Preserving Privacy in Person-Level Data for the American Community Survey

Rolando A. Rodríguez, U.S. Census Bureau
Michael H. Freiman, U.S. Census Bureau
Jerome P. Reiter, Duke University and U.S. Census Bureau
Amy Lauger, U.S. Census Bureau

Joint Statistical Meetings
August 1, 2018

The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

A roadmap for our research

1. Data protection goals for Census Bureau data
    - Protect data against all possible attacks
    - Quantify disclosure risk and data quality
2. Current methods do not fully achieve the goals
3. Differential privacy can achieve these goals
    - Assumed currently infeasible for American Community Survey (ACS)
4. Current synthesis research is intermediate step toward goals

The Census Bureau has multiple goals in protecting data

In the current data environment, we must protect data against known attacks and attacks not yet known to us

We need to quantify the risk and data quality that our disclosure methods allow

Our methods should be transparent, so that users can account for the effect of disclosure on their inferences

Traditional disclosure methods will not be as effective in the future

Currently, ACS protected with a variety of ad hoc approaches, with some parameters kept secret

Current methods cannot be mathematically demonstrated to be safe

Better database reconstruction algorithms and increased computing power will increase risk

“Big data” means the Census Bureau can no longer assume it knows all of the data an attacker could use

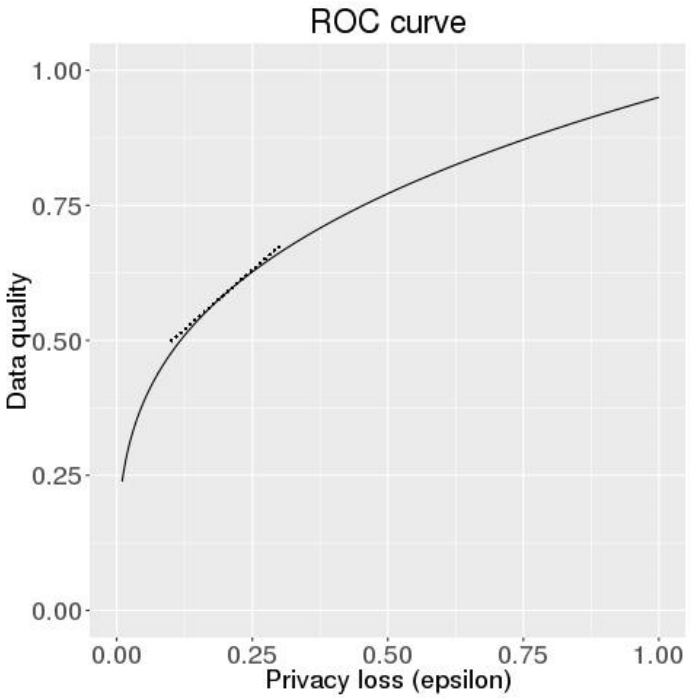
Formal privacy provides the guarantees traditional methods lack

A formal privacy framework, e.g., differential privacy, defines and quantifies the privacy loss from data releases

Algorithms used to protect the data must be proven to limit the privacy loss to no more than a certain “budget”

The Census Bureau is researching using formal privacy for more data products

The ROC curve shows the tradeoff in setting the privacy budget



The choice of privacy budget is a tradeoff between data usefulness and privacy loss

More privacy requires more perturbation

More utility requires less privacy

The appropriate point on the curve is a subjective decision

Making the ACS formally private is particularly challenging

The ACS collects data from 2.3 million housing units per year and publishes 12 billion estimates annually, including 5-year estimates

The ACS uses a complex sample weighting approach

Some challenges are shared with the Census

- Statistics are produced for small geographic areas
- Some within-household relationships are important and should be reflected in the protected data

We aim to make the ACS formally private in the future, but that is not our current focus

Model-based synthetic data will improve on current methods but will not be formally private

We generate variables sequentially, not yet incorporating weights

Each variable in the synthesis is created using the previously synthesized variables for that record

We synthesize categorical variables with a classification tree

- Grow a tree on the previous variables
- Draw a value at random from the appropriate leaf

Current research analyzes the results of a synthesis

Synthesis method:

- We start with original values of sex, age, relationship to householder
- We synthesize school enrollment, grade level attending, educational attainment

Data source: 2014 ACS Public Use Microdata Sample

unweighted data from Oregon

We create 896 synthetic implicates and 4,480 bootstrap draws from the data

Education variables have properties that are useful to study

School enrollment, grade level attending and educational attainment are useful to study because:

- They have very close relationships with each other, particularly for those enrolled in school
- They are approximately bounded depending on age
- Values are often aggregated up to higher levels that are of more interest than the more detailed data

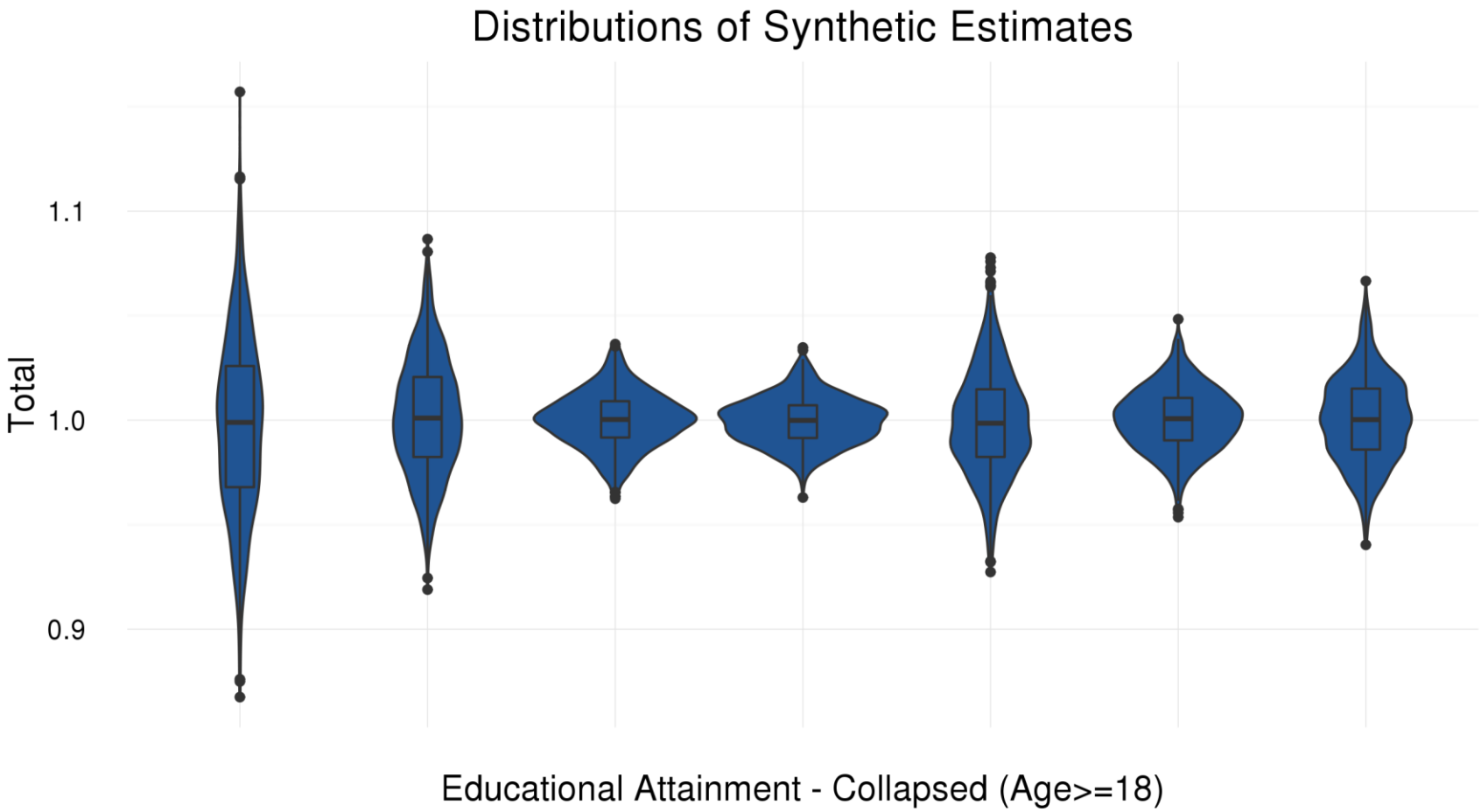
Tabular estimates look right on average

For various tables, we made violin plots of the ratio of the value of each cell of the table based on the synthetic data and the original data

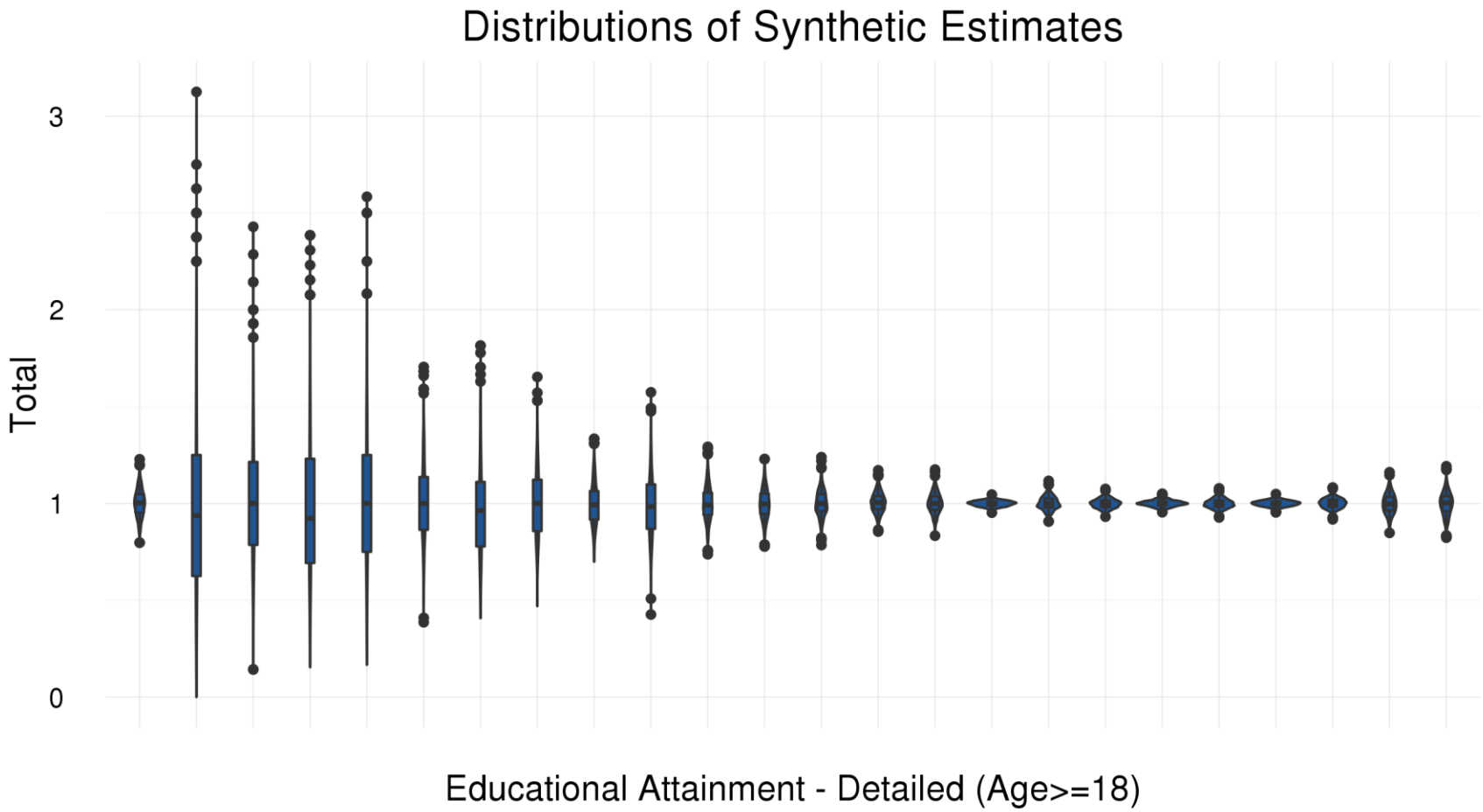
Generally, average number of records in a cell across synthetic implicates was approximately correct

Relative variance was larger for sparse cells

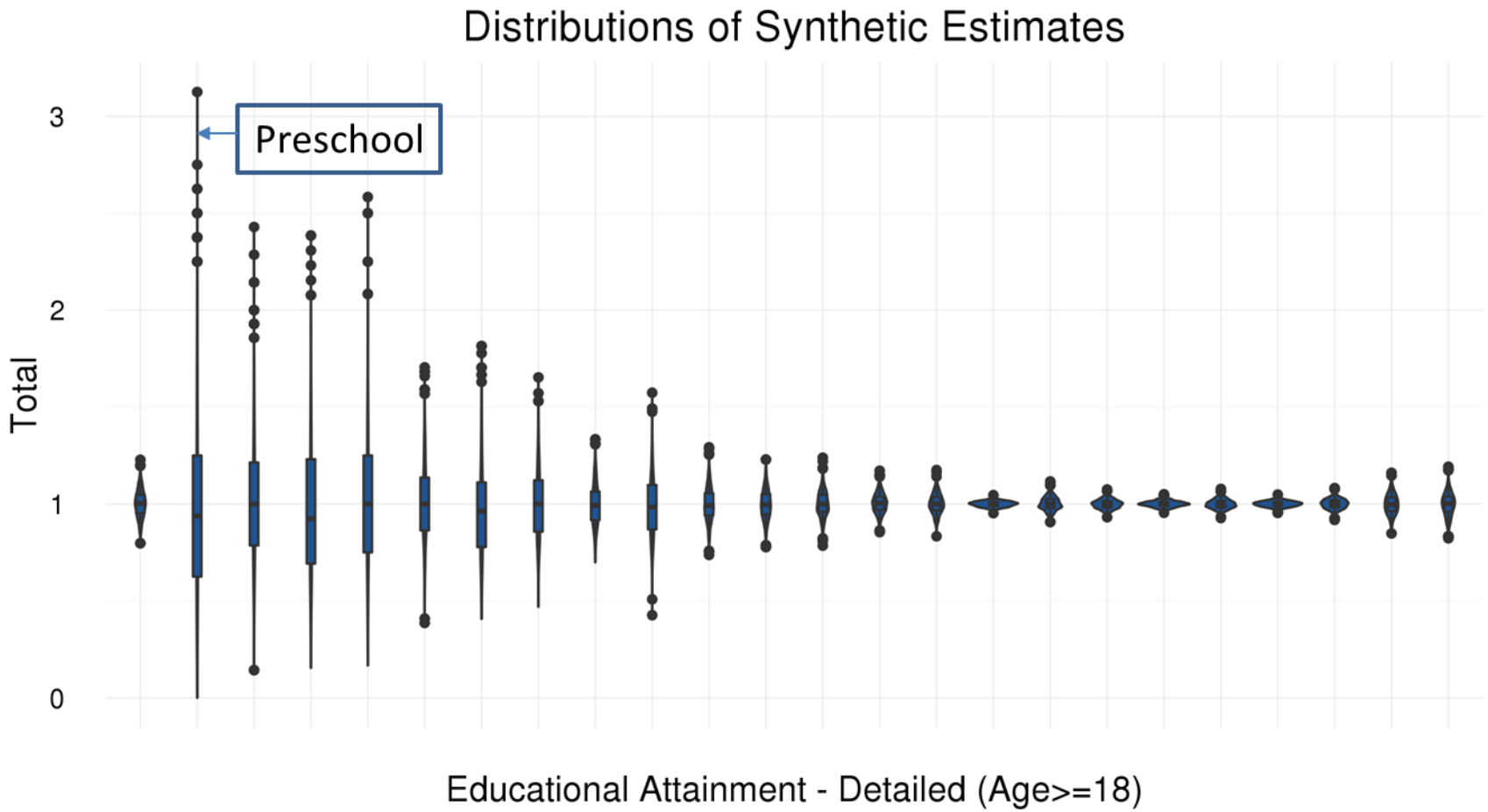
Violin plots of relative synthetic versus original estimates for cells in table of educational attainment – collapsed categories



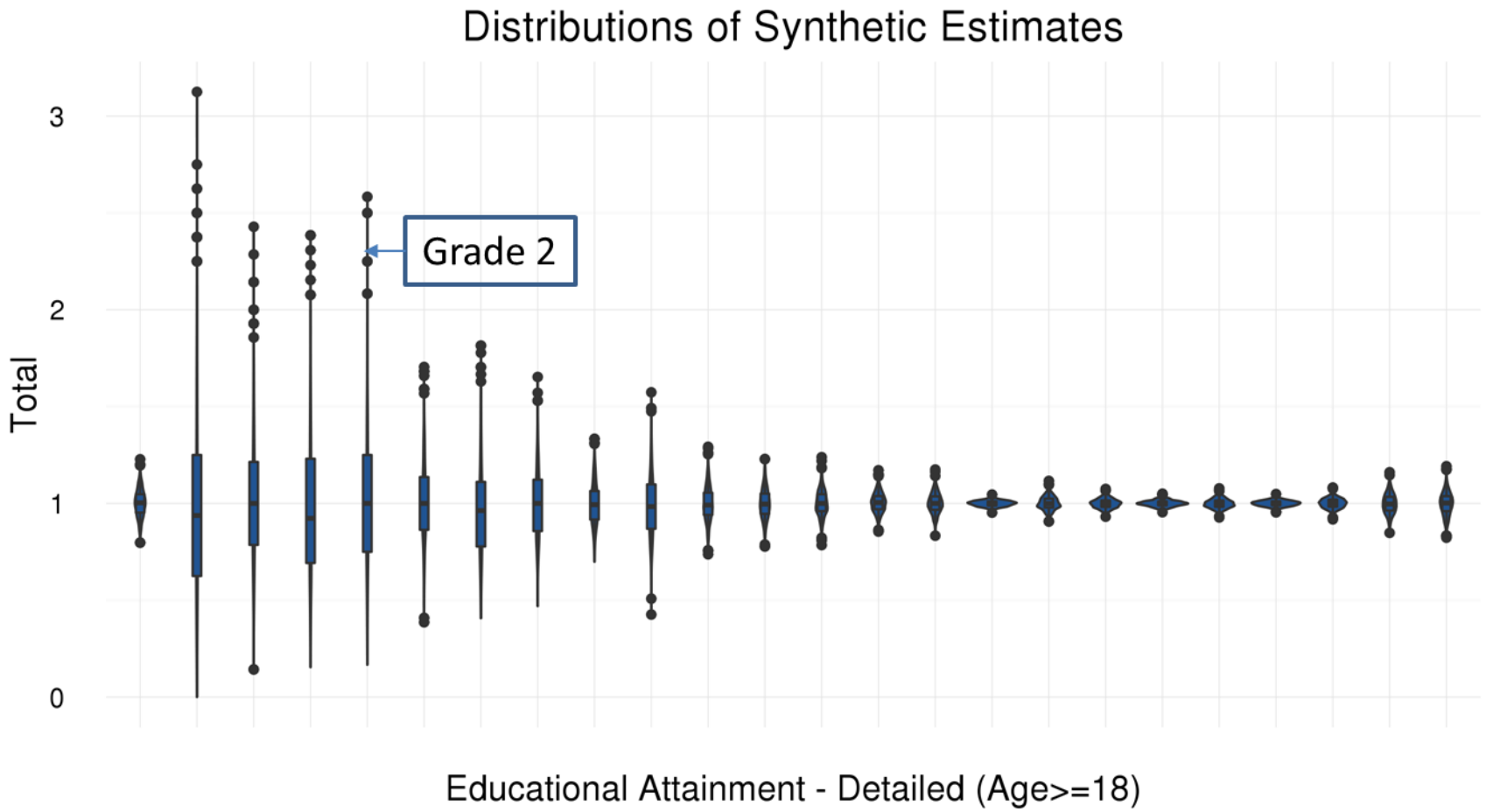
Violin plots of relative synthetic versus original estimates for cells in table of educational attainment – detailed categories



Violin plots of relative synthetic versus original estimates for cells in table of educational attainment – detailed categories



Violin plots of relative synthetic versus original estimates for cells in table of educational attainment – detailed categories

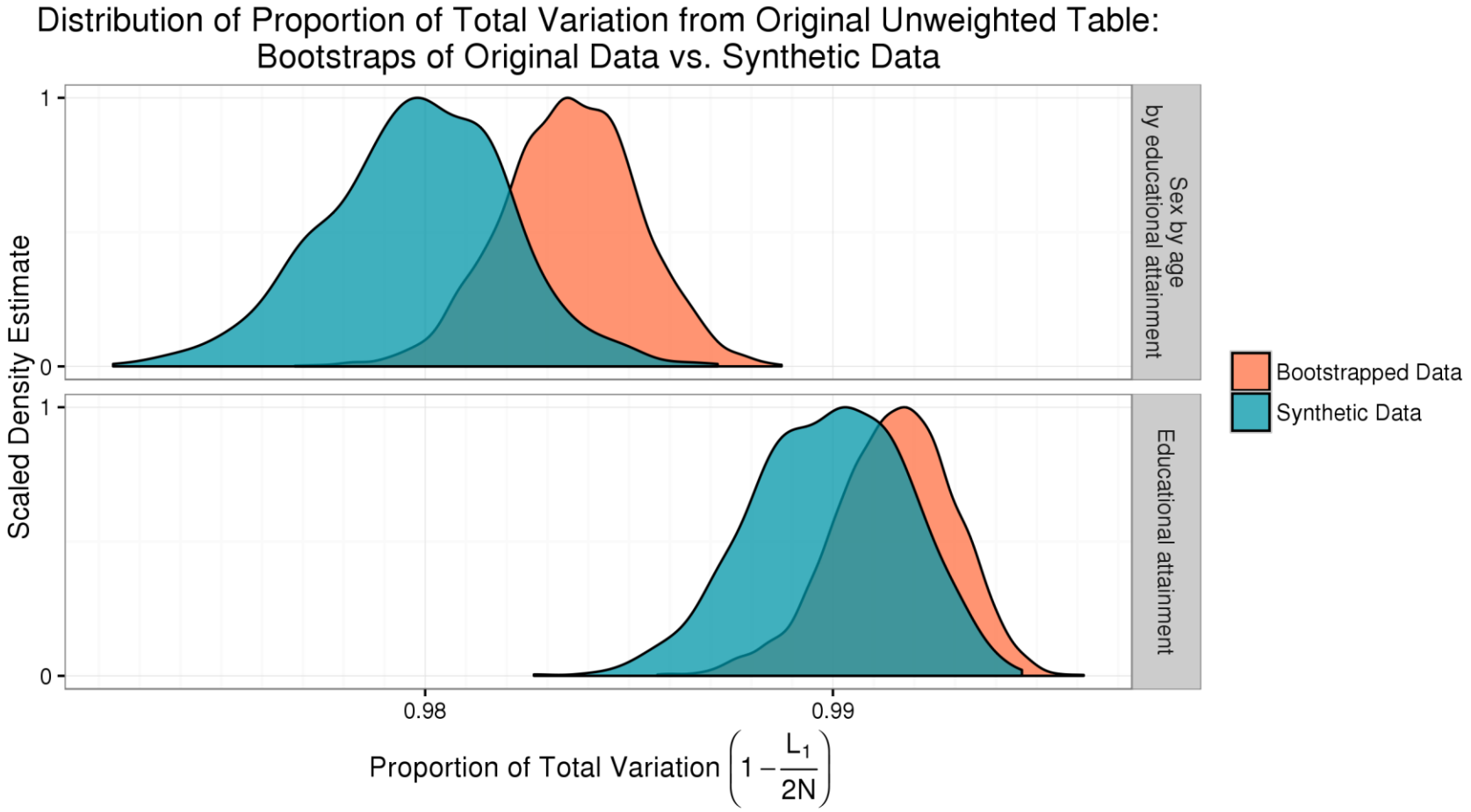


Quality of data

For each of several tables, we compute a measure of data quality: $1 - \frac{L_1}{2N}$, where L_1 is the L1 distance between the original table and the synthetic table and *N* is the sample size

We compare the distribution of the metric for the synthetic implicates to the distribution for the bootstrap draws

L1 graphs show some synthetic data have quality comparable to the original data



We seek to improve how the model captures correlations between variables

We have made progress in synthesizing variables, but we still want to improve our capturing of correlations between variables

The tree method is designed to preserve the strongest relationships in the data, including non-linear relationships among more than two variables

Some relationships of smaller magnitude may be important to preserve because of ways the data are used

The path forward presents unresolved challenges

Test data synthesized so far reflect only some of the original data's properties

We need to incorporate weights into the final synthetic data

We need to be able to synthesize data at levels lower than the state

We need more metrics and benchmarks to assess suitability of various models

We need to research the feasibility of formal privacy for this dataset

Michael Freiman
michael.freiman@census.gov