Effects of Differentially Private Noise Injection on Survey Operations

Quentin Brummet
NORC at the University of Chicago

Committee on National Statistics Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations
December 11, 2019

NORC at the UNIVERSITY of CHICAGO

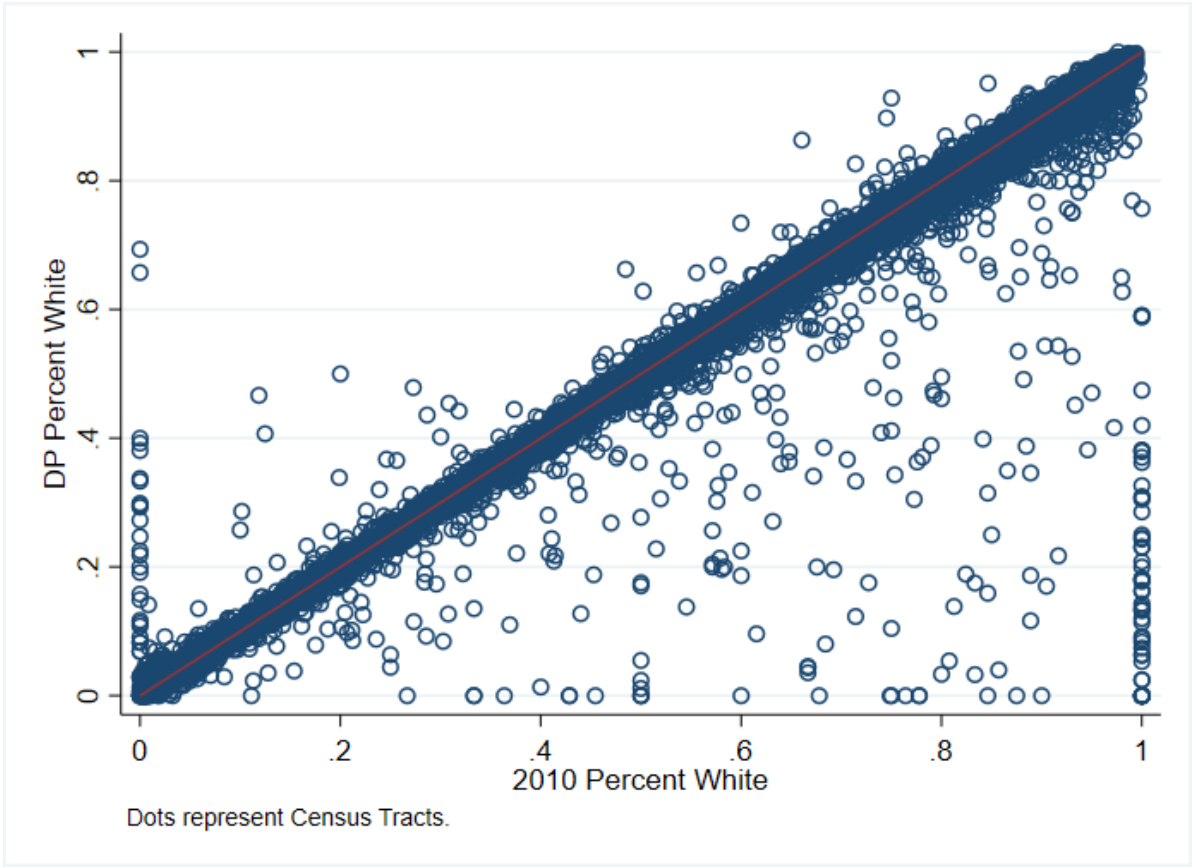
Motivation

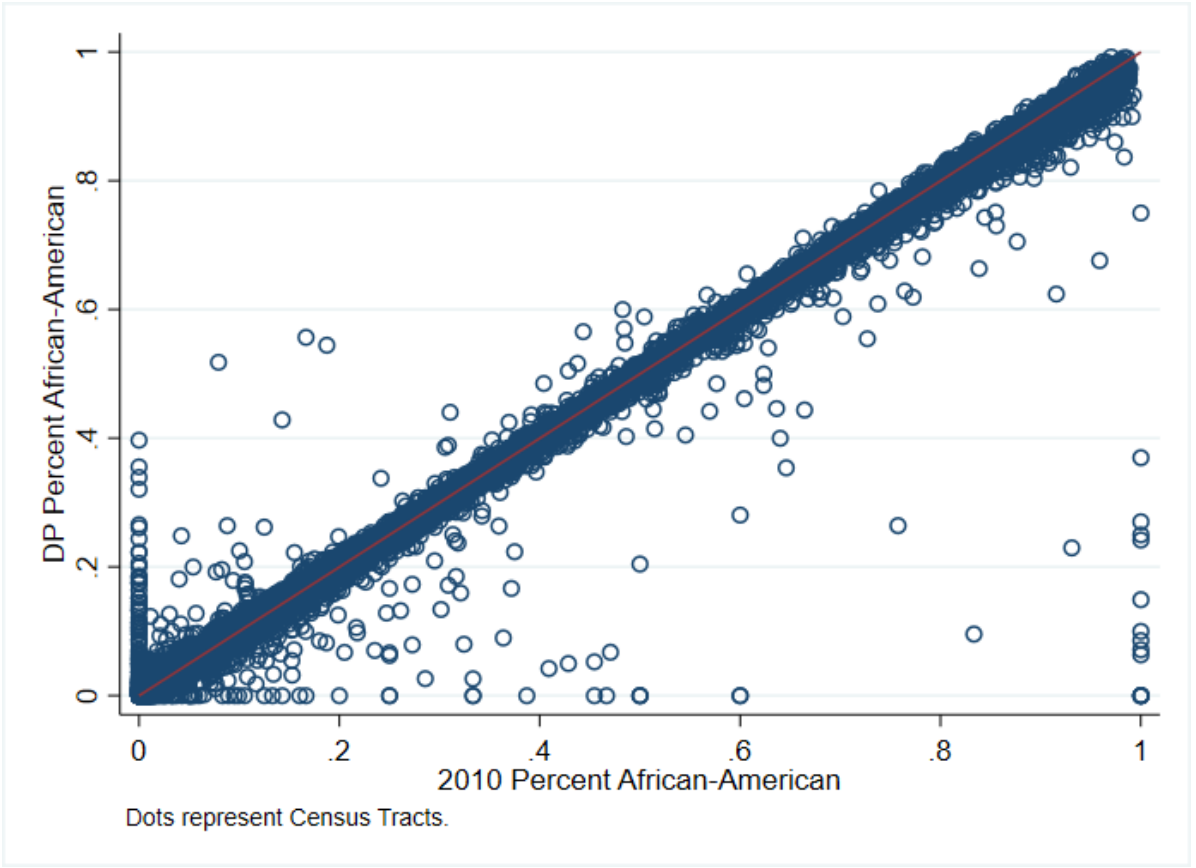
- Computational advances and the growing availability of data make re-identification of respondents from federal statistical data a growing possibility
- There are a number of decisions to make in applying DP to a large statistical release such as the 2020 Census
- How DP is applied (and what ϵ is chosen) depends on both privacy concerns and the effect of additional noise on the utility of data

Survey Operations and Census Data

- Many pieces of survey operations rely on Census data
  1. Sample design
     - How accurately can we predict whether a given household will have a specific characteristic?
     - Typically uses on-spine geographies, especially Census tracts
  2. Fielding
     - Example: prioritization of cases using block- and block group-level Census data on vacancy and household composition
     - This sort of use will likely be difficult with DP data, although the impact is hard to quantify
  3. Post-data collection (e.g., weighting and imputation)
     - Typically use aggregates at high levels of geography, which are less affected by DP
- The rest of the talk will focus specifically on sample design

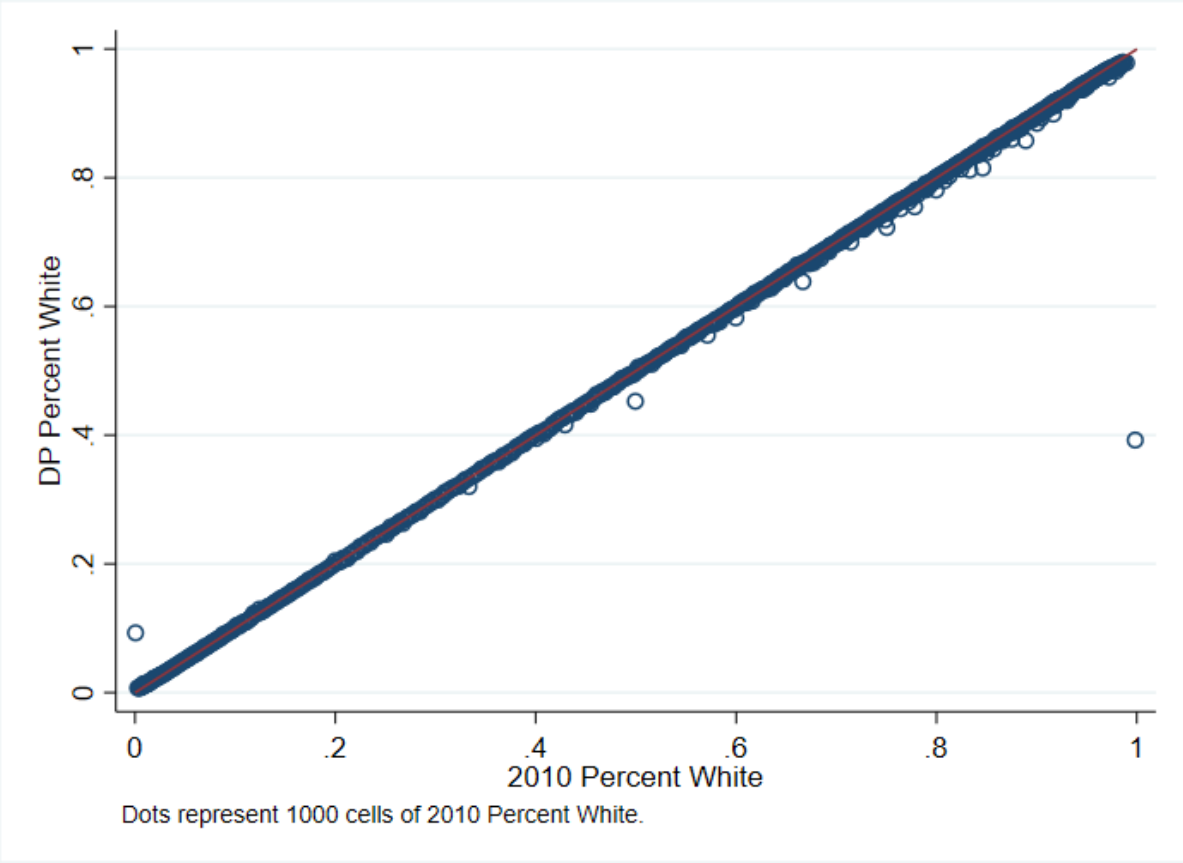
Tract-Level Racial Composition in 2010 Census and DP Data

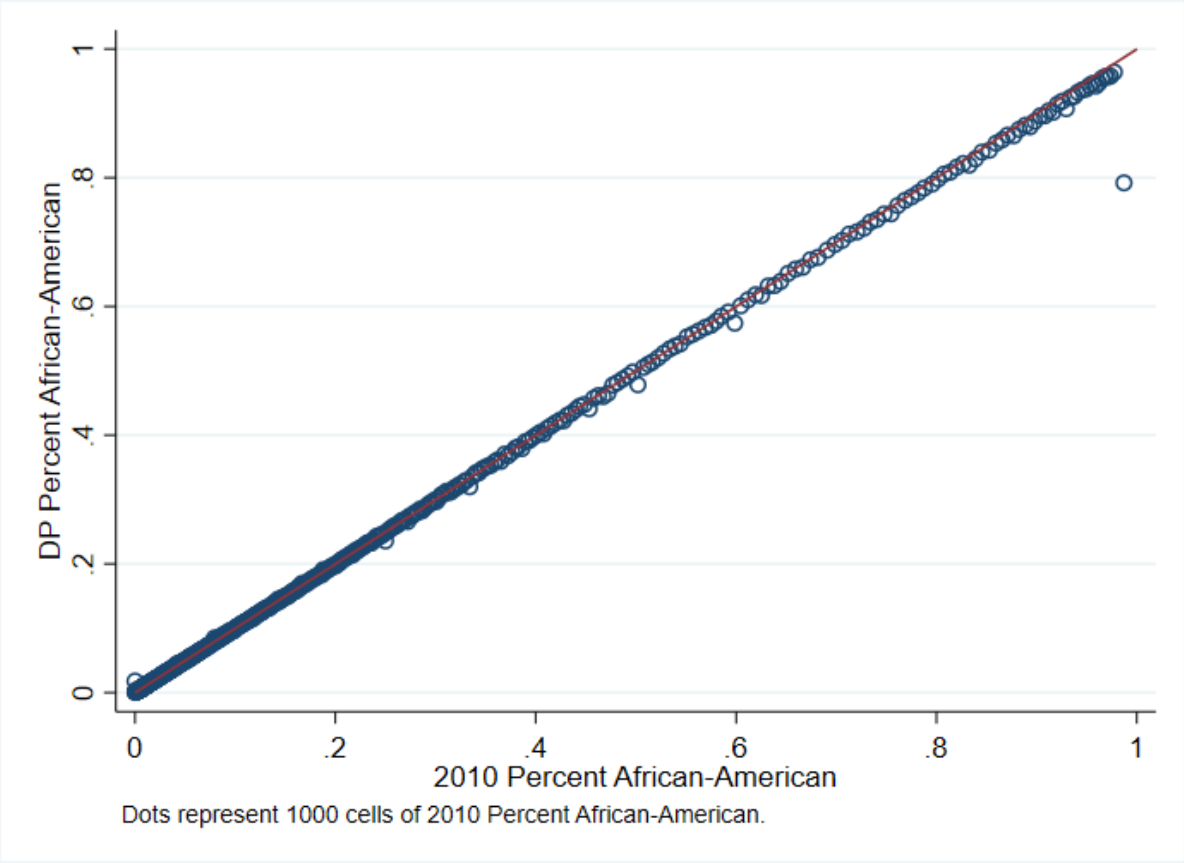




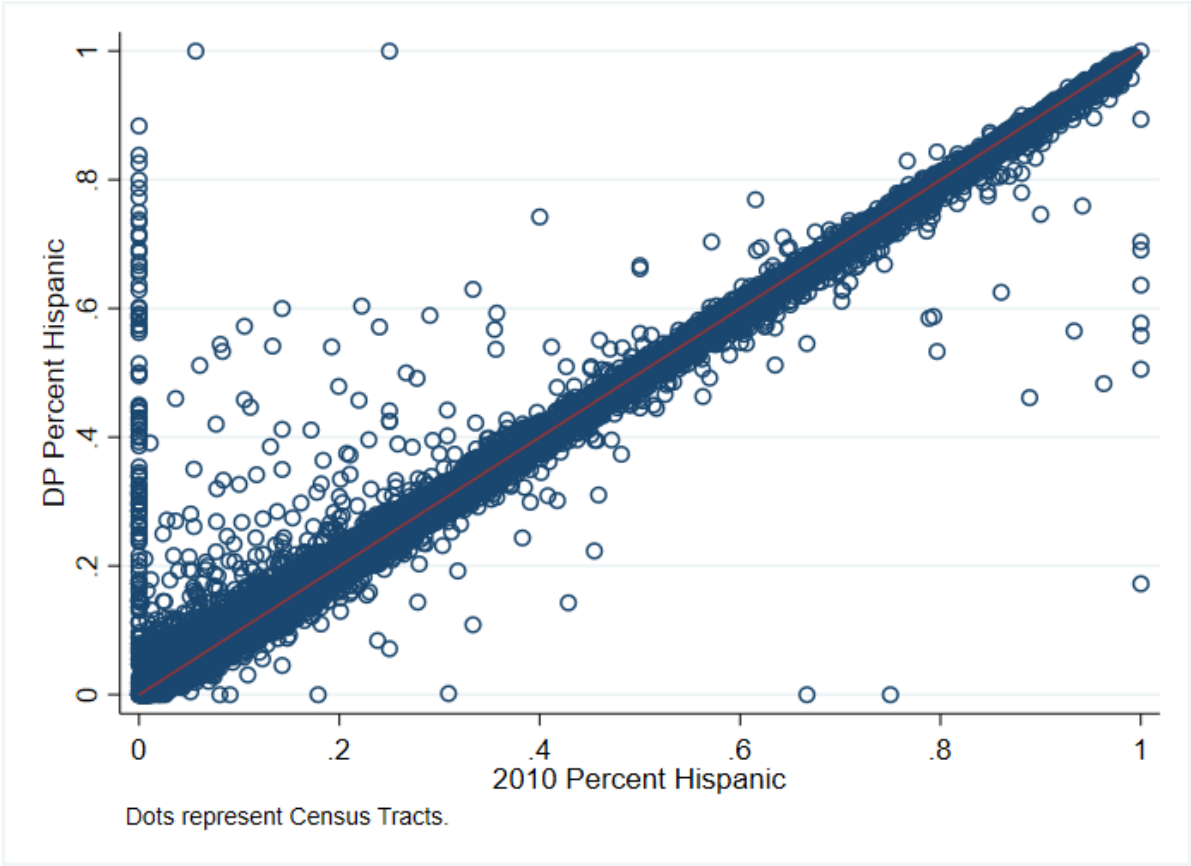
NORC at the UNIVERSITY of CHICAGO

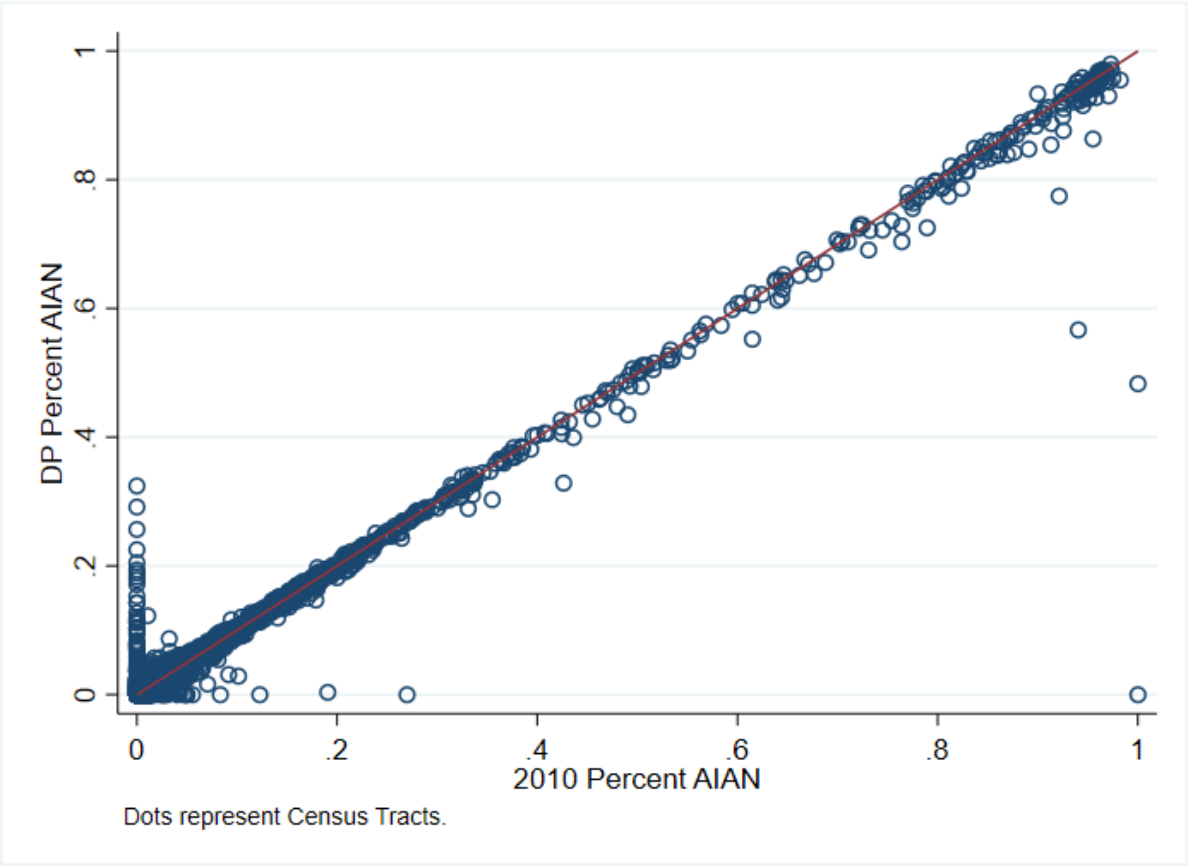
Tract-Level Racial Composition in 2010 Census and DP Data





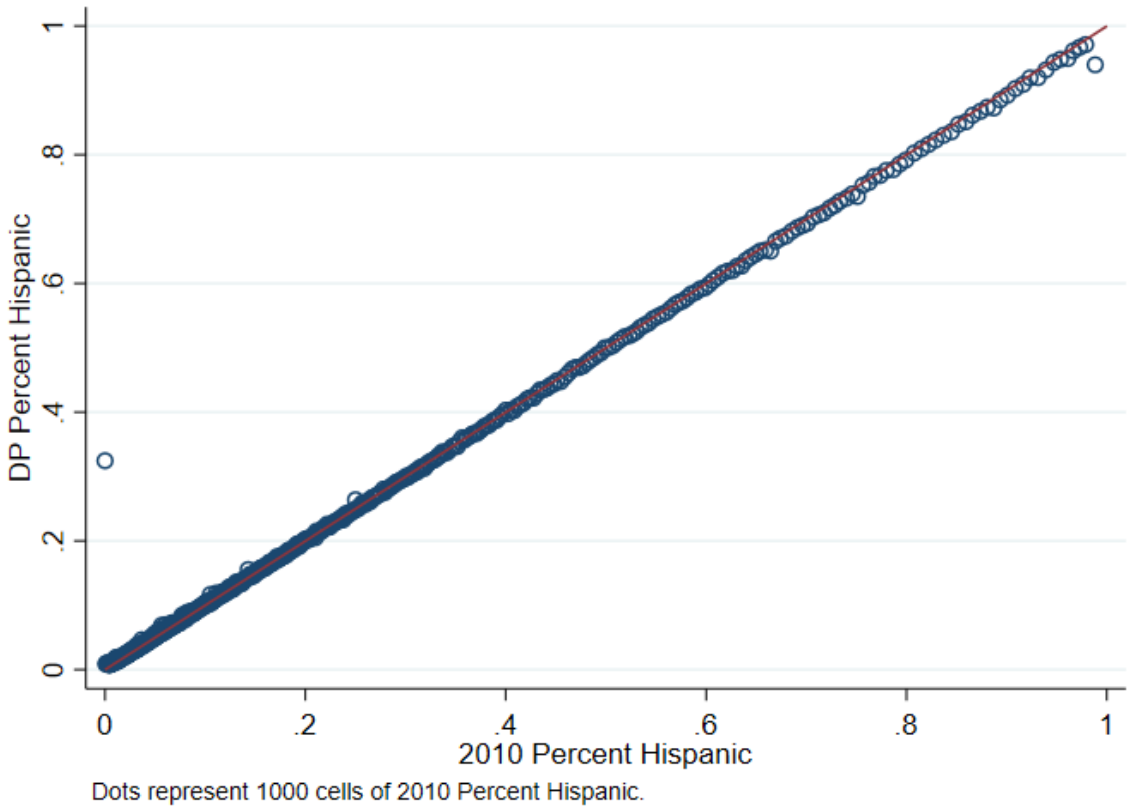
Tract-Level Racial Composition in 2010 Census and DP Data

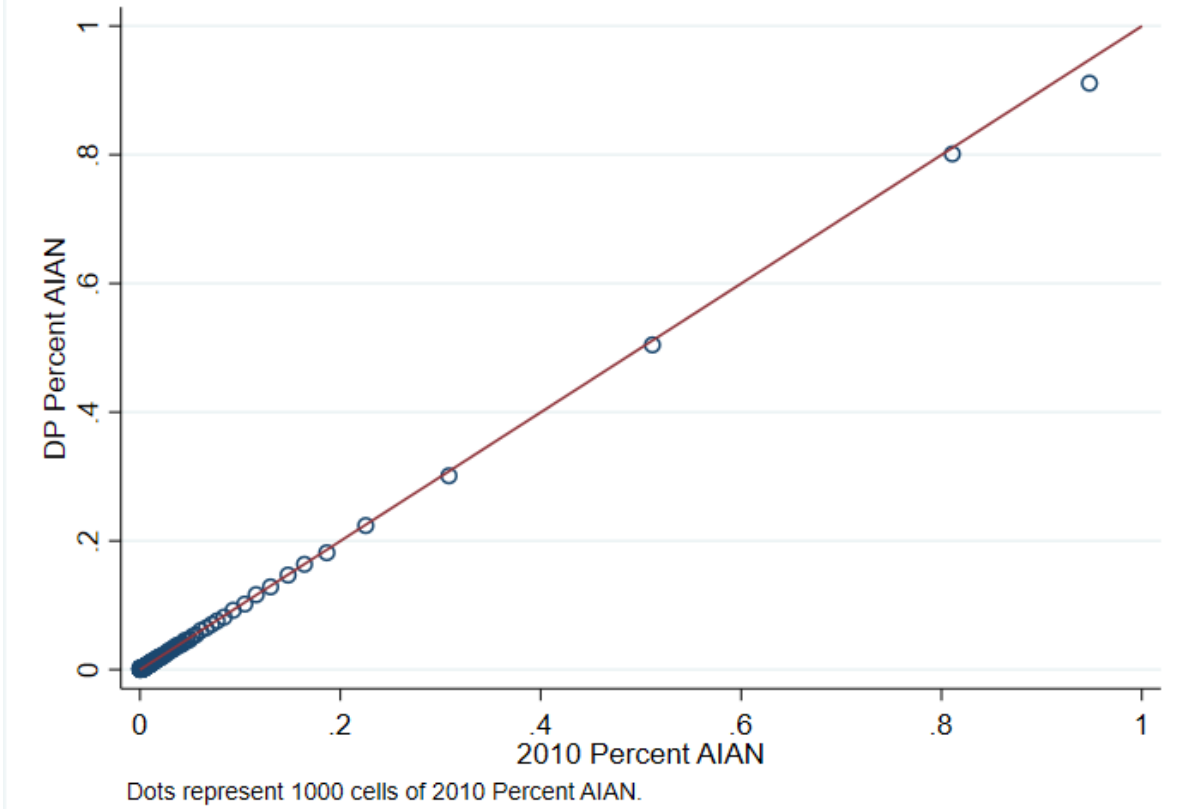




NORC at the UNIVERSITY of CHICAGO

Tract-Level Racial Composition in 2010 Census and DP Data





NORC at the UNIVERSITY of CHICAGO

Analysis of DP Noise and Sample Design

- To understand the effects of the additional noise from DP, will run through a hypothetical example of a survey with DP and 2010 Census data

- Analysis will compare DP demonstration data with previously released 2010 Census data
  - DP data from Minnesota Population Center and Cornell Institute for Social and Economic Research
  - 2010 Census data from SF1, which includes disclosure avoidance protections

Surveying a Rare Population

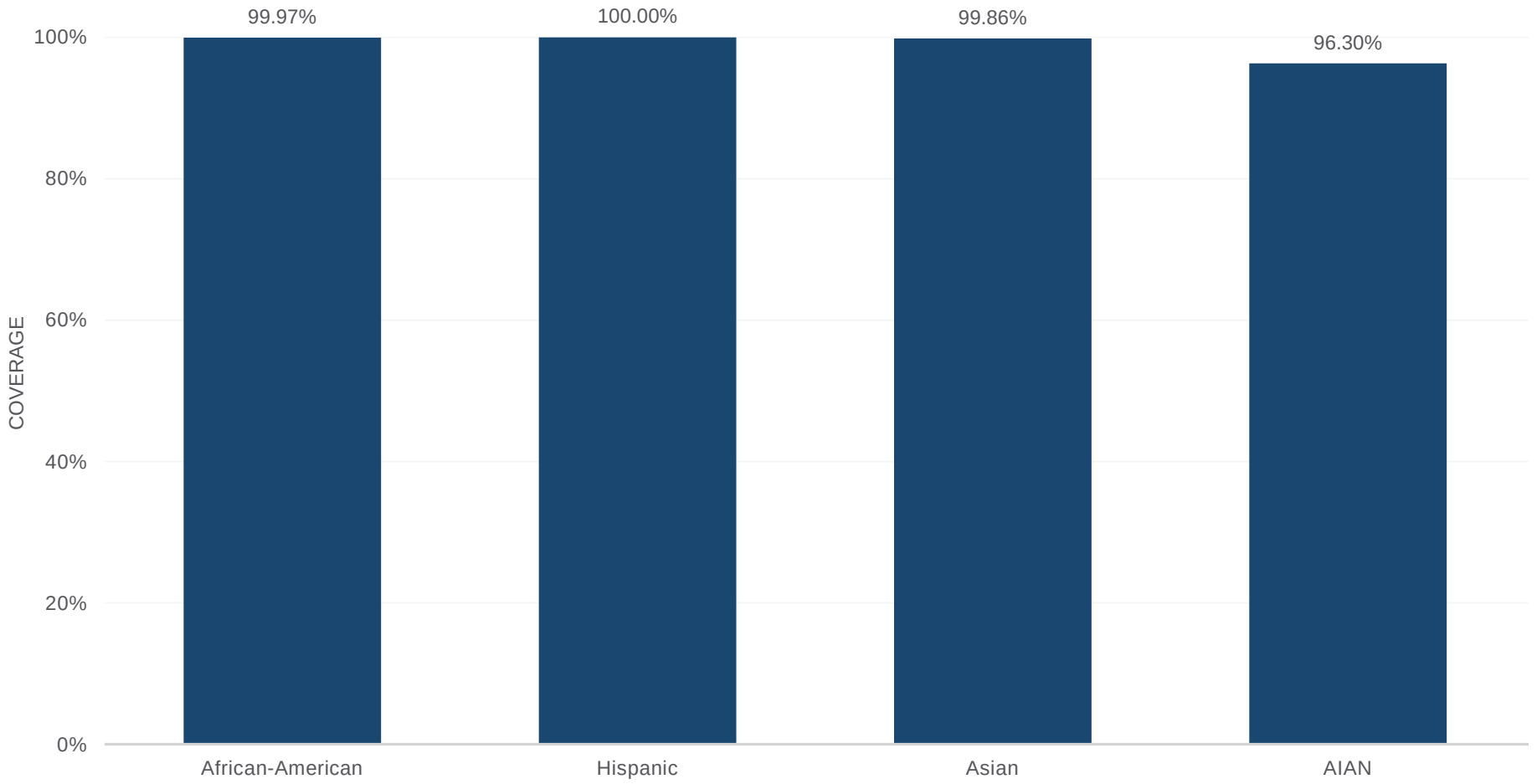
- Construct a hypothetical sample frame including areas with greater than 0 population for a given group
- Divide the sample into two strata based on if the enumeration district has least 30% of a given population
- Two stages of sampling:
  - Sample census tracts:
    - Probability proportional to size of the target population in the low-density stratum
    - Probability proportional to 5 times the size of the target population in the high-density stratum
- After tracts have been sampled, screen an equal number of individuals from each enumeration district in order to obtain a sample

Outcomes

1. Coverage: Given that the sampling frame will only include census tracts with greater than zero population for a given group, what fraction of the target population will not be included in the sampling frame?

2. Efficiency: What is the projected increase in survey costs?

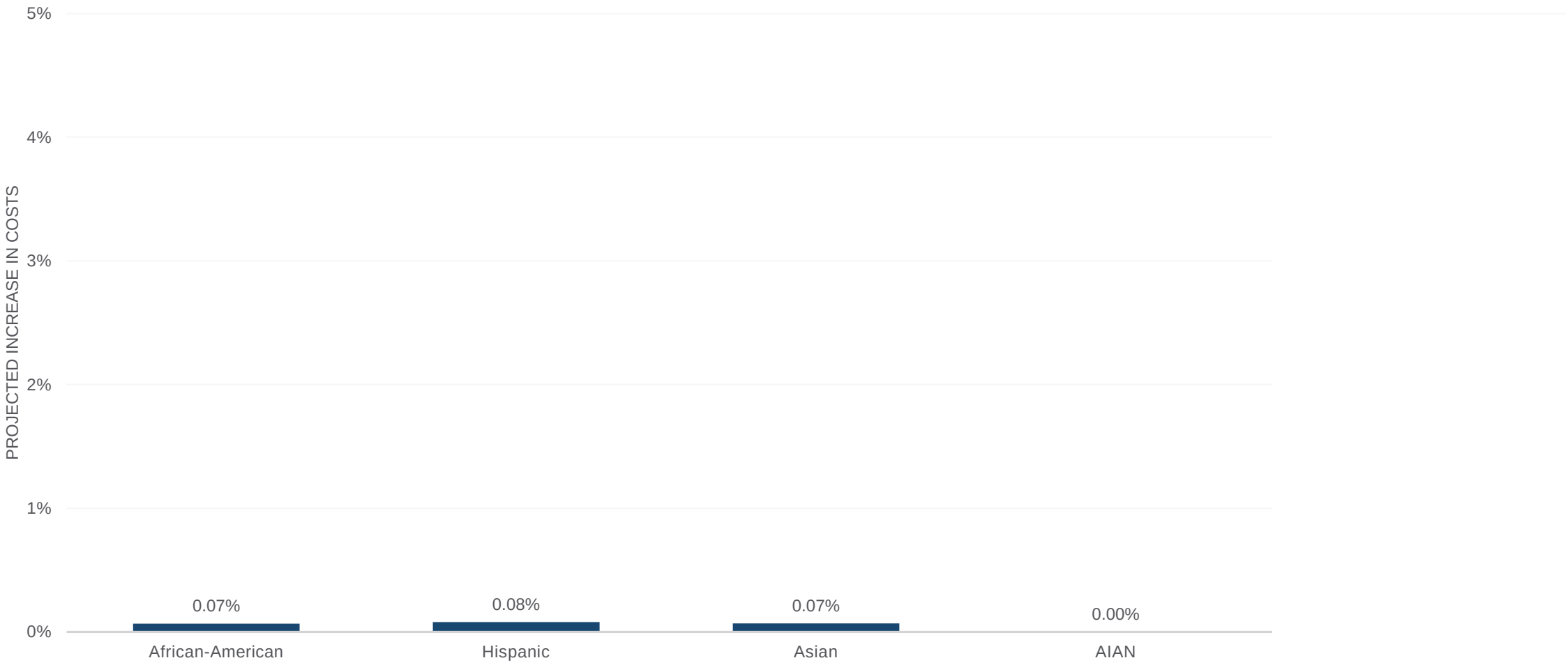
Projected Coverage for a Survey of Major Racial/Ethnic Groups



Note: Bars represent the percent of the given population in the 2010 Census data that resides in an enumeration district with >0 individuals in the DP data.

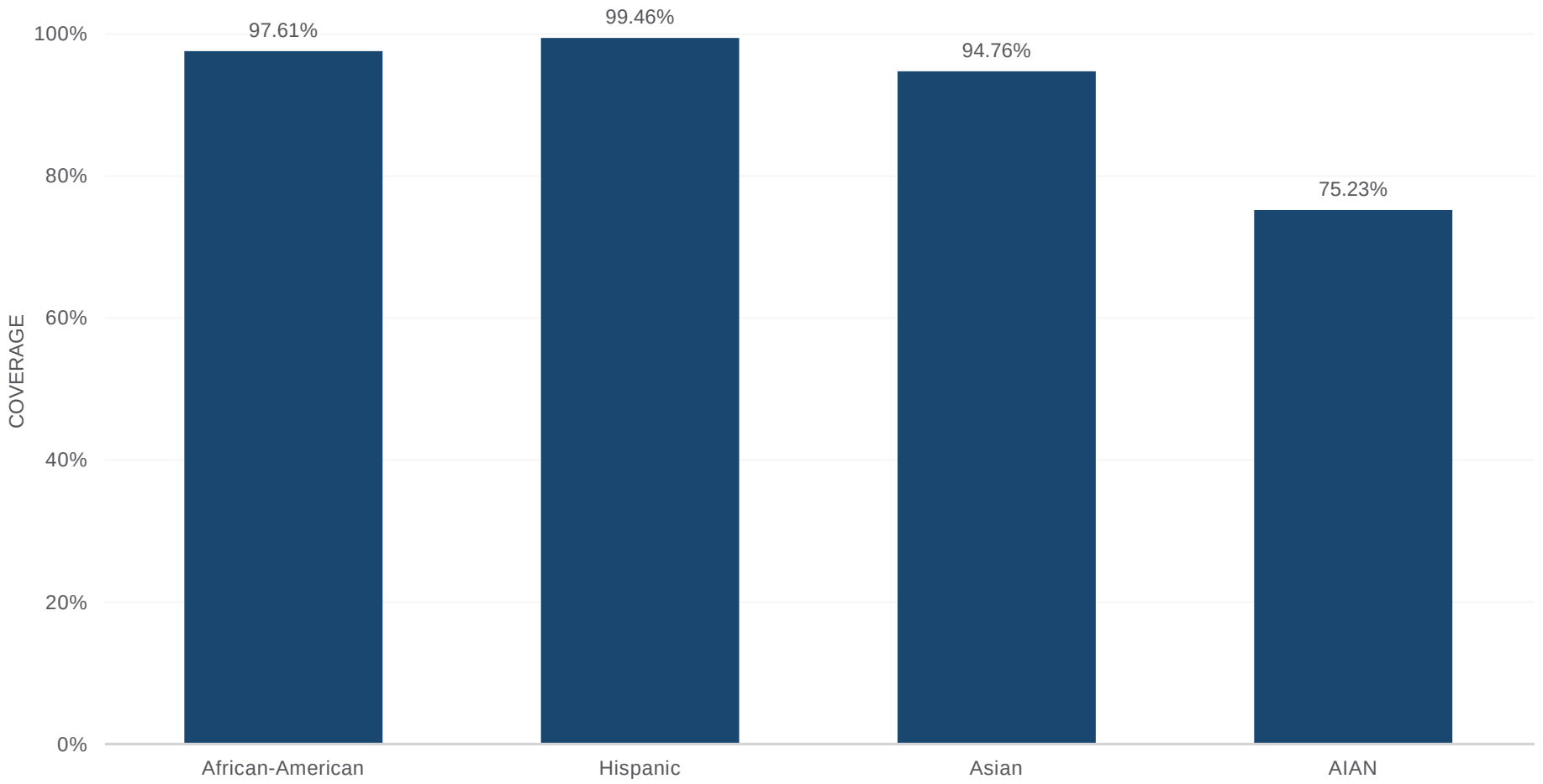
NORC at the UNIVERSITY of CHICAGO

Projected Increase in Costs for a Survey of Major Racial/Ethnic Groups



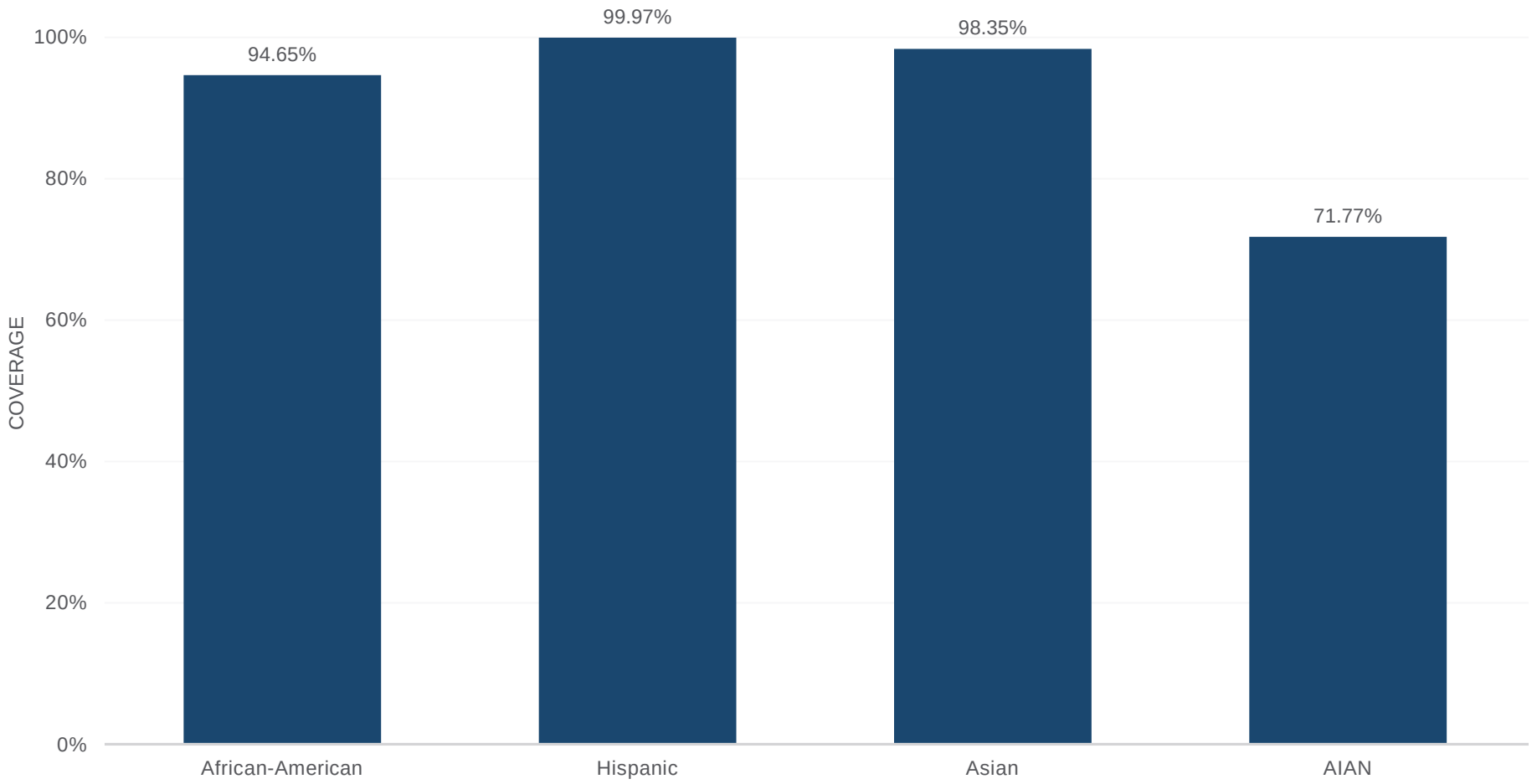
Notes: Graph displays the projected increase in survey costs to achieve a given sample size for a given group. Census tracts are sampled with probability proportional to size of the target population, sampling tracts with at least 30% of a given group at five times the rates. We assume that all individuals agree to participate in the survey and that the only costs are screening interviews to identify individuals to participate in the survey.

Projected Coverage for a Survey of 25-49 Year Old Males



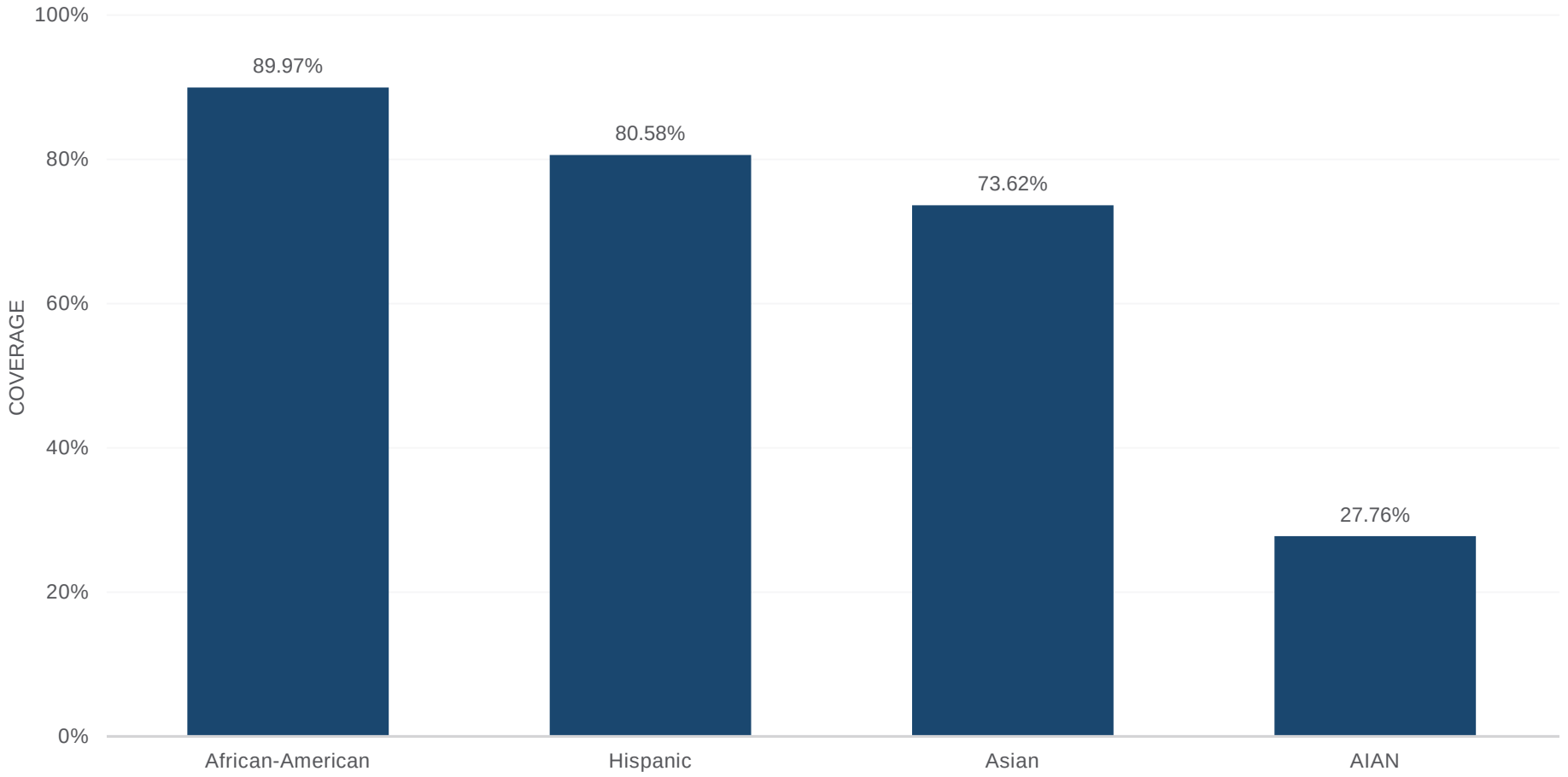
Note: Bars represent the percent of the given population in the 2010 Census data that resides in an enumeration district with >0 individuals in the DP data.

Projected Coverage for a Survey of 25-49 Year Old Males in California



Note: Bars represent the percent of the given population in the 2010 Census data that resides in an enumeration district with >0 individuals in the DP data.

Projected Coverage for a Survey of 25-49 Year Old Males in West Virginia



Note: Bars represent the percent of the given population in the 2010 Census data that resides in an enumeration district with >0 individuals in the DP data.

Discussion

- Uses of block- and block group-level data on household composition and vacancy for fielding purposes may be difficult with DP data
- Effects of DP noise on survey costs in the context of sample design are projected to be minimal
  - For example, the oversampling of non-white households in a nationally representative survey would likely be unaltered
- In the context of smaller demographic groups, DP data can result in coverage issues
- Weighting relies on accurate population benchmarks, and any error in the population totals will propagate to other products

Other Considerations

- Factors working in favor of DP having a small impact:
  - On-spine geography, especially tracts
  - Averaging out of noise across geographic areas

- Caveats:
  - The American Community Survey is currently used for many survey purposes
  - It is still unclear how users might account for the error induced by DP
  - The level of noise added for a given epsilon is not constant and might change for the production 2020 Census

Brummet-Quentin@norc.org

Thank You!

NORC at the UNIVERSITY of CHICAGO

insight for informed decisions™