Private Tabular Survey Data Products through Synthetic Microdata Generation

Terrance D. Savitsky [1] Matthew R. Williams [2]
Monika (Jingchen) Hu [3]

[1] U.S. Bureau of Labor Statistics (Office of Survey Methods Research)

[2] NSF (National Center for Science and Engineering Statistics)

[3] Vassar College (Mathematics and Statistics Department)

FCSM Conference
November, 2021

Overview

- Goal: Generate private data for tabular release
  - e.g. Tables of counts and salaries
  - point estimates and SE estimates
- Approach: Synthesize data with privacy guarantee
  - Model both outcome y and survey weights w
  - Two ways:
  - 1. Model under observed sample distribution
  - 2. Model under population distribution
- Results:
  - Compare synthesizers with additive noise mechanism

Outline

Differential privacy

- $D \in \mathbb{R}^{n \times k}$ be a database in input space $\mathcal{D}$
- Mechanism $\mathcal{M}() : \mathbb{R}^{n \times k} \rightarrow O$.
- $\mathcal{M}$ is ϵ-differentially private if

$$\frac{Pr[\mathcal{M}(D) \in O]}{Pr[\mathcal{M}(D') \in O]} \leq \exp(\epsilon),$$

- Probability $\mathcal{M}(D')$ assigns to O changes by max of $\exp(\epsilon)$ after deleting 1 row
- For all $D, D^{'} \in \mathcal{D}$ that differ by 1 row.

$\mathcal{M}$ = Additive Noise

- An output statistic $f(D)$; e.g., total employment
- Global sensitivity
  $\Delta_G = \sup_{D,D' \in \mathcal{D}:\ \delta(D,D')=1} \mid f(D) - f(D') \mid$
- Laplace Mechanism for additive noise, scaled to be proportional to Δ_G/ϵ with $\epsilon-$DP guarantee
- Survey weights dramatically increase Δ_G.
- Adding noise disrupts tabular constraints

$\mathcal{M}$ = Pseudo posterior distribution

$$\xi^{\boldsymbol{\alpha}(\boldsymbol{y})}(\theta \mid \boldsymbol{y}) \propto \prod_{i=1}^{n} p(y_i \mid \theta)^{\alpha_i} \times \xi(\theta)$$

- Down weight each likelihood by $\alpha_i \in [0, 1]$
- α_i lower when disclosure risk higher
- Sensitivity of $\xi^{\boldsymbol{\alpha}(\boldsymbol{y})}(\theta \mid \boldsymbol{y})$ based on $f_\theta(\boldsymbol{y}) = \log \prod_{i=1}^{n} p(y_i \mid \theta)^{\alpha_i}$.
- $\Delta_{\boldsymbol{\alpha}} = \sup_{\boldsymbol{y},\boldsymbol{y}' \in \mathcal{Y}^n : \delta(\boldsymbol{y},\boldsymbol{y}')=1} |\alpha(\boldsymbol{y}) \times f_\theta(\boldsymbol{y}) - \alpha(\boldsymbol{y}') \times f_\theta(\boldsymbol{y}')|$
- Each posterior draw with $\epsilon_{\boldsymbol{y}} = 2\Delta_{\boldsymbol{\alpha}}$ produces one synthetic $\boldsymbol{y}^*$
- Synthetic $\boldsymbol{y}^*$ produces survey tables with same privacy guarantee

Data from Survey sampling procedure

- A sample S of n individuals taken from population U of size N
- Each individual in U assigned selection probability $P(\omega_i = 1 \mid \mathcal{A}) = \pi_i$
- Estimate area statistics with survey weights $w_i = 1/\pi_i$, to reduce bias
- Survey w designed to correct bias
- Incorporating privacy leads to distortion (bias)

Outline

Two Synthesizing Models

- Synthesis of a local survey database $(\boldsymbol{y}_n, \boldsymbol{w}_n | \boldsymbol{X}_n, \boldsymbol{\alpha}_n)$:
- A Fully Bayes model for observed sample (FBS) models $(\boldsymbol{y}_n, \boldsymbol{w}_n | \boldsymbol{X}_n, \boldsymbol{\alpha}_n)$ under a multinormal pseudo likelihood.
- A Fully Bayes model for the population (FBP) that forms the exact likelihood for $(\boldsymbol{y}_n | \boldsymbol{X}_n, \boldsymbol{\alpha}_n), (\boldsymbol{w}_n | \boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{\alpha}_n)$ in the observed sample.

$$(y_i, w_i | x_i, \alpha_i, \omega_i = 1) = \frac{Pr(\omega_i = 1 | y_i, x_i, w_i) \times (y_i, w_i | x_i, \alpha_i)}{Pr(\omega_i = 1 | x_i, w_i)}$$

- FBS produces synthesized $\boldsymbol{y}_n^*$ without sampling bias, so discard weights to build tabular statistics
- FBS requires use of both $(\boldsymbol{y}_n^*, \boldsymbol{w}_n^*)$.

Estimation Algorithm

1. Estimate unweighted $\boldsymbol{\theta}$ with model,
   $\xi\left(\boldsymbol{\theta}|\mathbf{y},\mathbf{w}\right) \propto \left[\prod_{i=1}^{n} \pi\left(y_i, w_i|\boldsymbol{\theta}\right)\right] \times \pi\left(\boldsymbol{\theta}\right)$

2. Compute weights,
   $\alpha_i = m\left(\sup_{\theta\in\Theta} f_\theta\left(y_i, w_i\right)\right) \propto 1/\sup_{\theta\in\Theta} f_\theta\left(y_i, w_i\right)$

3. Re-estimate $\boldsymbol{\theta}$ using weights, α_i in
   $\xi^{\boldsymbol{\alpha}}\left(\theta \mid \mathbf{y},\mathbf{w}\gamma\right) \propto \left[\prod_{i=1}^{n} \pi\left(y_i, w_i \mid \theta\right)^{\alpha_i}\right] \pi\left(\theta \mid \gamma\right)$

4. Compute log-likelihood bound,
   $\sup_{y_i, w_i \in \mathcal{D}^n} \sup_{\theta\in\Theta} \left|\alpha(y_i, w_i) f_\theta(y_i, w_i)\right| \leq \Delta_\alpha$

5. Gives us privacy guarantee, $\epsilon \leq 2\Delta_\alpha$

6. Generate synthetic data, $\left(\mathbf{y}^*, \mathbf{w}^*\right) \sim \pi_{\boldsymbol{\alpha}}\left(\mathbf{y}^*, \mathbf{w}^*|\mathbf{y},\mathbf{w}\right)$

Outline

Computing Sensitivity

- The local sensitivity $\Delta^c_{f,g}$ for count of field f and gender g (cell count):

$$\Delta^c_{f,g} = \max_{i \in \mathcal{S}_{f,g}} w_i - \min_{i \in \mathcal{S}_{f,g}} w_i$$

- The local sensitivity $\Delta^a_{f,g}$ for average salary of field f and gender g (cell average):

$$\Delta^a_{f,g} = \frac{\max_{i \in \mathcal{S}_{f,g}} w_i y_i - \min_{i \in \mathcal{S}_{f,g}} w_i y_i}{\sum_{i \in \mathcal{S}_{f,g}} w_i - (\max_{i \in \mathcal{S}_{f,g}} w_i - \min_{i \in \mathcal{S}_{f,g}} w_i)}$$

- $\Delta^c_* = \max_{f,g} \Delta^c_{f,g}$ and $\Delta^a_* = \max_{f,g} \Delta^a_{f,g}$
- Generate noise $\mathrm{Laplace}(0, \Delta_{f,g*}^{c,a}/\epsilon)$ added to cell count and average salary

Outline

- Differential privacy
- Two synthesizers
- Laplace Mechanism
- Survey of Doctoral Recipients Application
- Simulation Study
- Concluding remarks

SDR application (10,355 obs): model fits

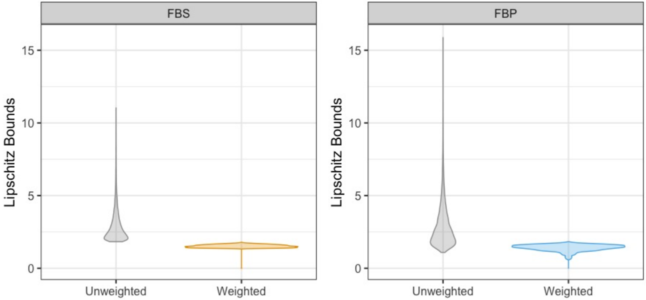

Figure: Distributions of record-level Lipschitz bounds of the non-private unweighted and the private weighted of FBS (left) and FBP (right) in the SDR application.

SDR application (10,355 obs): utility evaluation

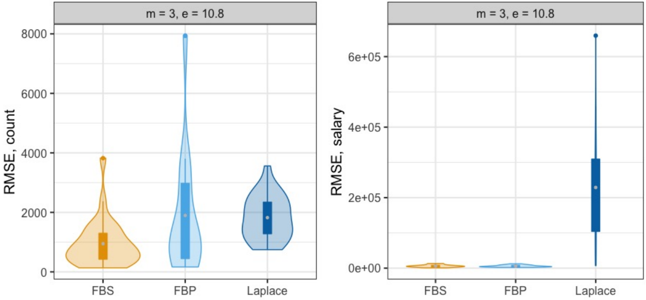

Figure: RMSE values of **counts** (left) and **average salary values** (right) of the three methods, FBS, FBP, and Laplace, applied to the SDR sample. Each violin plot represents a distribution of RMSE values over 27 cells. Results are based on $m = 3$ synthetic datasets by FBS and FBP, achieving $\epsilon_{\boldsymbol{y}_n} = 10.8$ for all three methods.

Outline

Differential privacy

Two synthesizers

Laplace Mechanism

Survey of Doctoral Recipients Application

Simulation Study

Concluding remarks

Simulation studies: simulation design

- Based on the 2017 SDR public use file
- Population $N = 100,000$ units of: salary (y_i), field of expertise and gender $(\boldsymbol{x}_i)$
- Salary $y_i \mid \boldsymbol{x}_i \sim \text{Lognormal}(\mu_i, 0.4)$ where μ_i is group-specific mean from the public use file
- Additive noise: $\text{noise}_i \sim \text{Lognormal}(0, 0.4)$
- Survey weights:

$$\begin{aligned} \log(\pi_i) &= \log(y_i) + \text{noise}_i \\ w_i &= 1/\pi_i \end{aligned}$$

- Take a stratified PPS sample of $n = 1000$ using fields as strata: $(\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{w}_n)$

Sensitivity before/after weighting by α

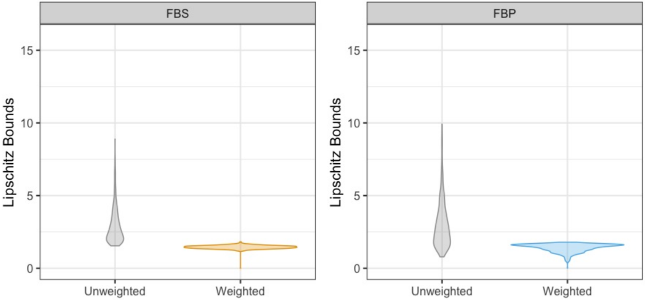

Figure: Distributions of record-level sensitivity bounds of the non-private unweighted and the private weighted of FBS (left) and FBP (right) in the simulation.

Smoothed weights $\mathbb{E}(w|y)$ Improves Efficiency

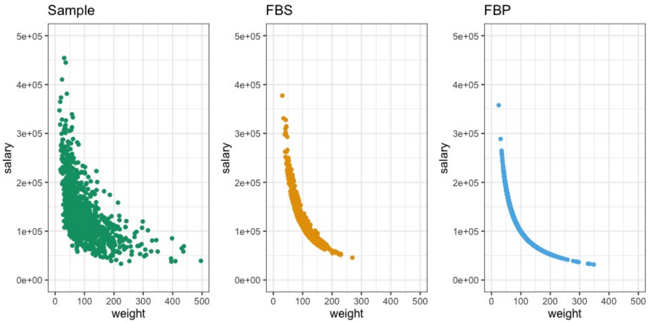

Figure: Comparison of the salary and weight bivariate distributions of confidential salary and weights in the sample (green and left), synthetic salary and smoothed weights from FBS (yellow and middle), and synthetic salary and smoothed weights from FBP (blue and right).

Utility of $\mathcal{M}$ = Synthesizers versus $\mathcal{M}$ = Laplace

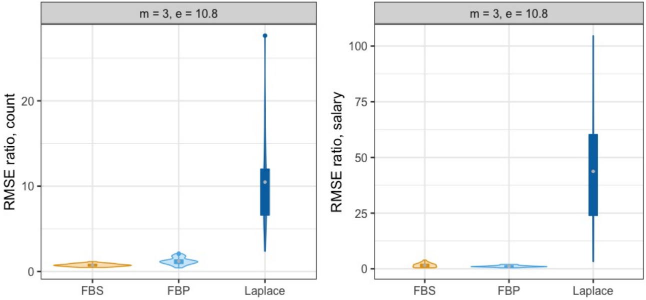

Figure: RMSE ratios of **counts** (left) and **average salary values** (right) of the three methods, FBS, FBP, and Laplace, applied to the selected sample. Each violin plot represents a distribution of RMSE ratios over 27 cells. Results are based on $m = 3$ synthetic datasets by FBS and FBP, achieving $\epsilon_{\boldsymbol{y}_n} = 10.8$ for all three methods.

Adding Synthetic Data Replicates, m, Improves Utility

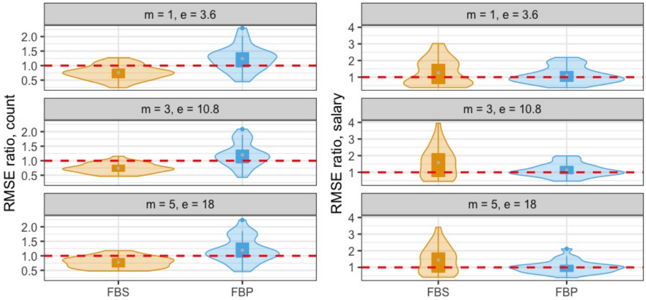

Figure: RMSE ratios of **counts** (left) and **average salary values** (right) of FBS and FBP, applied to the selected sample. A red dashed line at RMSE ratio = 1 is included for reference. Each violin plot represents a distribution of RMSE ratios over 27 cells. Results are based on $m = \{1, 3, 5\}$ synthetic datasets by FBS and FBP, achieving $\epsilon_{\boldsymbol{y}_n} = \{3.6, 10.8, 18\}$ for both methods.

Outline

Differential privacy

Two synthesizers

Laplace Mechanism

Survey of Doctoral Recipients Application

Simulation Study

Concluding remarks

Summary

- Formal privacy for data collected under an informative sampling design
- We recommend the FBS: easy to estimate and produces low RMSE
- The synthetic data is privacy protected and obeys all constraints without any post processing
- No interactive queries required
- The synthetic data may be used for other purposes
- arXiv link to manuscript: `https://arxiv.org/abs/2101.06188`

References

- Leon-Novelo, L. G. and Savitsky, T. D. (2019). Fully Bayesian estimation under informative sampling, *Electronic Journal of Statistics*, 13, 1608-1645.
- Savitsky, T. D., Williams, M. R. and Hu, J. (2020). Bayesian pseudo posterior mechanism under differential privacy. arXiv:1909.11796.
- Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys, *Survey Methodology*, **18**, 209-217.