Creating a Synthetic Public Use File and Validation Server

Progress Report

September 21, 2018

Len Burman, Surachai Khitatrakun, James R. Nunns, Philip Stallworth, Kyle Ueyama

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Goals

- Produce synthetic data file with the same record layout as IRS Administrative Data that:
  - Protects the confidentiality of tax return information
  - May be used for statistically valid analysis for certain research purposes
  - May be used as a “training data set” to develop programs to run on confidential data
- Develop a safe procedure for selected researchers to remotely submit programs to perform statistical analysis on administrative data when the synthetic files are inadequate

Propose to create a fully synthetic dataset

- In a fully synthetic dataset, all of the data are synthesized
  - If there are *k* variables, $Y_1, \ldots, Y_k$, create synthetic $\widehat{Y_1}$ drawn from empirical distribution of Y_1; $\widehat{Y_2}$ conditional on $\widehat{Y_1}$ and empirical distribution of ε_2; and so on until $\widehat{Y_k}$ is synthesized based on $\widehat{Y_1}, \ldots, \widehat{Y_{k-1}}$
- Minimizes disclosure risk since all data are synthesized
- Concerns about data quality
  - Very little experience with fully synthetic datasets
  - Never been tried on a dataset this large

Initial step: create a synthetic nonfiler database

- High-quality nonfiling data are necessary for policy simulations
- Working plan is to synthesize individuals, then create tax units
- Construct a set of individual nonfilers using information returns and SSA information for people who do not file a tax return
- Challenge #1: Identify the relevant set of nonfiling individuals
  - Exclude foreigners and residents of U.S. Possessions, people who died before filing year
  - Create records for people without information returns
- Challenge #2: Creating nonfiling tax units from individual records
  - Marry some individuals and assign children to form tax units based on information from survey data

Some issues

- What to do about people who apparently should have filed a tax return?
  - How to determine filing requirements given data limitations (e.g. single vs. married filing jointly thresholds, significantly lower threshold for self-employed and small businesses)?
  - RAS studies have found that they account for a significant share of nonfilers' tax liability (both income and payroll tax)
  - Some might file late, some may be due to erroneous identification of "nonfilers" (e.g., incorrect SSN matching), and some might be identified in IRS enforcement activities
- Studies of census data matched to IRS data show them imperfect for identifying tax liability and filing status. Can we improve?

Data Cleaning

- Conducting a synthesis using the full RTF is preferable to one based upon the INSOLE from a privacy perspective.
- However, the RTF records would need to be cleaned as they are on the INSOLE in order to conduct a proper synthesis.
- Using the INSOLE and their underlying RTF records as a training set, we are beginning to develop a machine learning algorithm to infer the corrections made on the INSOLE file by human editors and apply those corrections to the entire RTF.

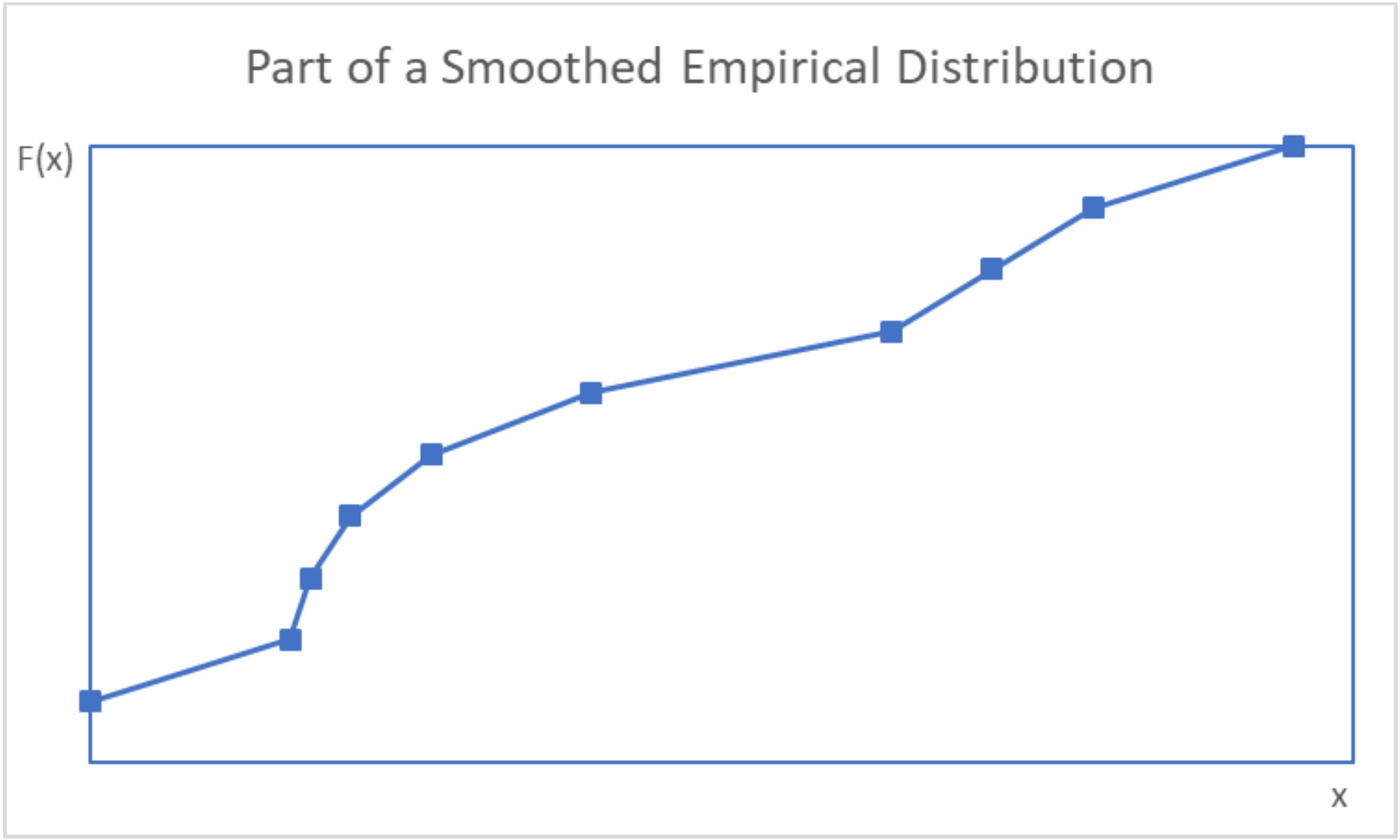
Creating the synthetic PUF

- Start with a cleaned version of the RTF (or an extract)
- Parametric estimation for some variables and nonparametric for others
- For example, CART for discrete variables such as number of children
- Regression for continuous variables (Tobit type estimators for censored variables, such as interest income)
  - Use a polynomial expansion on right-hand side variables to capture some nonlinear relationships between variables

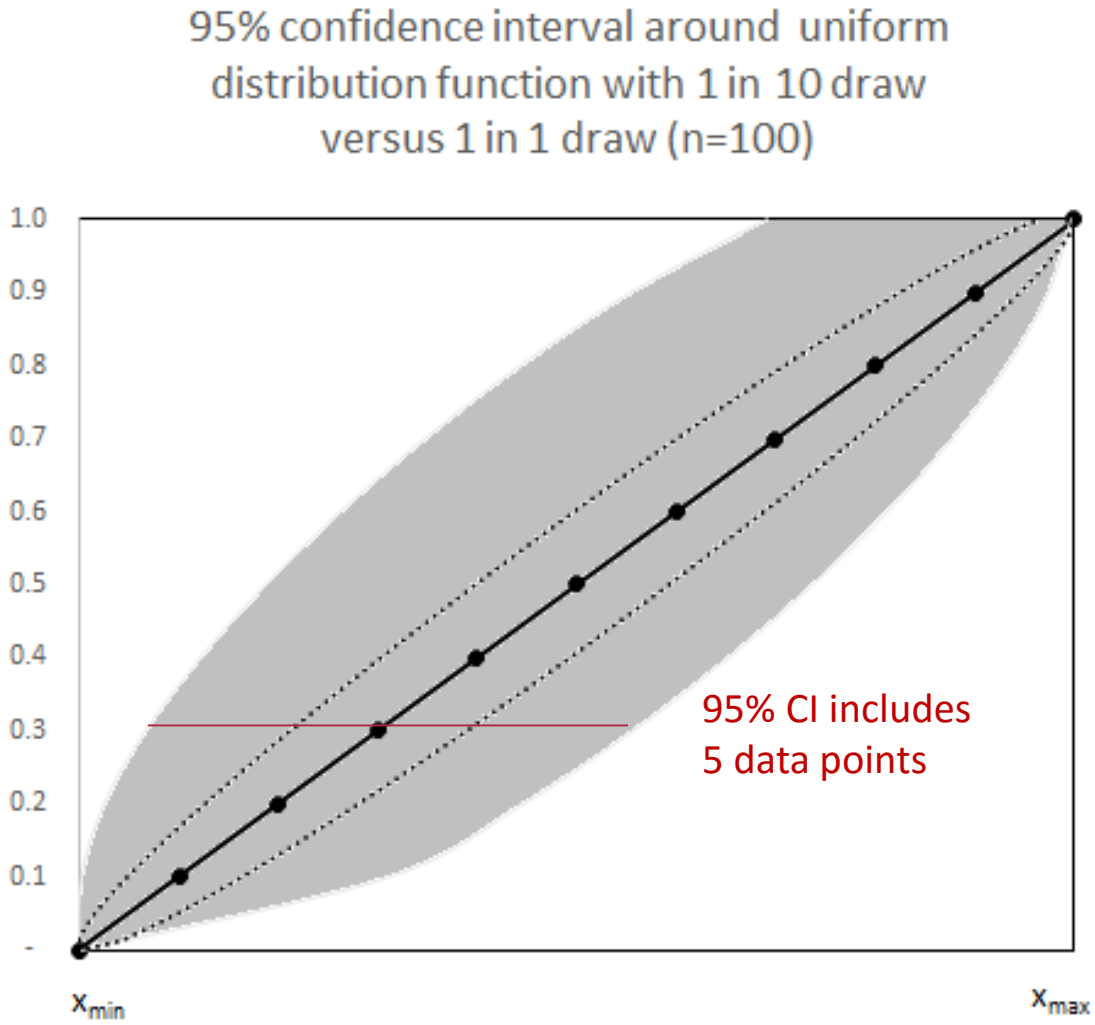
Protecting privacy

- Because data are fully synthetic, there is no risk of disclosing actual values from tax returns
- However, process may disclose information about the distribution of variables
- Drawing from smoothed version of distribution, sampling, and regression-based method provide substantial protection from disclosure for most of sample
- Special treatment of the tails of distribution (outliers)
- Top coding discrete variables (e.g., number of children)
- Some variables might have to be suppressed or combined

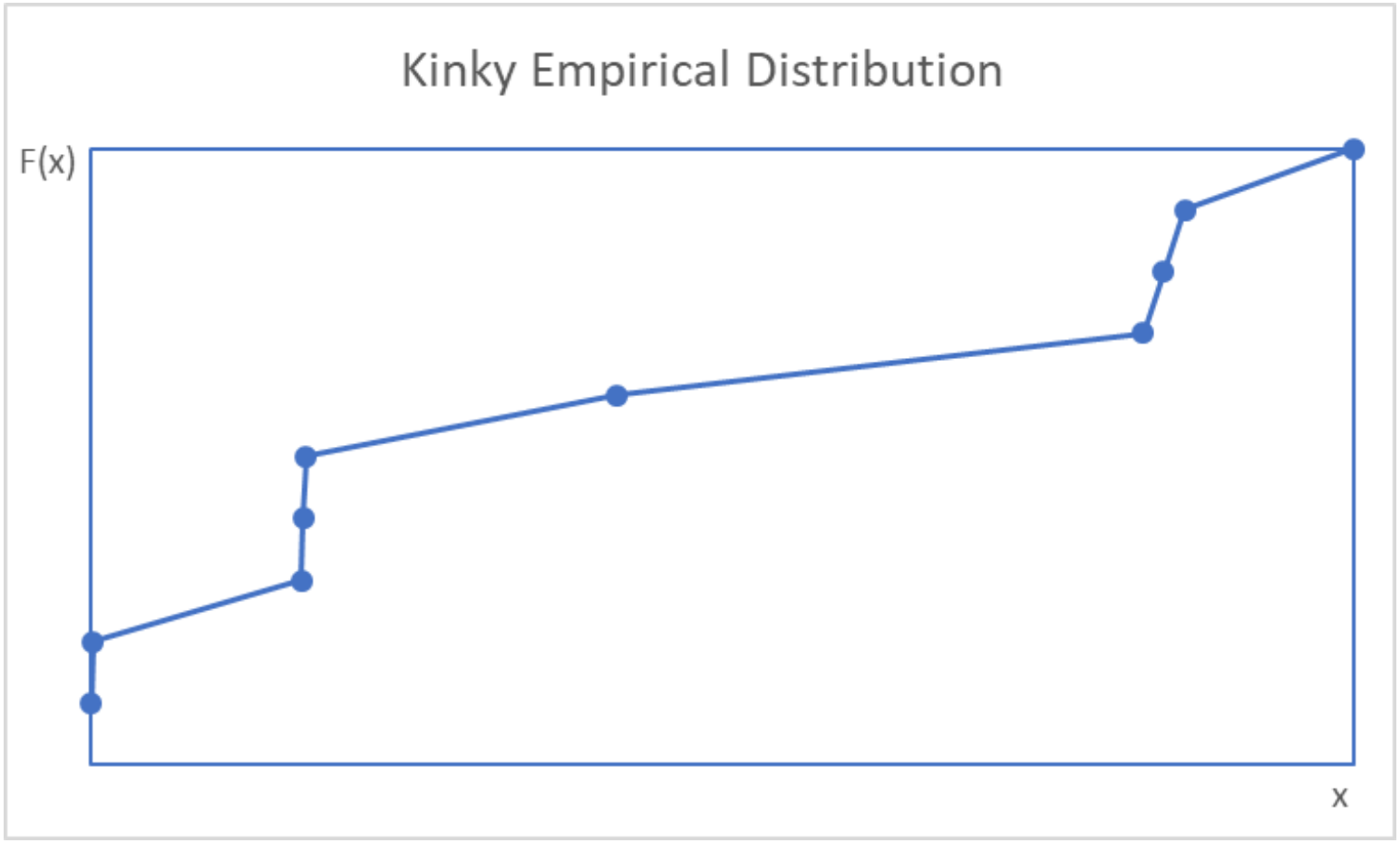
Smoothing the empirical distribution



How smoothing the empirical distribution and subsampling preserves privacy



Special concern with kinky empirical distribution (multiple observations with same values)



- Nonzero chance of selecting multiple identical values
- Address by rounding or slightly blurring

Dealing with outliers

- Procedure above could disclose information about extreme values (since no synthesized value could be greater than maximum)
- Further smooth the empirical distribution in tails
  - For example, draw from a distribution with mean μ and variance σ^2 for largest 100 values
  - Maximum of this distribution can be greater than the population maximum
  - Intruder could only infer information about mean and variance (2 parameters) in tail rather than the 100 potentially identifiable observations
  - As the tail distribution becomes more skewed, the procedure becomes *less* revealing about the distribution of individual observations

Remote Access of Administrative Data

- Researchers would develop their programs using the synthetic dataset and submit the programs electronically to the IRS
  - Output subjected to disclosure review before release to researcher
- Procedure would be similar to access to the confidential version of the SIPP
- Disclosure risk could be reduced by basing estimates on random subsamples of the restricted dataset (i.e., by drawing random samples with replacement from the full dataset)
  - Programs could include generated random seed so that a particular analysis could be replicated, but any new analysis would start with a different seed
- Costs defrayed by fees