An Integrated Approach to Providing Access to Confidential Data

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

Acknowledgements

- Research supported by
  - National Science Foundation
    - ACI 14-43014, SES-11-31897, CNS-10-12141
  - Alfred P. Sloan Foundation: G-2015-20166003

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

A vision we are working towards

- Integrated system for access to confidential data including
  - unrestricted access to **fully synthetic data** (ideally satisfying some formal privacy criterion), coupled with
  - means for approved researchers to access confidential data via **remote access** (or FSRDC), 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

Verification Servers

- Verification servers (Reiter et al. 2009, *Computational Statistics and Data Analysis*)
  - Separate system with confidential and synthetic data
  - User submits query to system for verification of particular analysis
  - Server reports back measure of similarity of analysis on confidential and synthetic data
- User can decide to publish if quality sufficient
- But quality measures can leak information

How to provide verifications

- 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
  - Measures of accuracy of prediction models

- R software package in development

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(basic pay) per employee-year on demographics including race. Modeled by gender.

- Some interesting results from the synthetic data for the analysis pooling all years
  - Median pay for Asian men about 2.8% lower than median pay for White men, holding all else constant
  - Black women have higher median pay than White women but not statistically significant

- Are the results 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 $B < -.01$
  - Or test if true B within tolerance of synthetic estimate
- Verification algorithm returns differentially private answer that reflects uncertainty due to noise
  - Goal: estimate the probability, $r = \Pr(B < -.01)$
  - Output: 95% credible interval or posterior mode for r
  - Examples:
    - interval for r is (.92, 1.0), conclude synthetic data result valid,
    - interval for r is (.00, .20), don't trust synthetic data result.

Verification measure

- Partition the confidential data into M disjoint subsets
- Compute coefficient of interest in each partition
- Count number of times, S, coefficient satisfies threshold
- Add noise to S drawn from Laplace distribution, where global sensitivity equals one, to get T
- Use Bayesian model to estimate posterior distribution of r given T
- Report posterior mode of r to user

Verification of males' regression (ε = 1, threshold = -.01 for each *B*)

Variable	Synthetic	Mode of p	Confidential
AI/AN	-.006 (4)	**.76**	-.019 (12)
Asian	-.028 (30)	**.99**	-.040 (43)
Black	-.021 (39)	**.99**	-.036 (61)
Hispanic	-.014 (22)	**.99**	-.029 (42)
Age	.033 (365)		.043 (480)
Age Sq.	-.00027 (269)		-.00036 (352)
Education	.013 (122)		.021 (180)

Verification of females' regression (ε = 1, threshold = -.01 for each *B*)

Variable	Synthetic	Mode of p	Confidential
AI/AN	-.009 (7)	**.97**	-.027 (19)
Asian	-.011 (13)	**.42**	-.010 (11)
Black	.00013 (.3)	**.003**	-.003 (8)
Hispanic	-.013 (19)	**.99**	-.021 (30)
Age	.023 (286)		.032 (404)
Age Sq.	-.00019 (205)		-.00027 (295)
Education	.014 (130)		.023 (198)

Illustrative application: Summary of verification results

- Males: synthetic data suggest all coefficients except for AI/AN less than -.01. Verification confirms!
- Females:
  - Synthetic data suggest coefficients for AI/AN near -.01 and Hispanic less than -.01. Verification confirms!
  - Synthetic data suggest coefficient for Black not less than -.01. Verification confirms!
  - Synthetic data suggest coefficient for Asian close to -.01. Synthetic data agree with .42.

When does verification measure work well?

- Works well for coefficients based on large sample sizes
    - Does not account for uncertainty appropriately otherwise
        - Methods for statistical significance can handle this
    - Undefined when regressions fail to fit in some partitions, e.g., because of sample size
        - Add third category, number of errors, and make differentially private version of count in that category as well.

Future directions

- Finding ways to get high quality results without high privacy budgets.
  - Considering multivariate quantities
  - Global privacy budget versus individual privacy budget versus analysis privacy budget – interactions with trust
- Accounting for uncertainty when estimates based on complex, modest-sized samples
- Developing software to implement the idea
- Paper on arXiv…. (search "Barrientos, Reiter")