	295,495	3,015,687
0.1	2,038,844	315,320	**10,681**	143,846	214,555	295,705	3,018,951
0.01	2,034,218	321,513	**42,068**	158,498	222,160	304,533	3,082,990

Fig. 3: Impact of DP on estimated population for each race in Connecticut

DP-sampling: Errors and Fairness (Fig. 4)

- **Standard Allocation:** Adding more noise *unexpectedly* reduces error variance for minorities because the strong positive bias overestimates minority population size, leading to higher survey allocation and improved fairness.
- **Phase 1 Only:** Insensitive to the variance of errors with respect to ε because the required number of samples does not depend on group size, maintaining consistent error variance regardless of noise levels.
- **Phase 1 and 2:** Slight changes in error variance with added noise, as noise affects region selection for Phase 2 based on prior population data, increasing the probability of incorrect region selection.

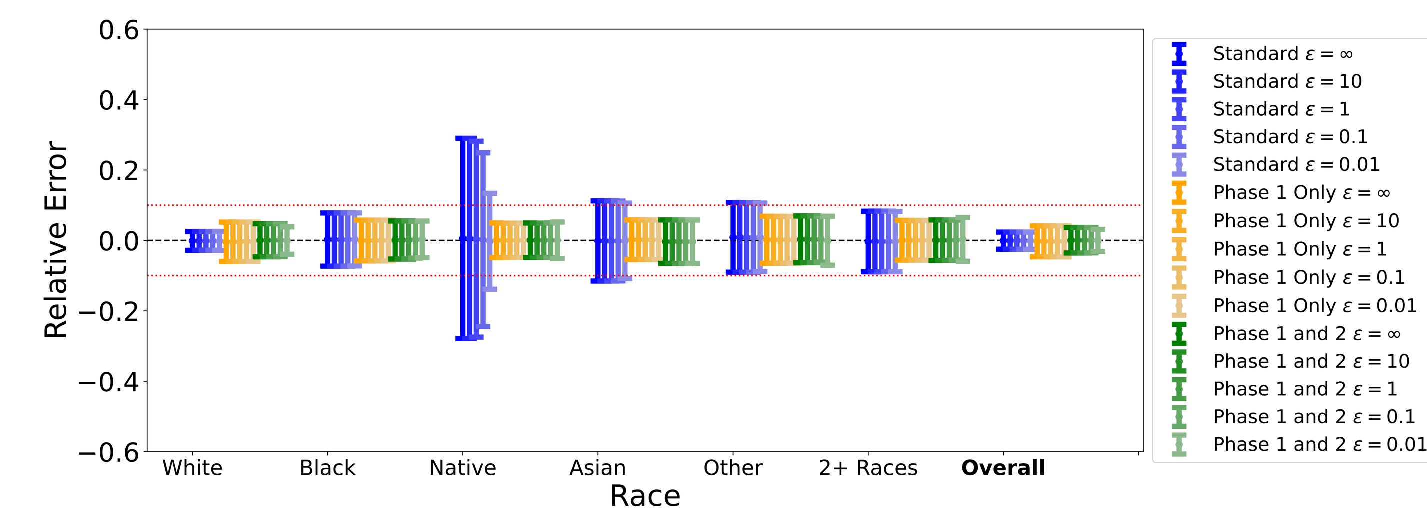


Fig. 4: Relative errors from estimating mean income in Connecticut using DP-noised $\tilde{N}_i^r$.

Conclusions

This work addresses unfairness in large surveys like the ACS, where traditional sampling methods disproportionately affect minority groups. We introduced an optimization-based framework to ensure fair error margins across all population segments while minimizing sampling costs. Surprisingly, we found that differential privacy can reduce unfairness by introducing beneficial positive biases for underrepresented populations. These findings demonstrate the effectiveness of our approach in enhancing fairness without compromising data utility or costs. Our results have significant implications for policy formulation and resource allocation, promoting equitable treatment of all demographic segments.

Acknowledgements

This research is partially supported by Dean's Undergraduate Engineering Fellowship from the University of Virginia. Its views and conclusions are those of the authors only.

References

[1] *American Community Survey and Puerto Rico Community Survey Design and Methodology.* Nov. 2022. URL: `https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/2022/acs_design_methodology_report_2022.pdf`.

[2] Cynthia Dwork et al. "Calibrating noise to sensitivity in private data analysis". English. In: *Theory Of Cryptography, Proceedings* 3876 (2006). 3rd Theory of Cryptography Conference ; Conference date: 04-03-2006 Through 07-03-2006, pp. 265–284. ISSN: 0302-9743.

[3] Matthew Spence. "What to Expect: Disclosure Avoidance and the 2020 Census Demographic and Housing Characteristics File". en. In: *United States Census Bureau* (May 2023). URL: `https://www.census.gov/newsroom/blogs/random-samplings/2023/05/what_to_expect_dhc.html`.