Fully Synthetic Data for Complex Surveys

Yajuan Si[1]
Joint work with Shirley Mathur[2] and Jerry Reiter[3]

University of Michigan [1]
University of Washington [2]
Duke University [3]

The 2023 FCSM Research & Policy Conference
Oct 25, 2023

Outline

Introduction

Two Methods for Generating Fully Synthetic Survey Data

Inferences

Simulation Studies

Discussion

Fully synthetic data

- Synthetic data approaches have been proposed for data confidentiality and privacy protection (Rubin, 1993).
- Releasing fully synthetic data sets can preserve confidentiality, since identification of units and their sensitive data can be difficult when the released data are not actual, collected values, where the agency
  1. randomly and independently samples units from the sampling frame to comprise each synthetic data set,
  2. imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and
  3. releases multiple versions of these data sets to the public.
- Methods for inferences from these multiply-imputed data files have been developed for a variety of statistical inference tasks.

Complex sample surveys

- Previous research on multiple imputation (MI) for missing data suggests that imputation models should account for the survey design features, such as stratification, clustering, and survey weights (Reiter et al., 2006).
- Similarly, when using multiple imputation for synthetic data, the models also should account for the survey design (Mitra and Reiter, 2006; Fienberg, 2010).
- The key challenge is properly incorporating weights in the synthesis model, which relates to the long-standing debate about the role of survey weights in model-based inferences (Pfeffermann, 1993, 2011; Little, 2004).

Existing methods

- Bayesian finite population inference approach conditional on design features available for all population units
- Weighted finite population Bayesian bootstrap (WFPBB) (Dong et al., 2014)
- Pseudo-likelihood approach (Pfeffermann, 1993; Savitsky and Toth, 2016; Kim et al., 2021), in which each individual's contribution to the likelihood function of a synthesis model is raised to a power that is a function of the survey weights

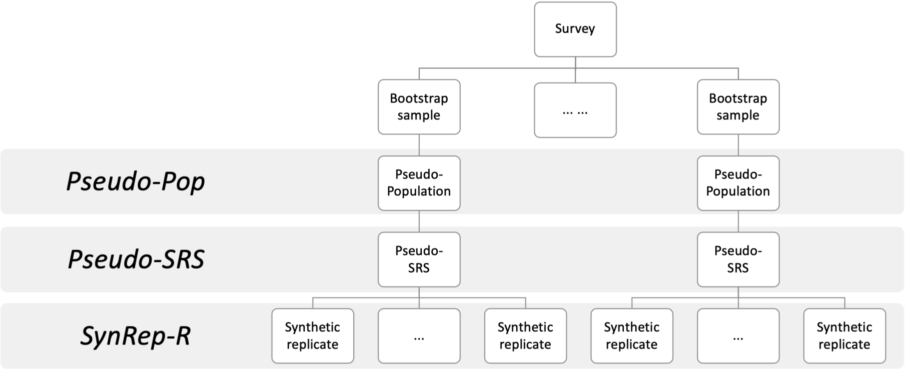
Potential concerns

- The Bayesian finite population inference approach, while theoretically principled, requires completing full populations, which can be cumbersome, and the availability of design variables for all records in the population, which may not be the case in some surveys.
- The WFPBB releases (multiple copies of) individuals' genuine data records, which creates obvious disclosure risks.
- Synthesizing weights does not have a theoretical basis; thus, it is unclear if this approach can adequately capture uncertainty from complex designs.
- Pseudo-likelihood approaches also may not estimate sampling variability correctly (Williams and Savitsky, 2021), and it is not clear how easily they can be implemented with machine learning synthesizers like classification and regression trees, which are commonly used in practical synthetic data projects.

New proposal (Mathur et al., 2023)

- We build on the WFPBB approach by first creating pseudo-populations that account for the survey weights.
- We then take simple random samples (SRSs) from each pseudo-population, estimate synthesis models from each SRS, and generate draws from these models to create multiply-imputed, fully synthetic public use files.
- We consider two synthetic survey data generation processes.
  1. We generate multiple synthetic data sets from each SRS; we call this *Synrep-R*.
  2. We generate one synthetic data set from each SRS; we call this *SynRep-1*.
- For both approaches, we derive multiple imputation combining rules that enable the estimation of variances.

Synrep-R



Survey

Bootstrap sample

... ...

Bootstrap sample

Pseudo-Pop

Pseudo-Population

Pseudo-Population

Pseudo-SRS

Pseudo-SRS

Pseudo-SRS

SynRep-R

Synthetic replicate

...

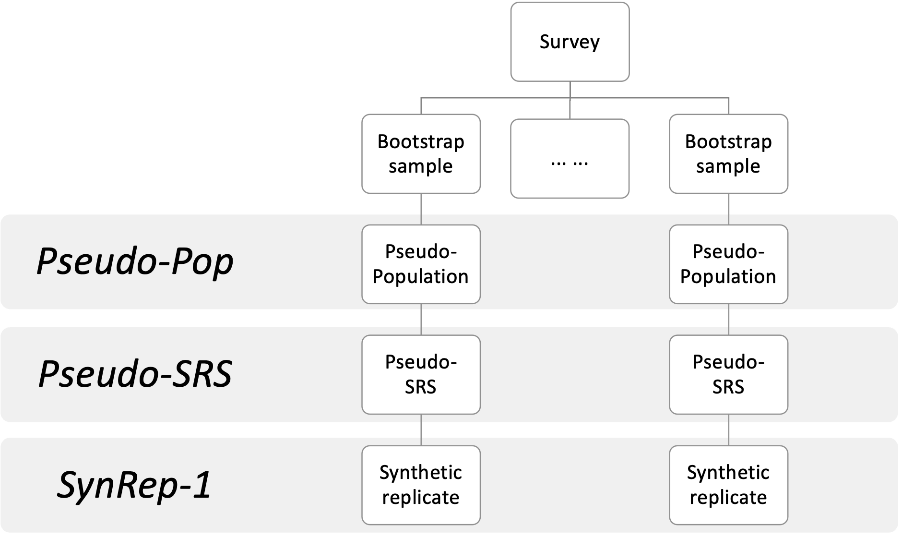
Synthetic replicate

Synthetic replicate

...

Synthetic replicate

Synrep-1



Survey

Bootstrap sample

… …

Bootstrap sample

Pseudo-Pop

Pseudo-Population

Pseudo-Population

Pseudo-SRS

Pseudo-SRS

Pseudo-SRS

SynRep-1

Synthetic replicate

Synthetic replicate

Synthetic data generation process

1. **Resample via Bayesian bootstrap**: To inject sufficient sampling variability, given the data from the "parent" sample $\mathcal{D}$, we generate *M* samples, $(\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(M)})$, each of size n using independent Bayesian bootstraps.
2. **Use the WFPBB to make pseudo-populations**: For each $\mathcal{S}^{(m)}$, we construct an initial Pólya urn using the set of $\{Y_i, w_i^{(m)}\}$. We then draw $N - n$ units using probabilities $(p_1^{(m)}, \ldots, p_n^{(m)})$ determined from

$$p_i^{(m)} = \frac{w_i^{(m)} - 1 + l_{i,k-1}^{(m)}(N-n)/n}{N - n + (k-1)(N-n)/n}, \tag{1}$$

for the kth draw, $k \in \{1, \ldots, N-n\}$, where $l_{i,k-1}^{(m)}$ is the number of bootstrap selections of Y_i among the elements present in the urn at the $k-1$ draw. The $N-n$ draws combined with the data in $\mathcal{S}^{(m)}$ comprise one pseudo-population, $\mathcal{P}^{(m)}$. We repeat this for $m = 1, \ldots, M$ to create $\mathcal{P}_{pseudo} = \{\mathcal{P}^{(m)} : m = 1, \ldots, M\}$.

Synthetic data generation process (cont.)

3. **Draw SRS from each pseudo-population**: For $m = 1, \ldots, M$, take a simple random sample $\mathcal{D}^{(m)}$ of size n from $\mathcal{P}^{(m)}$. Let $\mathcal{D}_{srs} = \{\mathcal{D}^{(m)} : m = 1, \ldots, M\}$.
4. **Generate synthetic data replicates:** For $m = 1, \ldots, M$, estimate a synthesis model using $\mathcal{D}^{(m)}$, and draw from the predictive distributions to form synthetic data replicates using either Step 4a or Step 4b.
   - 4a. **SynRep-R:** For $m = 1, \ldots, M$, draw $R > 1$ synthetic replicates $\mathcal{D}_{syn}^{(m,r)}$ of size n, where $r = 1, \ldots, R$, using each $\mathcal{D}^{(m)}$. We release $\mathcal{D}_{syn} = \{\mathcal{D}_{syn}^{(m,r)} : m = 1, \ldots, M; r = 1, \ldots, R\}$ including indicators of which m each $\mathcal{D}_{syn}^{(m,r)}$ belongs to.
   - 4b. **SynRep-1:** For $m = 1, \ldots, M$, draw one synthetic data sample $\mathcal{D}_{syn}^{(m)}$ of size n from each $\mathcal{D}^{(m)}$. Release $\mathcal{D}_{syn} = \{\mathcal{D}_{syn}^{(m)} : m = 1, \ldots, M\}$.

Combining rules for *SynRep-R*

For each $\mathcal{D}_{syn}^{(m,r)}$, let $q_{syn}^{(m,r)}$ be the point estimate of Q, and let $v_{syn}^{(m,r)}$ be the estimate of the variance associated with $q_{syn}^{(m,r)}$. The analyst needs to compute the following quantities.

$$\bar{q}_{syn}^{(m)} = \sum_{r=1}^{R} q_{syn}^{(m,r)}/R, \qquad \bar{q}_{syn} = \sum_{m=1}^{M} \bar{q}_{syn}^{(m)}/M \tag{2}$$

$$b_{syn} = \sum_{m=1}^{M} (\bar{q}_{syn}^{(m)} - \bar{q}_{syn})^2/(M-1) \tag{3}$$

$$w_{syn}^{(m)} = \sum_{r=1}^{R} (q_{syn}^{(m,r)} - \bar{q}_{syn}^{(m)})^2/(R-1), \qquad \bar{w}_{syn} = \sum_{m=1}^{M} w_{syn}^{(m)}/M \tag{4}$$

$$\bar{v}_{syn} = \sum_{m=1}^{M} \sum_{r=1}^{R} v_{syn}^{(m,r)}/MR \tag{5}$$

$$T_r = \left(1 + M^{-1}\right) b_{syn} - \bar{v}_{syn} - \bar{w}_{syn}/R. \tag{6}$$

We compute approximate 95% intervals for Q as $\bar{q}_{syn} \pm t_{0.975,M-1}\sqrt{T_r}$.

Combining rules for *SynRep-1*

The analyst computes each $(q_{syn}^{(m)}, v_{syn}^{(m)})$ by acting is if $\mathcal{D}_{syn}^{(m)}$ is a SRS of size n from the population. We require the following quantities for inferences.

$$\bar{q}_{syn} = \sum_{m=1}^{M} q_{syn}^{(m)}/M \tag{7}$$

$$b_{syn} = \sum_{m=1}^{M} (q_{syn}^{(m)} - \bar{q}_{syn})^2/(M-1) \tag{8}$$

$$\bar{v}_{syn} = \sum_{m=1}^{M} v_{syn}^{(m)}/M \tag{9}$$

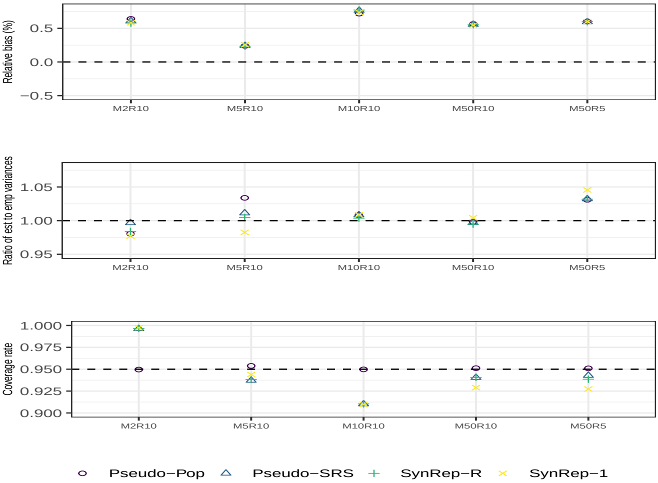
$$T_m = \left(1 + M^{-1}\right) b_{syn} - 2\bar{v}_{syn}. \tag{10}$$

We compute approximate 95% intervals for Q as $\bar{q}_{syn} \pm t_{0.975,M-1}\sqrt{T_m}$.

Repeated sampling properties

- We conduct simulation studies to compare the repeated sampling performances of different inferential methods, estimating the finite population mean.
- Alternative estimators include:
  1. *Pseudo-Pop* as the procedure that uses a point estimator of $\bar{Q}$ and variance estimator of $(1 + 1/M)B$ computed with the WFPBB-generated pseudo-populations $(\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(M)})$
  2. *Pseudo-SRS* as the procedure that uses a point estimator of $\bar{q}$ and variance estimator of Raghunathan et al. (2003) computed with $(\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(M)})$
  3. *SRSsyn* as the procedure that generates synthetic data by using the unweighted sample mean and standard deviation as plug-in parameters in a normal distribution
  4. *Direct* using the unweighted sample mean and standard deviation from $\mathcal{D}$, i.e., ignoring the survey weights
  5. *HT* as the Horvitz and Thompson (1952) estimator and its estimated variance using $\mathcal{D}$.

Under a PPS design with different (M, R) values



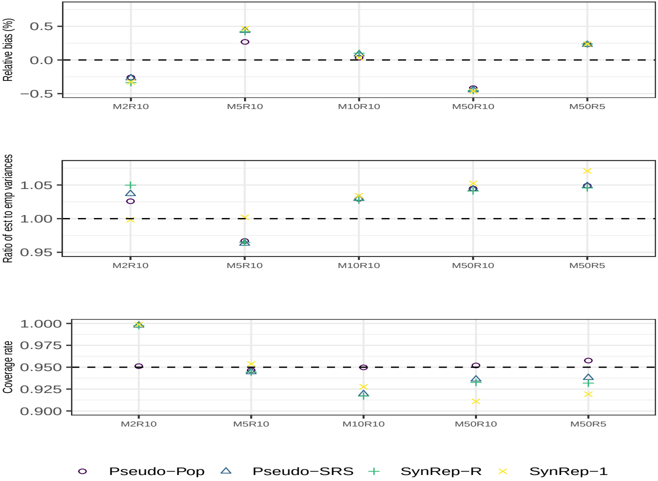
Direct and *SRSsyn* are not plotted because they result in large biases ($\approx 39\%$) and very low coverage rates (mostly close to 0). The *HT* method is not plotted but results in unbiased estimates and near nominal coverage rates, as expected.

Issues with negative variances

Table: Percentage of negative variance estimates in the PPS simulation studies. When $M = 50$, all variance estimates are positive.

(M, R)	Method	Percentage of negative variances (%)
M2R10	Pseudo-SRS	36
M2R10	SynRep-2	37
M2R10	SynRep-1	43
M5R10	Pseudo-SRS	11
M5R10	SynRep-2	13
M5R10	SynRep-1	21
M10R10	Pseudo-SRS	3
M10R10	SynRep-2	3
M10R10	SynRep-1	10

Under a SRS design



Estimates from *SynRep-R* and *SynRep-1* do not lose much efficiency relative to the estimators from *Pseudo-SRS*.

Discussion

- *SynRep-R* and *SynRep-1* represent a general strategy for constructing fully synthetic data from complex samples: use the WFPBB to undo the complex design, then replace the confidential values with synthetic data.
- There are many related topics worth further investigation, e.g.,
  1. use model-based approaches to smooth the weights
  2. extend to multivariate data and for various estimands of interest, e.g., subdomain means and regression coefficients
  3. account for missing survey data
  4. posterior sampling inference
  5. comparison with other proposed methods for full synthesis with complex samples

Acknowledgements

The work was funded by the U.S. National Science Foundation grant (SES 2217456) and a pilot project from the Michigan Center on the Demography of Aging with funding from the National Institute on Aging (P30 AG012846).

References

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, 40(1):29–46.

Fienberg, S. E. (2010). The relevance or irrelevance of weights for confidentiality and statistical analyses. *Journal of Privacy and Confidentiality*, 1:183–195.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite university. *Journal of the American Statistical Association*, 47(260):663–685.

Kim, H. J., Drechsler, J., and Thompson, K. J. (2021). Synthetic microdata for establishment surveys under informative sampling. *Journal of Royal Statistical Society, Series A*, 184:255–281.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling inference for finite population sampling. *Journal of the American Statistical Association*, 99:546–556.

Mathur, S., Si, Y., and Reiter, J. P. (2023). Fully synthetic data for complex surveys. https://arxiv.org/abs/2309.09115.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In Domingo-Ferrer, J. and Franconi, L., editors, *Privacy in Statistical Databases*, pages 177–188. New York: Springer-Verlag.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317–337.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, 37(2):115–136.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology*, 32:143–150.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:462–468.

Savitsky, T. D. and Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10(1):1677–1708.

Williams, M. R. and Savitsky, T. D. (2021). Uncertainty estimation for pseudo-Bayesian inference under complex sampling. *International Statistical Review*, 89(1):72–107.