An Integrated System for Confidential Data Access

Jerry Reiter
Department of Statistical Science
Information Initiative at Duke
Duke University
jerry@stat.duke.edu

Acknowledgements

- Research supported by
  - National Science Foundation
    - ACI 14-43014, SES-11-31897, CNS-10-12141
  - National Institutes of Health: R21-AG032458
  - Alfred P. Sloan Foundation: G-2015-20166003
  - US Bureau of the Census

- Any views expressed are those of the author and not necessarily of NSF, NIH, the Sloan Foundation, or the Census Bureau

The vision we are working towards

- Integrated system for access to confidential data including
  - unrestricted access to **fully synthetic data**, coupled with
  - means for approved researchers to access confidential data via **remote access** solutions, glued together by
  - **verification servers** that allow users to assess quality of inferences from the synthetic data.

Synergies of integrated system

- Use synthetic data to develop code, explore data, determine right questions to ask
- User saves time and resources when synthetic data good enough for her purpose
- If not, user can apply for special access to data
- This user has not wasted time
  - Exploration with synthetic data results in more efficient use of the real data
  - Explorations done offline free resources (cycles and staff) for final analyses

Synthetic data: Where are we now?

- Available data products (released by Census Bureau)
  - Synthetic Longitudinal Business Database,
  - Synthetic Survey of Income and Program Participation
  - OnTheMap
- Off-the-shelf software to generate synthetic data? Not yet.
- General plug-and-play routines?
  - *Model based synthesis* – yes, but hard to characterize disclosure risks beyond re-identification
  - *Formally private synthesis* – much theoretical development, but not much practical experience for complex datasets

Verification servers: Where are we now?

- Allowable verifications depend on user characteristics
- We have developed verification measures that satisfy **differential privacy**
  - Plots of residuals versus predicted values for regression
  - ROC curves in logistic regression
  - Statistical significance of regression coefficients
  - Tests that coefficients exceed user-defined thresholds
- R software package in development
- Open question: how to scale up while respecting privacy budgets

Illustrative application: The OPM Synthetic Data Project

- Created fully synthetic version of the OPM CPDF-EHRI status file
  - Longitudinal work histories of civil servants from 1988 to 2011
  - Simulate careers, demographics, grades and steps, salaries, ….
  - Only available to OPM and Duke IRB approved researchers at the moment

Illustrative application: Verification of regression

- Regress log salary on demographics, including gender and race
- Hypothetical results from the synthetic data (dummy numbers as we are vetting final analyses):
  - Median salaries for Asian men are about 1.5% lower than median salaries for white men, holding all else constant
  - Huge sample sizes, so statistically significant
- Is the result from the synthetic data believable?

Illustrative application: Verification of regression

- User defines a threshold that represents a result of practical significance
  - Test if true coefficient for Asian male $B < -.01$
- Verification software returns differentially private answer that reflects uncertainty due to noise
  - Goal: estimate the probability, $p = \Pr(B < -.01)$
  - Output: 95% credible interval for p
  - Examples:
    - interval for p is (.92, 1.0), conclude synthetic data result valid
    - interval for p is (.52, .64), don't trust synthetic data result