UNCERTAINTY IN DEMOGRAPHIC AND SOCIOECONOMIC DATA

THE USE OF DIFFERENTIAL PRIVACY FOR DISCLOSURE CONTROL, AND ITS POTENTIAL IMPACT ON AGE AND RACE/ETHNICITY COUNT

DAVID VAN RIPER, UNIVERSITY OF MINNESOTA
SETH SPIELMAN, UNIVERSITY OF COLORADO

SOURCES OF RACE/ETHNICITY DATA

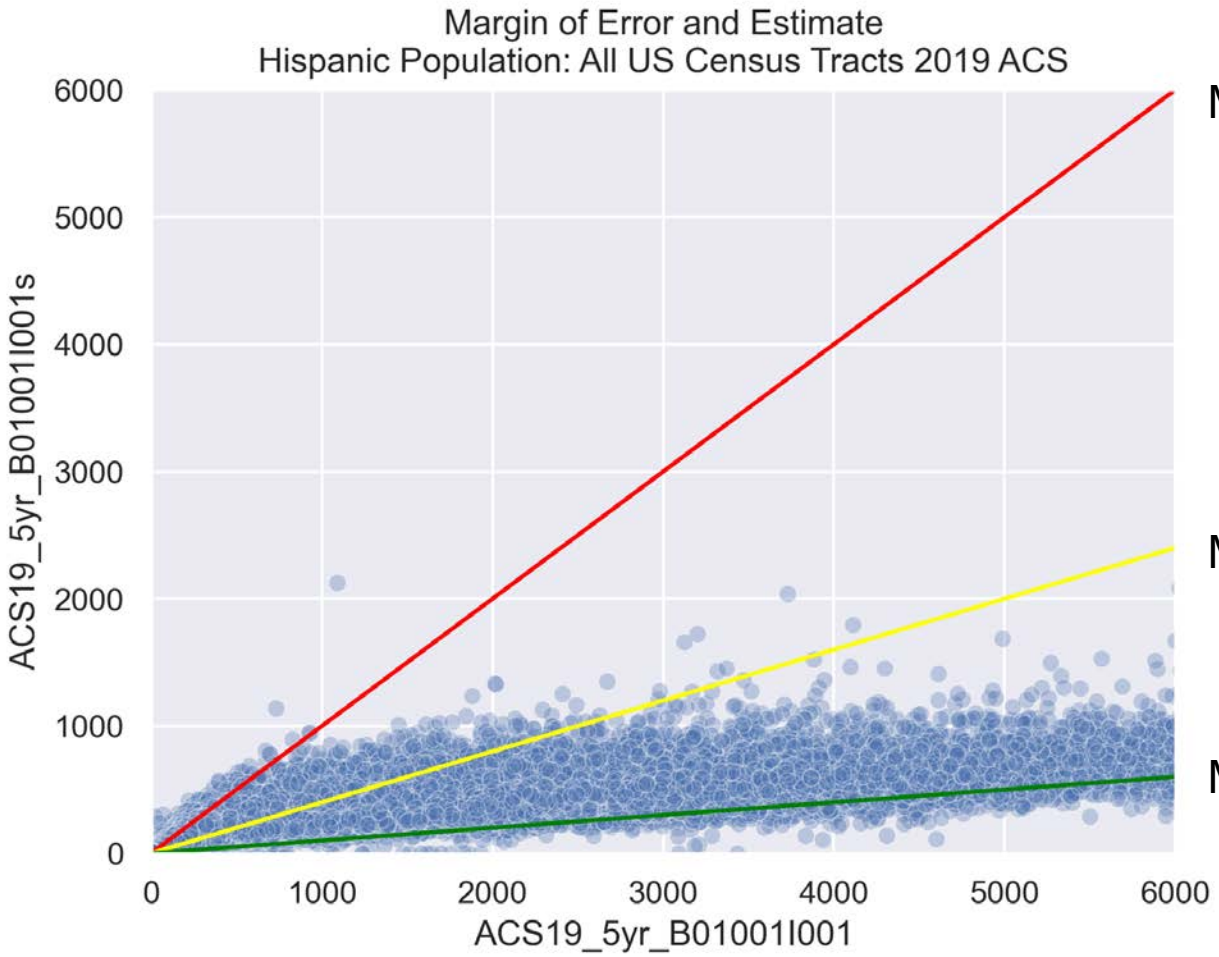
American Community Survey

- ACS is a sample of housing units.
- 3.5M housing units are sampled, about 2M responses are collected (~60% response rate).
- ACS population counts are estimates, from a sample, and and as a result carry uncertainty.

Decennial Census

- The Decennial Census is complete enumeration of the US population.
- Published data have always obfuscated responses to prevent reidentification.
- For 2020 Census is adopting a formal privacy framework which injects noise into the data.

ACS ESTIMATE QUALITY: HISPANIC POP BY TRACT



Margin of error 100% of the estimate

15% of all Census Tracts have a margin of error that is 100% of more of the estimate.

Margin of error 50% of the estimate

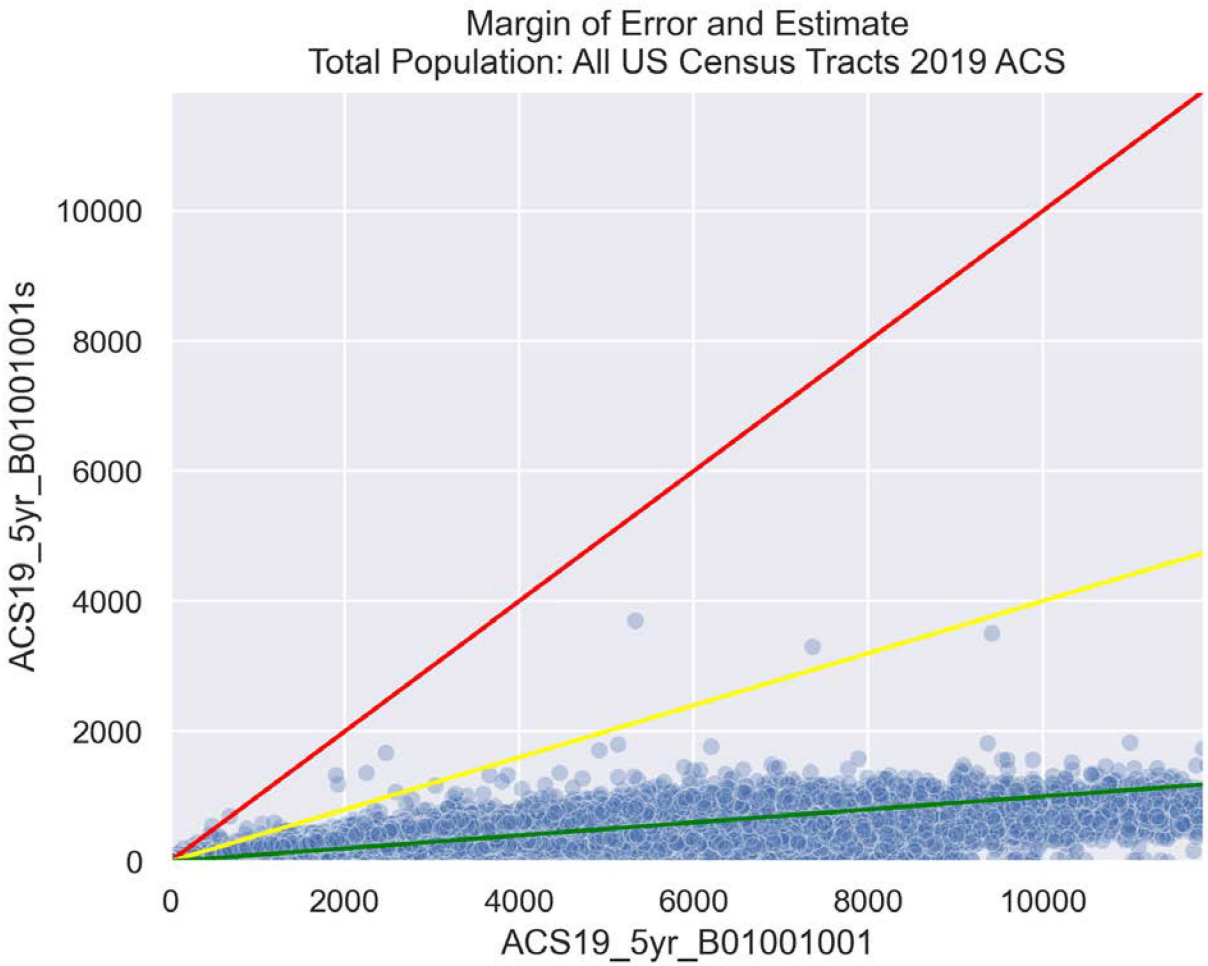
53% of all Census Tracts have a margin error that is 50% of more of the estimate.

Margin of error 10% of the estimate

99% of all Census Tracts have a margin of error that is 10% of more of the estimate.

Note: Zero estimates included in figure but excluded from % of tracts > than 100%, 50%, 10% thresholds
Figures truncate outliers, excluding the largest 1% of areas

ACS ESTIMATE QUALITY: TOTAL POP BY TRACT



Margin of error 100% of the estimate

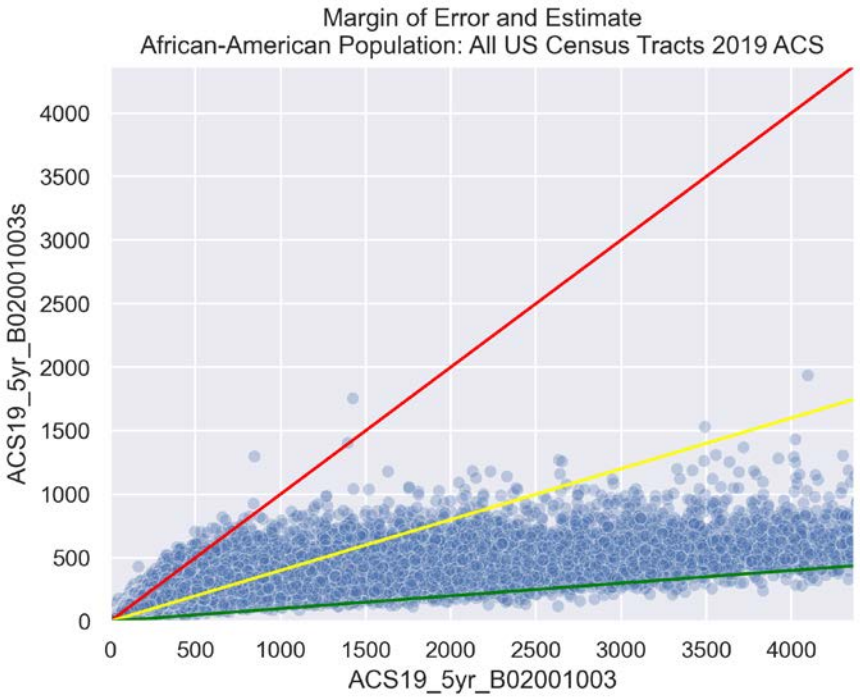
Less than 1% of Census Tracts have a margin of error that is 100% or more of the estimate.

Margin of error 50% of the estimate

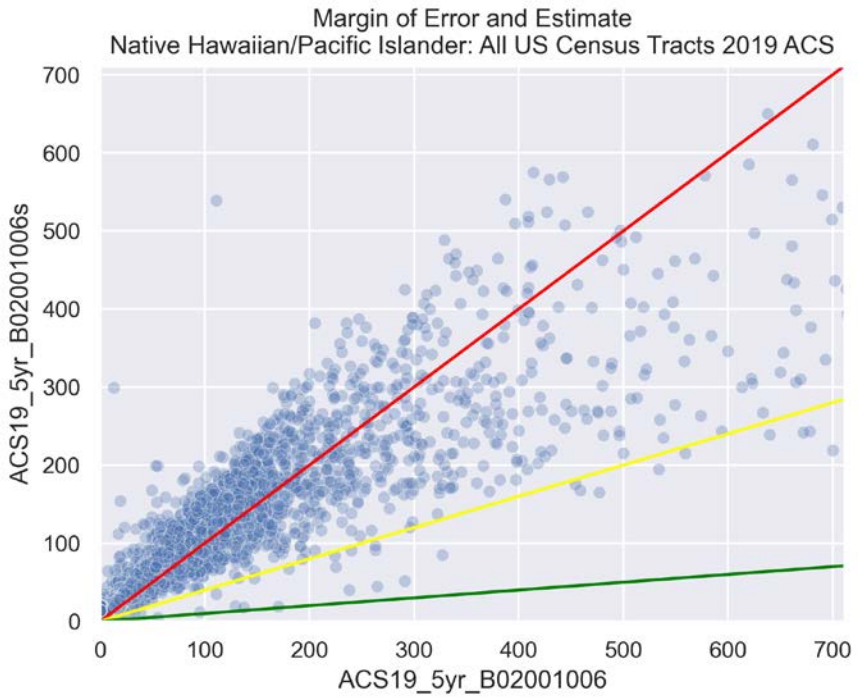
Less than 1% of all US Census Tracts have a margin error that is 50% of more of the estimate.

Margin of error 10% of the estimate

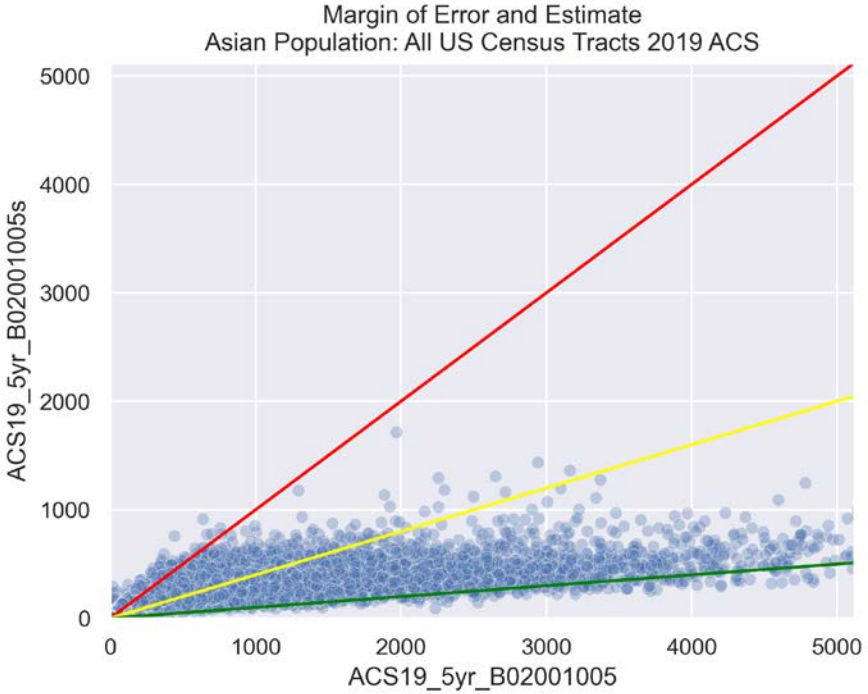
35% of all US Census Tracts have a margin of error that is 10% of more of the estimate.



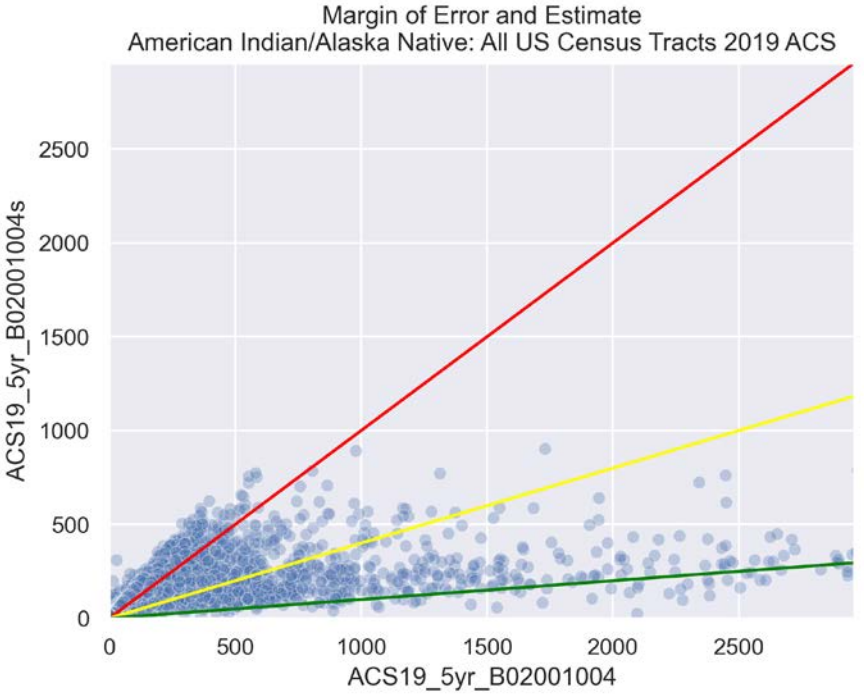
African-American Population
moe > 100% of estimate: 26.33% of tracts
moe > 50% of estimate: 61.55% of tracts
moe > 10% of estimate: 99.63% of tracts



Native Hawaiian/Pacific Islander
moe > 100% of estimate: 89.61% of tracts
moe > 50% of estimate: 98.92% of tracts
moe > 10% of estimate: 99.98% of tracts

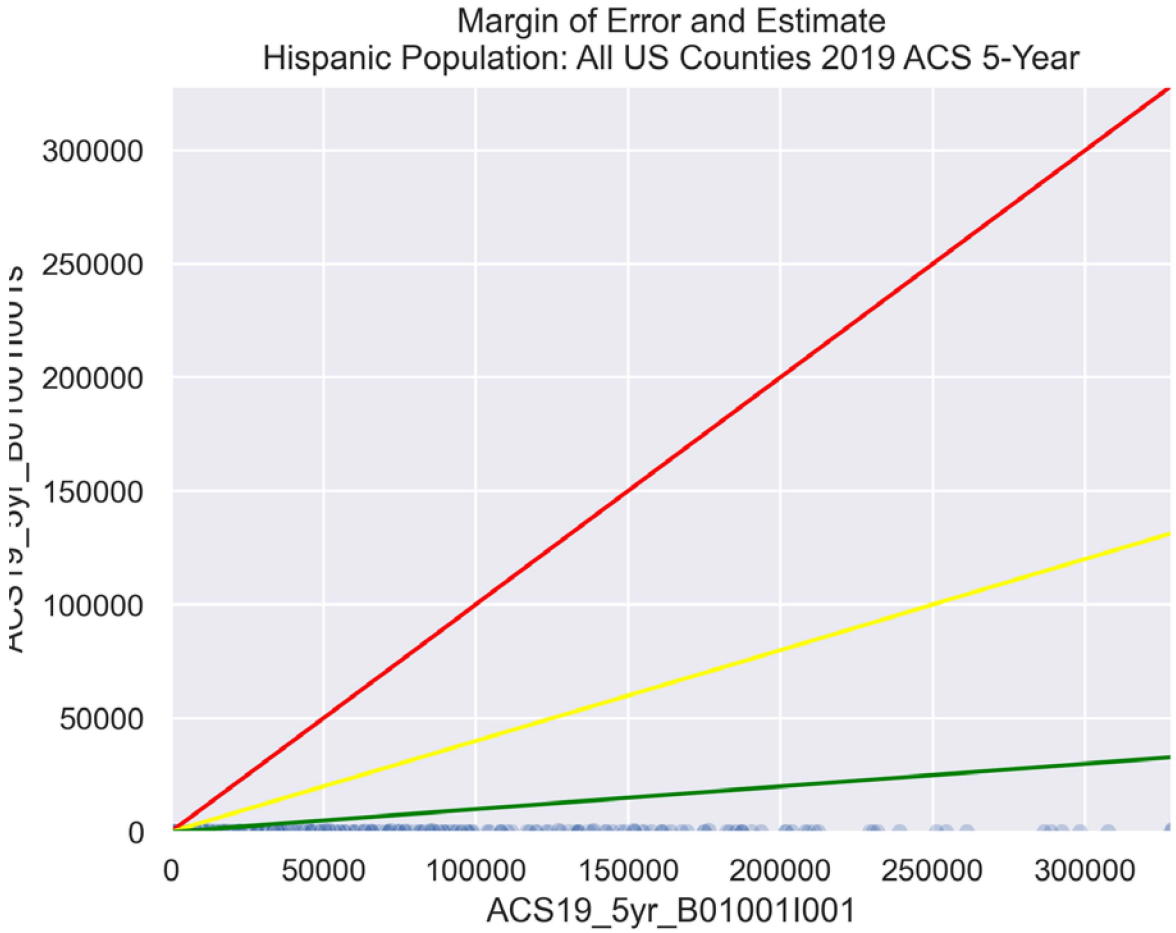


Asian Population
moe > 100% of estimate: 37.75% of tracts
moe > 50% of estimate: 75.71% of tracts
moe > 10% of estimate: 99.68% of tracts



American Indian/Alaska Native Pop
moe > 100% of estimate: 78.29% of tracts
moe > 50% of estimate: 96.26% of tracts
moe > 10% of estimate: 99.74% of tracts

ACS ESTIMATE QUALITY: HISPANIC POP BY COUNTY



Margin of error 100% of the estimate

3% of all counties have a margin of error that is 100% of more of the estimate.

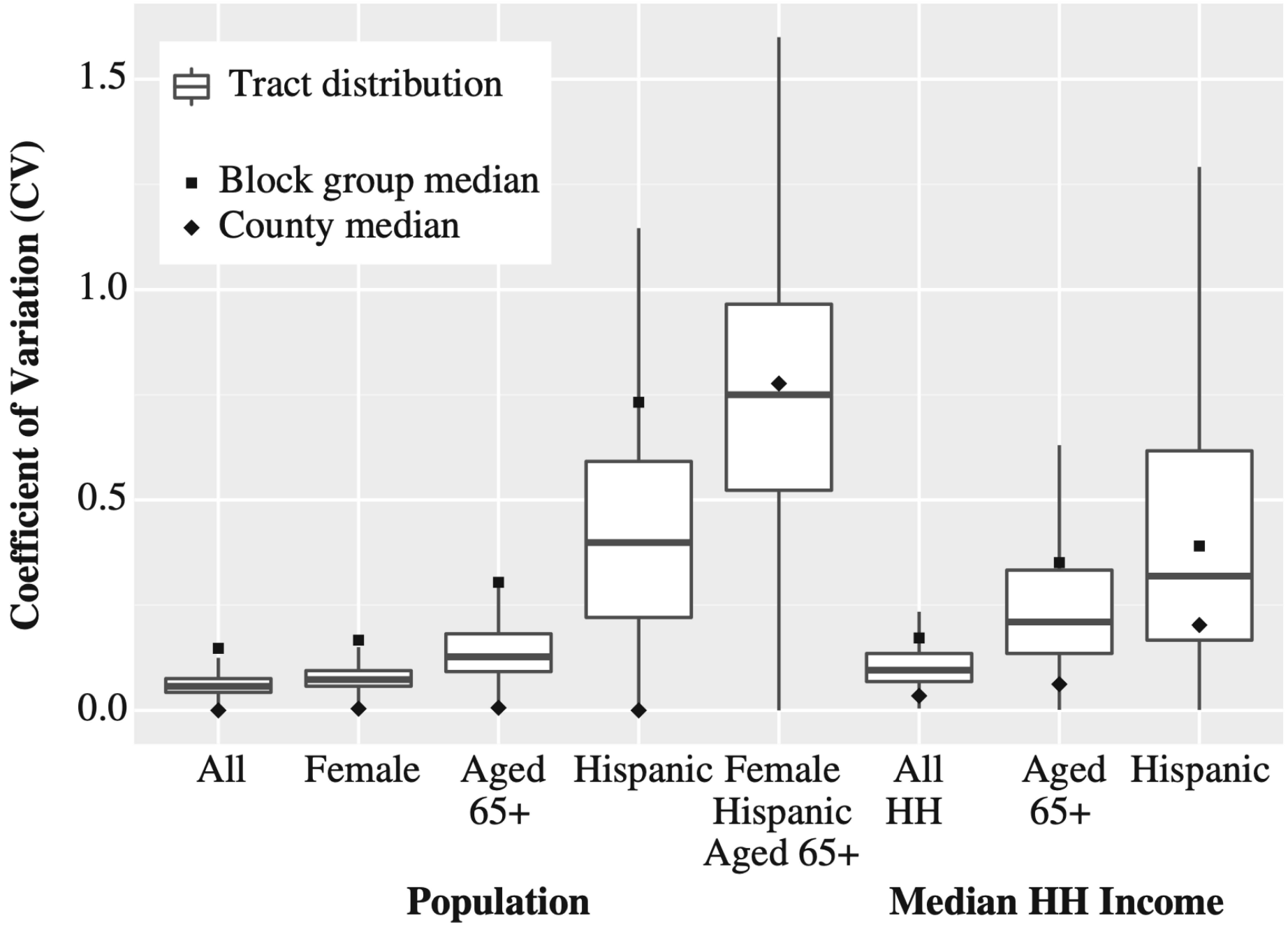
Margin of error 50% of the estimate

10% of counties have a margin error that is 50% of more of the estimate.

Margin of error 10% of the estimate

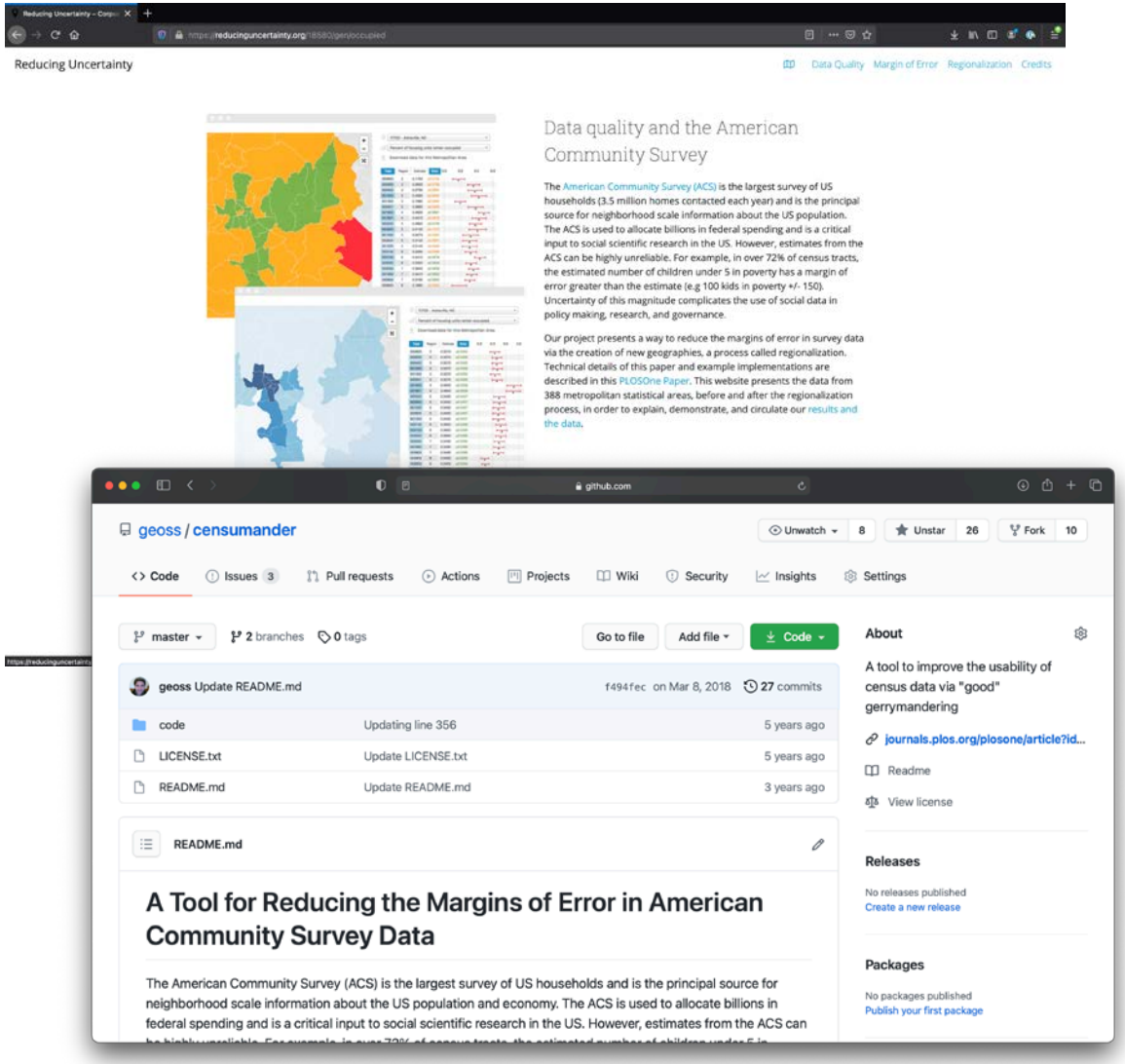
13% of all counties have a margin of error that is 10% of more of the estimate.

Note: Figure truncate outliers, excluding the largest 5% of counties


Tract distribution
Block group median
County median
Coefficient of Variation (CV)
1.5
1.0
0.5
0.0
All
Female
Aged 65+
Hispanic
Female Hispanic Aged 65+
All HH
Aged 65+
Hispanic
Population
Median HH Income

A SOLUTION?

- Census tracts can be too small to provide reliable racial/ethnic group estimates. Counties can be too large, meaningful intra metropolitan socio-spatial variation is lost.
- NYC Approach: Define custom “Neighborhood Tabulations Areas” bespoke groups of tracts.
- We’ve developed software and a website to generate “optimal” geographies based on user defined constraints (margin of error, population size).
- There are methods to refine census/ACS estimates using ancillary data sources.



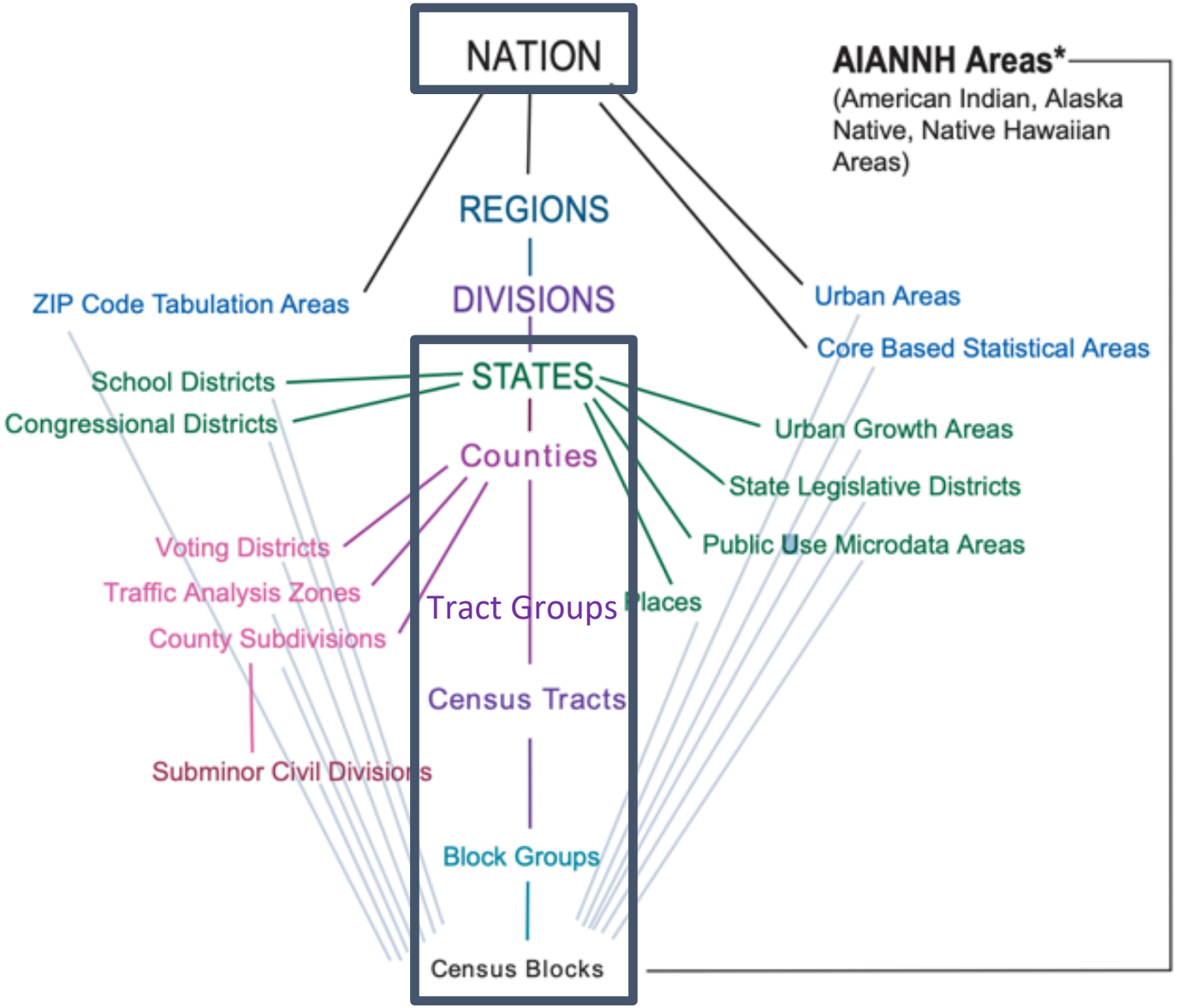
DECENNIAL 2020: A NEW APPROACH TO DISCLOSURE AVOIDANCE

Swapping (2010 and earlier)

Noise infusion (2020)

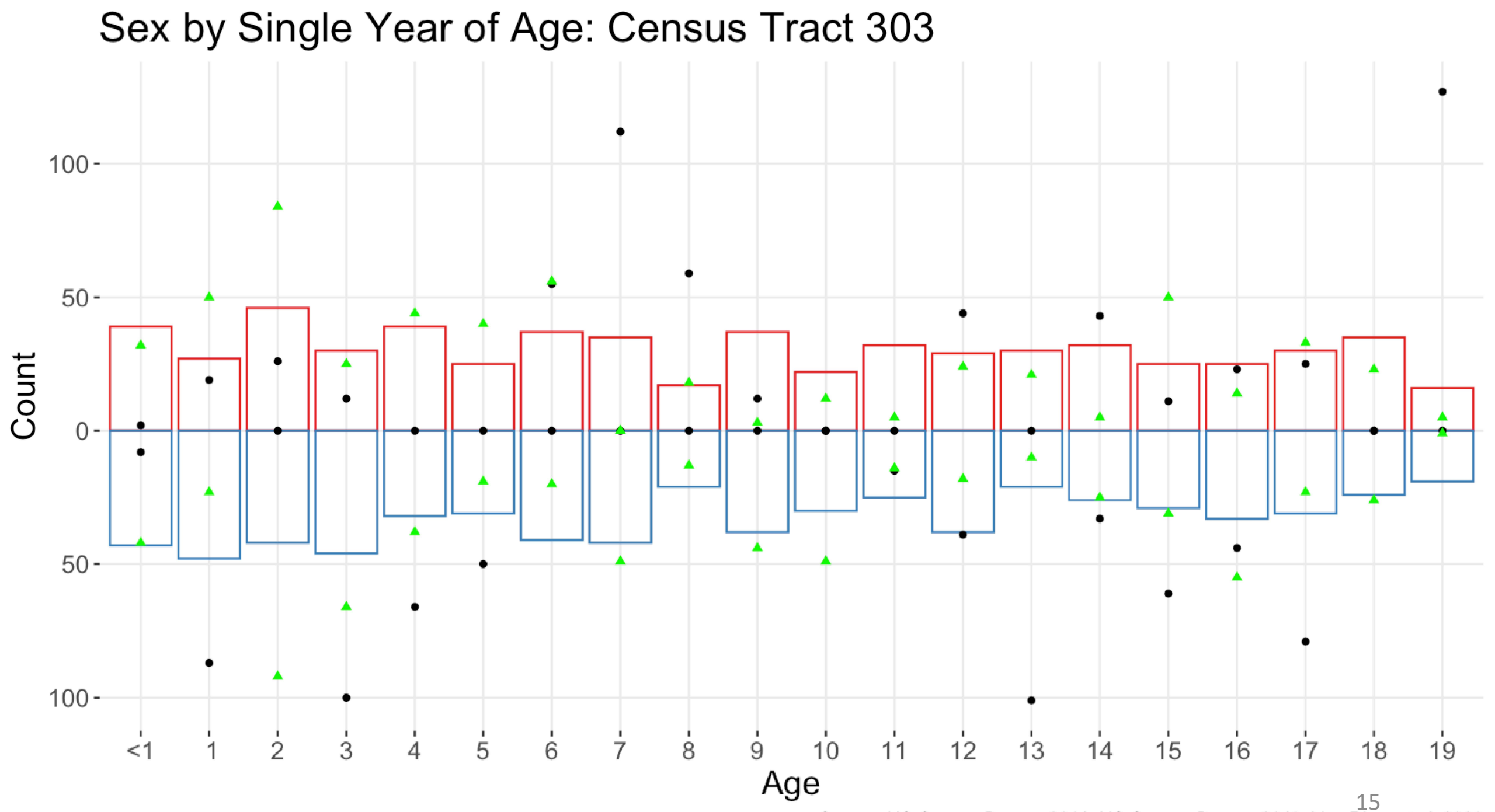
	School Attendance		
	Never	Attending	Past
Male	3 - 1 = **2**	12 + 0 = **12**	33 + 1 = **34**
Female	4 + 8 = **12**	17 + 2 = **19**	31 - 2 = **29**
	N = 100	N = 108	

DECENNIAL 2020: POLICY DECISIONS

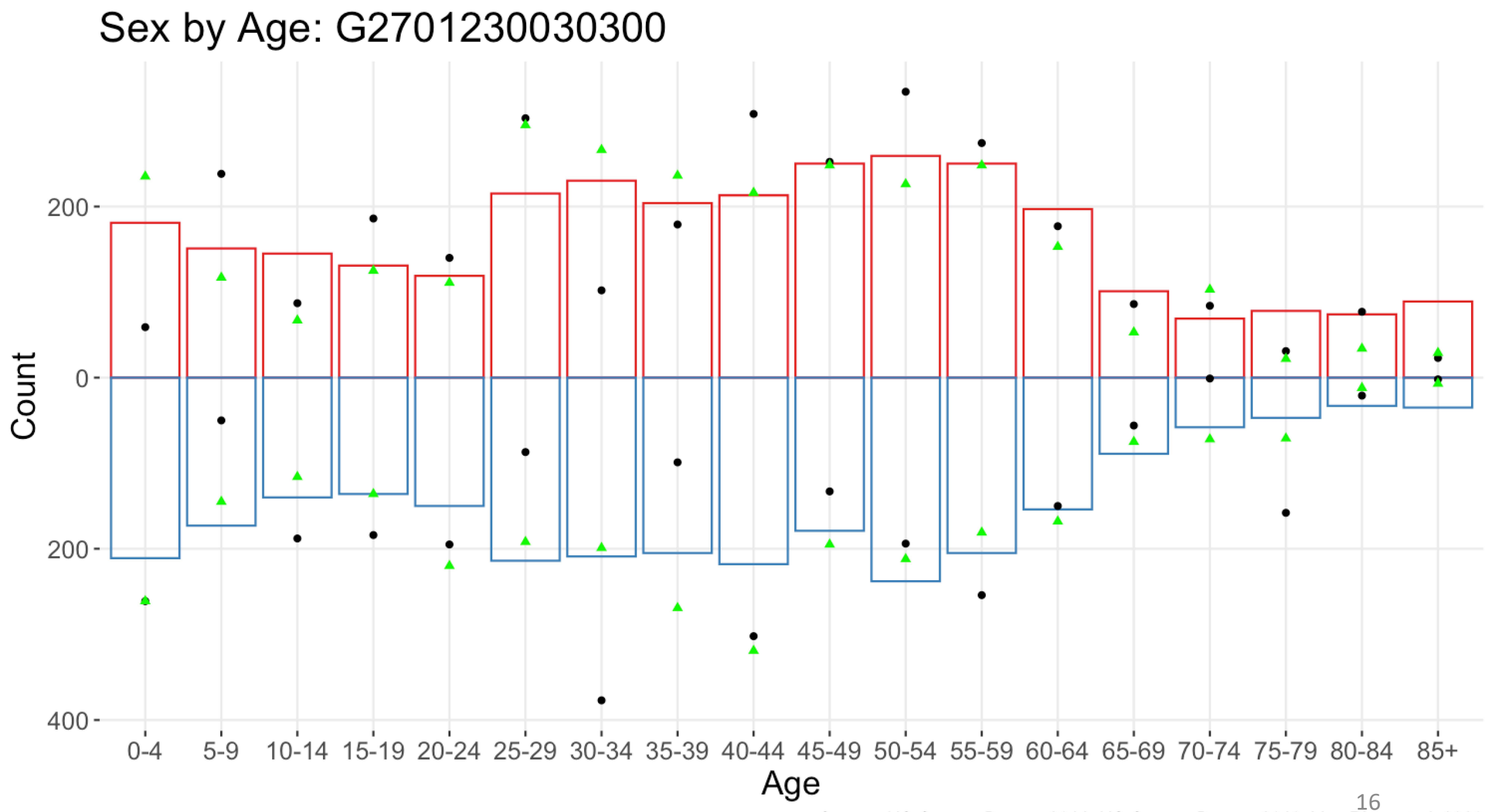

NATION
AIANNH Areas*
(American Indian, Alaska Native, Native Hawaiian Areas)
REGIONS
DIVISIONS
ZIP Code Tabulation Areas
Urban Areas
Core Based Statistical Areas
STATES
School Districts
Congressional Districts
Urban Growth Areas
Counties
State Legislative Districts
Public Use Microdata Areas
Voting Districts
Traffic Analysis Zones
County Subdivisions
Tract Groups
Places
Census Tracts
Subminor Civil Divisions
Block Groups
Census Blocks

Query	Allocation (%)
Voting age * Hispanic * Race * Citizen	50
Household – Group quarters	20
Detailed	10
Sex * Age (single year of age)	5
Sex * Age (4-year age bins)	5
Sex * Age (16-year age bins)	5
Sex * Age (64-year age bins)	5

RESULTS



Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020

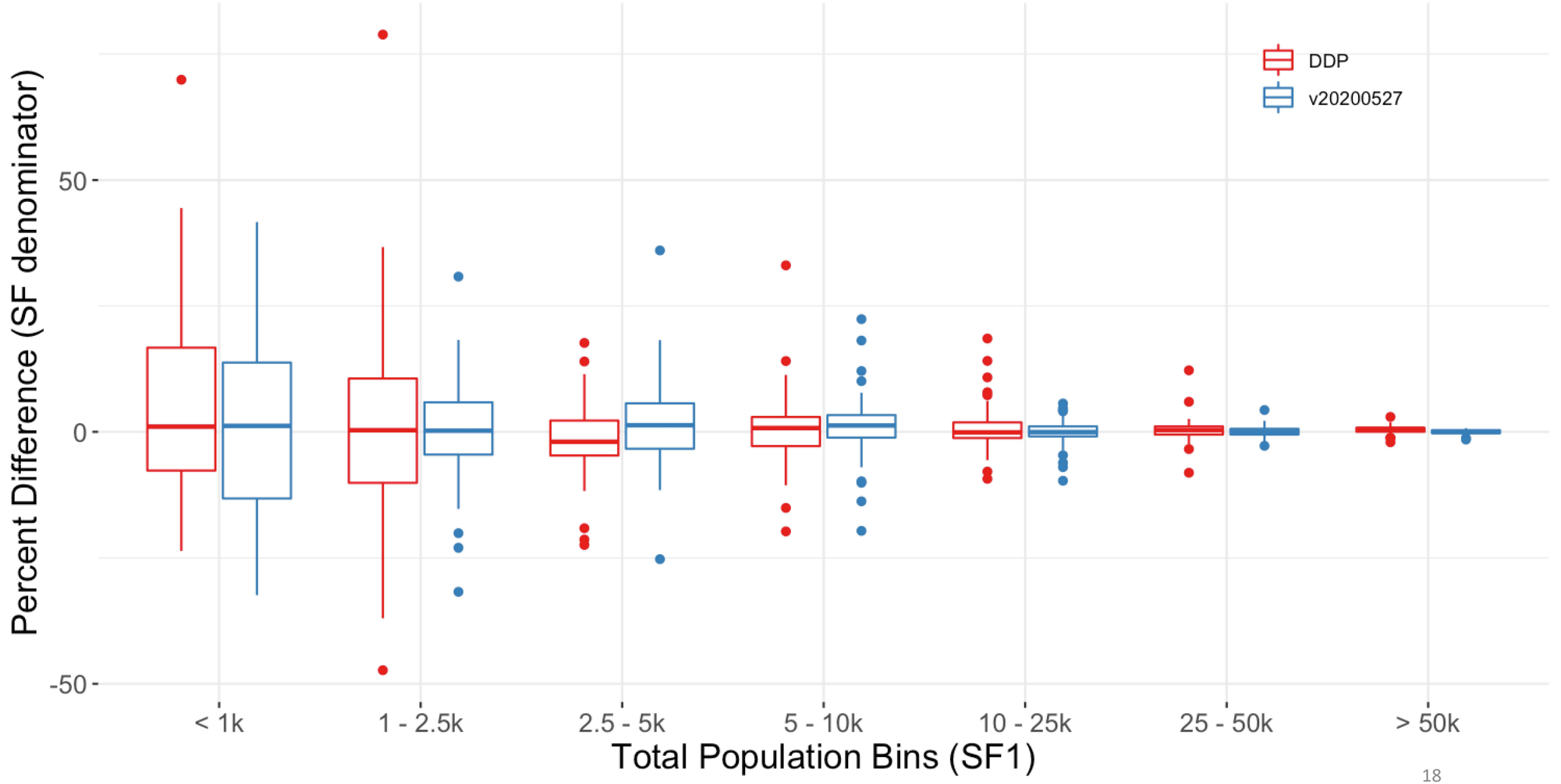


Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020

REAL-WORLD PUBLIC HEALTH EXAMPLE

- Asthma ED visit rates
  - Asthma ED visits in 2010 for Massachusetts towns
    - 0-4, 5-14, 15-34, 35-64, 65+ age bins
  - Age counts (denominators)
    - 2010 Summary File 1
      - Gold standard
    - Vintage 1 (October 2019)
    - Vintage 2 (May 2020)
  - How do rates based on diff. private different denominators compare to rates based on 2010 SF1 data?

Percent Difference in Age-Adjusted Asthma ED Visits in 2010 (MA towns)



Source: Van Riper et al. 2020; US Census Bureau 2019; Massachusetts Department of Health 2020

2020 DECENNIAL STILL A MOVING TARGET

- Additional demonstration data release scheduled for April 30
  - Only data on race, ethnicity, and voting age
- Demonstration data on sex, age, race/ethnicity forthcoming
  - No firm timeline, though, from Census Bureau
- Scientists should study guidance on handling uncertainty in decennial counts
  - Handbook on differential privacy will be available
  - Unlikely to get measure of uncertainty for decennial counts

CONCLUSION

- From ACS and Decennial one can expect good city-level rates/population estimates
  - But within-city or county is harder to understand
- Geographic and demographic resolution matter
  - Units with larger counts will be more accurate
  - Demographic groups with larger counts will be more accurate
- It is possible to process publicly released data to improve estimates.
  - Particularly for ACS data
  - Less certain about decennial data

QUESTIONS OR FEEDBACK:

vanriper@umn.edu

seth.spielman@colorado.edu

Code and data:

https://github.com/geoss/cdc-ncvhs-covid-2021