Synthetic Data Generation for Survey Data

Discussion

Jingchen (Monika) Hu

Vassar College

FCSM 2023

Talk #1: Fully synthetic data for complex surveys

- Two proposals for generating fully synthetic data for complex surveys: Synrep-R and Synrep-1
- Both proposals: 1) resample via Bayesian bootstrap (M samples), 2) use the WFPBB to make pseudo-populations, and 3) draw SRS from each pseudo-population
- These two proposals differ in how to generate synthetic data replicates
  - Synrep-R: generating multiple (R) synthetic data sets from each SRS; total R*M
  - Synrep-1: generating one synthetic data set from each SRS: total M
- Newly developed combining rules for both proposals to enable variance estimation
- Questions:
  - Advice for values of M and R in practice?
  - Can popular synthesizers such as CART be easily implemented within these two proposals? Need for more synthesis models development?
  - What to do if the goal is to release fully synthetic populations?

Talk #2: Synthetic population generation for nested data using differentially private posteriors

- The work is built on the Pseudo Posterior Mechanism for DP synthetic data
  - Record-indexed privacy weights
  - Allowing surgical downweighting of high-risk records
  - Asymptotic DP
- Current work is extending
  - The neighborhood concept: from one record to one group of records
  - To latent variables: challenge of working with integrated likelihood
  - The weighting approach: challenge of two sets of weights (individual- and group-levels)
- Preliminary simulation results show little gain in privacy or utility for additional downweighting of groups after downweighting of individuals
- Questions:
  - Any intuition of why there is little gain in privacy or utility shown in the preliminary results?
  - Any intuition for when group sizes differ?

Talk #3: Private table statistics using synthetic microdata generation

- The work is built on the Pseudo Posterior Mechanism for DP synthetic data
- Current work is incorporating survey weights by modeling and use table statistics as data utility metrics
- Two fully synthesizing models: FBS for sample and FBP for population
- SDR application shows better performance of FBS in terms of count table and better performance of FBP in terms of average salary table
- Simulation studies with (more) informative sampling design show even better performance of FBS and FBP as Laplace Mechanism destroys utility
- Questions:
  - How challenging is it to extend the FBS and FBP for multivariate outcome variables?
  - How to effectively choose m to balance the privacy-utility tradeoff?

Talk #4: Calibration procedure for estimates obtained from posterior approximation algorithms, with application to domain-level modeling

- Goal is to develop a practical tool for the evaluation of the fit and correction of a bias in the posterior variance of the ADVI-based point estimator
- Proposal of a resampling-based approach
  - Assumption: model is correct but the fitting algorithm may produce systematic errors in uncertainty measures
  - Goal: to evaluate and potentially correct the fit
  - Method: adjustment of the pivot with pivot-based CIs and re-scaled CIs
- Simulation and CES results are positive
- Questions:
  - Are you considering extension of the method when the correct model assumption does not hold?
  - Can the proposed methods extend beyond the FH and CCFH models?

Thank you! Questions?

Jingchen (Monika) Hu

jihu@vassar.edu

Vassar College