A Comparison of Strategies for Generating Synthetic Data for Complex Survey

by

Min Chen

Department of Statistical Science
Duke University

Defense Date: April 1, 2024

Approved:

Jerome Reiter, Supervisor

Jerome Reiter

Jason Xu

David Banks

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2024

ABSTRACT

A Comparison of Strategies for Generating Synthetic Data for Complex Survey

by

Min Chen

Department of Statistical Science
Duke University

Defense Date: April 1, 2024

Approved:

Jerome Reiter, Supervisor

Jerome Reiter

Jason Xu

David Banks

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2024

Abstract

Synthetic data is a type of method for protecting data privacy. In the context of disseminating confidential data for public utilization, some statistical agencies employ the generation of fully synthetic datasets. This practice is applied to census and administrative records. It is important to note that many research datasets come from surveys with complex sampling methods, which is not ignorable when constructing synthetic data. The thesis presents an illustration for the comparison of three different synthetic data strategies. Each of them has different procedures to generate the synthetic data. Two of them are based on the bootstrap methods, one is Bayesian bootstrap, and the other is regular bootstrap. The third method is based on the posterior inference with pseudo-likelihood. Using simulation studies with probability proportional to size sampling, we show that all three methods can result in accurate estimates of the mean of a finite population. However, when estimating the sampling statistic's variance, only the method based on the Bayesian bootstrap method can provide an approximately unbiased estimate in these simulations.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to first express my gratitude to my respectful professor Prof. Jerry Reiter, who gave me valuable instructions on the synthetic data topics. I will also express my respect for Prof. Jason Xu and Prof. Eric Laber, who instructed me on doing researches on stochastic modeling and clinical trial, respectively. In addition, I am glad to have Prof. David Banks be my master thesis committee member.

1. Introduction

The synthetic data approach has piqued the interest of statisticians across academia and governmental statistical offices. For instance, Raghunathan et al. (2003) devised strategies for amalgamating point and variance estimates from various synthetic datasets to facilitate valid statistical inferences. It is important to note that their aggregation principles vary from those applied to multiple imputations of missing data, as originally proposed by Rubin (1987). Moreover, Reiter (2002) illustrated that these methodologies can yield accurate inferences in a broad range of survey design scenarios. In addition, statistical agencies are also increasingly leveraging it to create publicly accessible datasets from confidential information.

Synthetic data is generated through statistical modeling of confidential data to produce new records. For example, Abowd et al. (2006) and Kinney et al. (2011) describe how the Census Bureau has released synthetic datasets for both the Survey of Income and Program Participation and the Longitudinal Business Database. In addition, Statistics Canada, Statistics New Zealand, the Australian Bureau of Statistics, the UK Office of National Statistics, and the Scottish Longitudinal Study, all showcased its growing adoption for ensuring data privacy while maintaining utility, according to the report of United Nations Economic Commission for Europe (2022).

Generally, there are two different types of synthetic data. One approach involves partially synthetic data, which retains the original surveyed units but replaces certain values, such as sensitive information with high disclosure risk or key identifiers, with synthetic values through multiple imputations (Little, 1993; Reiter, 2003). The other approach generates fully synthetic data, where the entire dataset is constructed synthetically, according to Raghunathan et al. (2003) and Rubin (1993). This thesis will concentrate on the method of fully synthetic data. The attractiveness of this approach lies in its ability to safeguard confidentiality, as it becomes challenging to identify units and their sensitive information when the data released does not correspond to the collected values. Furthermore, employing suitable imputation and estimation techniques grounded in the principles of multiple

imputation (Rubin, 1987), this method enables users to conduct valid analyses on different estimands using conventional statistical methods and software designed for complete datasets (Reiter, 2005).

In the realm of synthetic data, most applications have been developed for census or administrative data. However, the landscape of research data is diverse, with a significant portion derived from surveys that employ sampling structures, like multi-stage stratified cluster samples. Kim et al. (2020) illustrates that the sample design becomes informative when dealing with complex sampling techniques, like stratified sampling or probability proportional to size (PPS) sampling. This means that the distribution of the observations y_i within the sample will deviate from the distribution observed in the original population.

Given this context, agencies need to incorporate survey design nuances when formulating synthetic data. Thus, it is reasonable for agencies to consider the survey design when generating synthetic data. Then, a pertinent question is, how should agencies construct synthetic data concerning complex survey designs? Through this project, we intend to assess extant techniques and offer recommendations for assimilating complex survey designs into synthetic data generation.

In this research, we mainly adopt the Bayesian finite population inference framework to guide the generation of synthetic data (Reiter, 2002; Rubin, 1993). This framework involves a series of inferential procedures that consider the nuances of the synthetic data generation process. Typically, this process encompasses three primary steps usually, (1) construct predictive models for survey variables, leveraging design variables like size metrics or cluster indicators, which the agency presumes are known for each unit in the population, (2) do imputation for the missing survey variables for the units that were not sampled within the population, and (3) draw a simple random sample from the fully populated data to produce a synthetic dataset.

We explore three different approaches to generating synthesis data in this thesis. The first method involves using the survey sample to create pseudo-populations. From these, pseudo-samples are drawn, leading to the generation of synthetic datasets, where we esti-

mate the data-generation model and release simulated values from the model as synthetic data. This method, mainly based on the Weighted Finite Population Bayesian Bootstrap (WFPBB), comes from Mathur et al. (2023). The second method follows a similar path. It also starts with the creation of a pseudo-population using the survey sample and draws pseudo-samples from these pseudo-populations. The synthetic dataset is subsequently produced through the estimated data generation from these pseudo-samples as the procedure of the WFPBB method. Despite the similar general process, the method, differs significantly in how the pseudo-populations are created, impacting variance estimation outcomes. This approach is grounded in the work of Mashreghi et al. (2016). Both methods apply the bootstrap technique, initially introduced by Efron (1979). Initially, this technique was used in traditional statistics for data that are independently and identically distributed (i.i.d.) from an unknown distribution. It helps estimate the accuracy of certain statistical measures, like variance or confidence intervals. However, survey data do not always follow the i.i.d. assumption. Thus, various bootstrap resampling techniques have been developed for survey sampling over the years. These techniques modify the original i.i.d. bootstrap method to consider survey-specific aspects.

The third method was referred to Kim et al. (2020) and Savitsky et al. (2022), which estimates an approximate posterior distribution for parameters from synthetic data models and then creates the fully synthetic data by drawing datasets from the models. To utilize the design information in survey weights, Kim et al. (2020) relied on the pseudo-likelihood approach when building the Bayesian model. Then, synthetic populations can be randomly drawn from the estimated distributions. The full conditional distributions of the Markov chain Monte Carlo algorithm can be applied for posterior inferential methods with the pseudo-posterior likelihood.

In this thesis, I describe three different strategies for generating synthetic data in the case of a complex survey and design a simulation to compare them in terms of the estimation precision for the population mean and the corresponding variance. The remainder of the thesis is presented as follows. Chapter 2 describes all three methods and their inference for

estimating the population quantities and their associated variance. Chapter 3 presents a simulation study and the corresponding result. Chapter 4 is the conclusion.

2. Synthetic Data Methods and Inference Process

We begin by designating $\mathcal{P}$ as the finite population within which we conduct our survey and the population size is N. Y is a $p \times 1$ survey vector we are interested in, X is the auxiliary variable representing the complex survey design variables such as stratum or size indicators, and for each $i = 1, \ldots, N$, we denote π_i as the selection probability for each unit, conditioned on X. Given the selection probability, the survey weight can be calculated as $\omega_i = 1/\pi_i$. For simplicity consideration, we here treat $p = 1$, so Y is a scalar.

Let $\mathcal{D}$ be a probability sample with size n, sampled from the population proportional to the survey weights. The $\mathcal{D}$ can be expressed as $\mathcal{D} = \{(\omega_i, X_i, Y_i) : i = 1, \ldots, n\}$. The synthetic data can be generated from the survey sample $\mathcal{D}$ in many different ways, and in this thesis, we denote the fully synthetic data as $\mathcal{S}$. After the generation, the data $\mathcal{S}$ will be released to the public instead of the confidential data to protect the data privacy.

In section 2.1, we discuss the Weighted Finite Population Bayesian Bootstrap (WFPBB) method. In section 2.2, we illustrate a method that is similar to the WFPBB, which we call the Weighted Pseudo Population Bootstrap Method. In section 2.3, we present the Pseudo-likelihood method.

2.1 WFPBB

The Weighted Finite Population Bayesian Bootstrap (WFPBB), as referenced in Dong et al. (2014), is closely aligned with the Bayesian finite population inference method. In steps (1) and (2), the agency creates complete populations by replicating the observed data based on their survey weights. However, Dong et al. (2014) does not take the simple random sampling performed in step (3) and model-based synthesis performed in step (4), which we describe below. Figure 2.1 illustrates the process of the WFPBB (Mathur et al., 2023).

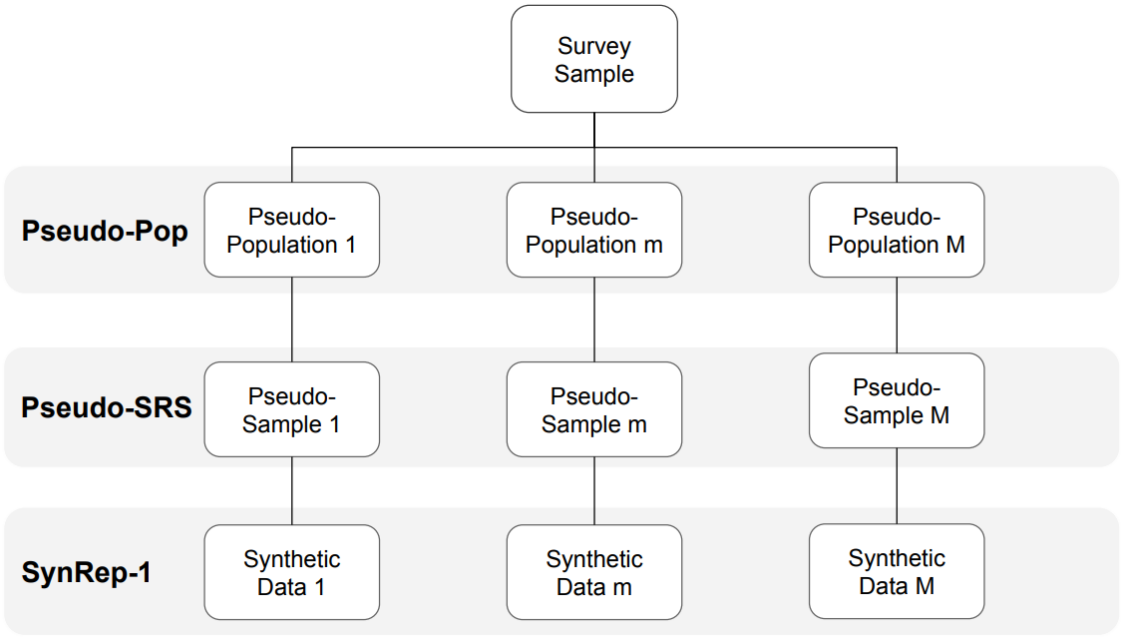

FIGURE 2.1: Process for generating fully synthetic datasets by WFPBB.

The modified WFPBB, which is summarized in section 2.1.1 below, is used to generate fully synthetic data, which closely follow the steps in Mathur et al. (2023), shown in detail below.

2.1.1 Method

The process of the WFPBB approach mainly contains four steps. Here we review the procedure presented in the work of Mathur et al. (2023).

1. **Resample by Bayesian Bootstrap:**

   To introduce sufficient sampling variability based on data from $\mathcal{D}$, we generate M samples, $(BS^{(1)}, \ldots, BS^{(M)})$, each of size n, using independent Bayesian bootstraps (Rubin, 1981). For each sample $BS^{(m)}$, where $m = 1, \ldots, M$, and for each element $i = 1, \ldots, n$, we define the weight $w_i^{(m)} = \frac{N w_i r_i^{(m)}}{\sum_{k=1}^{n} w_k r_k^{(m)}}$. Here, $r_i^{(m)}$ denotes the number of times that the i-th element from $\mathcal{D}$ appears in the bootstrap sample $BS^{(m)}$, w_i is the original weight of the i-th element in $\mathcal{D}$, and N is the population size. The constant

c serves as a normalizing factor to ensure that the sum of the new weights equals N.

2. **Creating the Pseudo-Population by WFPBB:**

   For each designated $BS^{(m)}$, an initial *Polya* urn is formulated, utilizing the set $\left\{ Y_i, w_i^{(m)} \right\}$. Subsequently, $N-n$ units are drawn based on the probabilities $(p_1^{(m)}, \ldots, p_n^{(m)})$, which are derived from the following equation:

$$p_i^{(m)} = \frac{w_i^{(m)} - 1 + l_{i,k-1}^{(m)} \left(\frac{N-n}{n}\right)}{N - n + (k-1) \left(\frac{N-n}{n}\right)}. \tag{2.1}$$

   This is for the k-th draw, where $k \in \{1, \ldots, N-n\}$, and $l_{i,k-1}^{(m)}$ denotes the number of bootstrap selections of Y_i from dataset $BS^{(m)}$ in the urn among the previous $k-1$ draws. The union of the $N-n$ draws and the data contained in $BS^{(m)}$ forms a single pseudo-population, represented as $\mathcal{P}^{(m)}$. This process is reiterated across $m = 1, \ldots, M$, resulting in the establishment of $\mathcal{P}_{pseudo} = \left\{ \mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(M)} \right\}$. In some cases, N is considerably large, and generating a same-size pseudo-population may not be requisite. This pragmatic approach can lead to substantial savings in memory consumption and an enhancement in computational efficiency. Typically, a pseudo-population comprising $50n$ records, rather than $N-n$ or N, would be sufficient.

3. **Draw SRS from each Pseudo-Populations**

   For $m = 1, \ldots, M$, take a simple random sample $\mathcal{D}_{srs}^{(m)}$ of size n from $\mathcal{P}^{(m)}$.

4. **Generate Synthetic Data:**

   For $m = 1, \ldots, M$, we estimate a synthetic model using $\mathcal{D}_{srs}^{(m)}$ and draw from the predictive distributions to form synthetic data replicates. In this thesis, for $m = 1, \ldots, M$, we draw one synthetic data sample $S^{(m)}$ of size n from the estimated synthetic model calculated by the $\mathcal{D}_{srs}^{(m)}$, and release $\mathcal{S} = \left\{ S^{(m)} : m = 1, \ldots, M \right\}$.

2.1.2 Inference

In this section, we review the derivations for inferences with the WFPBB method as derived in Mathur et al. (2023). Our review closely follows their work.

For any population quantity Q, for instance, Q is the population mean $\bar{Y}$, we want to get the posterior distribution $P(Q|\mathcal{S})$. Following the steps in Raghunathan et al. (2003), we derive the following integral based on each level of the data synthesis process from Figure 2.1.

$$\begin{aligned} P(Q|\mathcal{S}) = \int\int\int & P(Q|\mathcal{S}, \mathcal{D}_{srs}, \mathcal{P}_{pseudo}, \mathcal{D}) \times \\ & P(\mathcal{D}|\mathcal{S}, \mathcal{D}_{srs}, \mathcal{P}_{pseudo}) \times \\ & P(\mathcal{P}_{pseudo}|\mathcal{S}, \mathcal{D}_{srs}) \times \\ & P(\mathcal{D}_{srs}|\mathcal{S})\, d\mathcal{D}\, d\mathcal{P}_{pseudo}\, d\mathcal{D}_{srs}. \end{aligned} \tag{2.2}$$

According to Mathur et al. (2023), when conditioning on $\mathcal{D}$, the values of $(\mathcal{S}, \mathcal{D}_{srs}, \mathcal{P}_{pseudo})$ do not provide information for Q, so we can replace the density $P(Q|\mathcal{S}, \mathcal{D}_{srs}, \mathcal{P}_{pseudo}, \mathcal{D})$ with $P(Q|\mathcal{D})$. When conditioning on $\mathcal{P}_{pseudo}$, the values of $(\mathcal{S}, \mathcal{D}_{srs})$ do not provide any information for $\mathcal{D}$. Thus, we can simplify the density by $P(\mathcal{D}|\mathcal{P}_{pseudo})$. When conditioning on $\mathcal{D}_{srs}$, the value of $\mathcal{S}$ does not provide additional information for $\mathcal{P}_{pseudo}$. Therefore, $P(\mathcal{P}_{pseudo}|\mathcal{S}, \mathcal{D}_{srs})$ could be written as $P(\mathcal{P}_{pseudo}|\mathcal{D}_{srs})$. Given these, the original formula can be rewritten as

$$P(Q|\mathcal{S}) = \int \left[\int \left[\int P(Q|\mathcal{D}) P(\mathcal{D}|\mathcal{P}_{pseudo})\, d\mathcal{D} \right] P(\mathcal{P}_{pseudo}|\mathcal{D}_{srs})\, d\mathcal{P}_{pseudo} \right] P(\mathcal{D}_{srs}|\mathcal{S})\, d\mathcal{D}_{srs}. \tag{2.3}$$

We start with $P(Q|\mathcal{P}_{pseudo}) = \int P(Q|\mathcal{D}) P(\mathcal{D}|\mathcal{P}_{pseudo}) d\mathcal{D}$. We presume that, given a sufficiently large M, this approximates a normal distribution. This assumption is typically justified in the context of large sample sizes, which align with the circumstances under which agencies are inclined to release public use datasets. Since normal distributions are fully

described by their means and variances, our estimations concentrate on the distributions of these initial two moments. Consequently, for each $m = 1, \ldots, M$, we define $Q^{(m)}$ as the calculated value of Q assuming the availability of the pseudo-population $\mathcal{P}^{(m)}$. Rubin (1987) illustrates that,

$$(Q|\mathcal{P}_{pseudo}) \sim t_{M-1}(\bar{Q}, (1 + M^{-1}B) \tag{2.4}$$

where $\bar{Q}$ is the average of the $Q^{(m)}$ values, calculated as $\bar{Q} = \frac{\sum_m Q^{(m)}}{M}$, and M is the number of measurements. Additionally, define $B = \frac{\sum_m (Q^{(m)} - \bar{Q})^2}{(M-1)}$, which represents a normalized sum of squared deviations from the average $\bar{Q}$. $t_v(\mu, \sigma^2)$ represents a t-distribution with ν degrees of freedom, where μ and σ^2 denote the location and variance, respectively. For the sake of simplification in the derivations, we approximate the t-distribution, as referenced in (2.4), with a normal distribution. This approximation is generally reasonable when M is relatively large (Raghunathan et al., 2003).

Following the work of Mathur et al. (2023), we next turn to $P(\mathcal{P}_{pseudo}|\mathcal{D})$, where we specifically need to consider $P(\bar{Q}, B|\mathcal{D}_{srs})$. For each m in the range from 1 to M, let $q^{(m)}$ represent the estimated value of $Q^{(m)}$ and $v^{(m)}$ denote the estimated sampling variance associated with $q^{(m)}$. These estimations could be computed if we had access to $\mathcal{D}^{(m)}$. Following Mathur et al. (2023), we operate under the assumption that the set $\{q^{(m)}, v^{(m)} : m = 1, \ldots, M\}$ is valid in the following sense,

1. For each m, the estimator $q^{(m)}$ is nearly unbiased for $Q^{(m)}$ and follows an asymptotic normal distribution. This is valid when repeatedly sampling from the pseudo population $\mathcal{P}^{(m)}$, which has a sampling variance denoted by $V^{(m)}$. Thus, we can express $(q^{(m)}|\mathcal{P}^{(m)})$ as being distributed normally with mean $Q^{(m)}$ and variance $V^{(m)}$, i.e., $(q^{(m)}|\mathcal{P}^{(m)}) \sim N(Q^{(m)}, V^{(m)})$.
2. The estimated sampling variance, represented by $v^{(m)}$, is nearly unbiased for $V^{(m)}$. Furthermore, the fluctuation in $v^{(m)}$ due to sampling is minor, suggesting that $(v^{(m)}|\mathcal{P}^{(m)})$ is approximately equal to $V^{(m)}$.

3. The variability of $V^{(m)}$ across M pseudo-populations is minor, leading to $V^{(m)} \approx V \approx \bar{v}$, where $\bar{v} = \frac{1}{M}\sum_m v^{(m)}$ represents the average of the sampling variances.

Leveraging normal Bayesian arguments based on these sampling distributions, it follows that

$$(Q^{(m)}|q^{(m)}, \bar{v}) \overset{\text{ind}}{\sim} N(q^{(m)}, \bar{v}) \tag{2.5}$$

$$(\bar{Q}|\bar{q}, \bar{v}) \overset{\text{ind}}{\sim} N(\bar{q}, \bar{v}/M) \tag{2.6}$$

where $\bar{q} = \sum_m \frac{q^{(m)}}{M}$.

To derive the distribution of $(Q|\mathcal{D}_{srs})$, we integrate the distribution in (2.4), which is approximated to be a normal distribution, over the distributions of $\bar{Q}$ and B. Given that the resultant distribution conforms to approximately a normal distribution, it necessitates only the first two moments for its characterization. Consequently, we obtain the following,

$$E(Q|\mathcal{D}_{srs}) = E(E(Q|\bar{Q})|\mathcal{D}_{srs}) = E(\bar{q}|\mathcal{D}_{srs}) = \bar{q}. \tag{2.7}$$

We also have,

$$\begin{aligned} Var(Q|\mathcal{D}_{srs}) &= E(Var(Q|\mathcal{P}_{pseudo})|\mathcal{D}_{srs}) + Var(E(Q|\mathcal{P}_{pseudo})|\mathcal{D}_{srs}) \\ &= (1 + M^{-1})E(B|\mathcal{D}_{srs}) + \frac{\bar{v}}{M}. \end{aligned} \tag{2.8}$$

This is the variance estimator presented in Raghunathan et al. (2003) and is useful for analysts working with $\mathcal{D}_{srs}$. However, in our case, we have gone a step further by replacing each $\mathcal{D}_{srs}^{(m)}$ with simulated values. As a result, we need to consider the distributions of $(\bar{q}, \bar{v}, B)$ and take an average over them.

To derive inferences for Q within this context, we use the approach suggested by Mathur et al. (2023) who apply the approach outlined by Raab et al. (2018). They noted that valid variance estimations can be achieved from single implicates with adjustments to the combining rules, particularly when the source data originate from a simple random sample, as seen with each $\mathcal{D}_{srs}^{(m)}$. The derivation of this process is detailed below.

For each $m = 1, \ldots, M$, define $q_{syn}^{(m)}$ as the point estimate of Q derived using $S^{(m)}$, and $v_{syn}^{(m)}$ as the corresponding estimated variance for $q_{syn}^{(m)}$. The analyst calculates each pair of $(q_{syn}^{(m)}, v_{syn}^{(m)})$ under the assumption that $S^{(m)}$ represents a simple random sample of size n from the population. The derivation of inferences will utilize these computed quantities.

$$\bar{q}_{syn} = \sum_{m=1}^{M} q_{syn}^{(m)} / M \tag{2.9}$$

$$b_{syn} = \sum_{m=1}^{M} (q_{syn}^{(m)} - \bar{q}_{syn})^2 / (M-1) \tag{2.10}$$

$$\bar{v}_{syn} = \sum_{m=1}^{M} v_{syn}^{(m)} / M. \tag{2.11}$$

Following the logic of Raab et al. (2018), Mathur et al. (2023) assume that $q_{syn}^{(m)} \sim N(q_{syn}^{(m)}, V^{(m)})$ and $V^{(m)} \approx \bar{v}$ for all m, so we have

$$(q^{(m)} | S^{(m)}, \bar{v}) \sim N(q_{syn}^{(m)}, \bar{v}) \tag{2.12}$$

$$(\bar{q} | \mathcal{S}, \bar{v}) \sim N(\bar{q}_{syn}, \bar{v}/M). \tag{2.13}$$

Since $\mathcal{D}$ is a complex sample, it yields sampling variances that could differ from the simple random sampling variances associated with $\mathcal{D}_{srs}$.

Since all the components are approximately normal distributions, $P(Q|\mathcal{S})$ is also an approximate normal distribution. For its expectation, we use (2.7) and (2.13) to obtain

$$E(Q|\mathcal{S}) = E(E(Q|\mathcal{D}_{srs})|\mathcal{S}) = E(\bar{q}|\mathcal{S}) = \bar{q}_{syn}. \tag{2.14}$$

As for the variance, we write the variance in terms of $(B, \bar{v})$ and then plug in point estimates of these terms. We have

$$\begin{aligned} Var(Q|\mathcal{S}, B, \bar{v}) &= E(((1 + M^{-1})B + \bar{v}/M)|\mathcal{S}, B, \bar{v}) + Var(\bar{q}|\mathcal{S}, B, \bar{v}) \\ &= (1 + M^{-1})B + \bar{v}/M + \bar{v}/M \\ &= (1 + M^{-1})B + 2\bar{v}/M. \end{aligned} \tag{2.15}$$

We proceed to establish estimates for $(B, \bar{v})$ to be incorporated into (2.15). For $\bar{v}$, we utilize $\bar{v}_{syn}$ as defined in (2.15). This approximation is justifiable, given that we replace the entirety of each $\mathcal{D}^{(m)}$ set with synthetic values. To estimate B, we observe that the sampling distribution of a randomly generated $q_{syn}^{(m)}$ across all data generation steps adheres to a normal distribution $N(Q, B + 2\bar{v})$. Consequently, we derive that

$$\frac{\sum_{m=1}^{M}(q_{syn}^{(m)} - \bar{q}_{syn})^2}{B + 2\bar{v}} \sim \chi^2_{M-1}, \tag{2.16}$$

and hence,

$$E\left(\frac{\sum_{m=1}^{M}(q_{syn}^{(m)} - \bar{q}_{syn})^2}{B + 2\bar{v}}\right) = M - 1. \tag{2.17}$$

Applying a method of moments and referring to the definition of b_{syn} in (2.10), Mathur et al. (2023) finds that $b_{syn} \approx B + 2\bar{v}_{syn}$, leading to $B \approx b_{syn} - 2\bar{v}_{syn}$.

Therefore, the approximate variance $Var(Q|\mathcal{S})$ can be estimated using T_m, where

$$T_m = (1 + M^{-1})b_{syn} - 2\bar{v}_{syn}. \tag{2.18}$$

Given this, we can calculate the 95% intervals for Q as $\bar{q}_{syn} \pm t_{0.975,M-1}\sqrt{T_m}$. Some T_m may be negative in some cases. If so, Mathur et al. (2023) suggests replacing the B by $\bar{v}$ in equation (2.15), so that the replaced T_m will be

$$T_m^* = (1 + 3/M)\bar{v}_{syn}. \tag{2.19}$$

2.2 Weighted Pseudo Population Bootstrap Method

The method we illustrate here is closely aligned with the work of Mashreghi et al. (2016), and the overall framework is similar to what we present in Figure 2.1. However, in this approach, we use a less complicated method than the WFPBB to create the pseudo-populations.

2.2.1 Method

The methodology is tailored for unequal (single-stage) probability sampling designs, to create the pseudo-populations and pseudo-samples. Chauvet (2007) introduced a method applicable to Poisson sampling, while Holmberg (1998) developed a technique for sampling proportional to size, commonly known as πPS. Both methods strive to mirror the original sampling design, similar to the approach used in simple random sampling without replacement. Using the simple random sample without replacement, a πPS sample of size n selects the i-th element with probability $\pi_i = \frac{nx_i}{t_X}$, where $t_X = \sum_{i=1}^{N} x_i$ is the total sum of the size variable. Therefore, the probability of selecting is proportional to the size variable. When the correlation between X and Y is significant, this sampling method is notably effective for estimating the total using the Horvitz-Thompson estimator (Murthy, 1970). Furthermore, during Poisson sampling, each population element is independently chosen for the sample with a probability of π_i, making the sample size variable. Therefore, we could describe two methods with the following steps,

1. **Creating the Fixed Part of the Pseudo Population:**

   Duplicate the pair (y_i, π_i) within a single $\mathcal{D}$, exactly $\left\lfloor \pi_i^{-1} \right\rfloor$ times for every i in $\mathcal{D}$, thereby forming $\mathcal{P}^f$, which constitutes the fixed component of the pseudo population.

2. **Completing the Pseudo Population:**

   To finalize the composition of the pseudo-population, denoted as $\mathcal{P}^{(m)}$, we conduct a Poisson sampling of $\mathcal{P}^c$ from the set $\{(y_i, \pi_i) : i \in \mathcal{D}\}$, employing an inclusion probability of $\pi_i^{-1} - \left\lfloor \pi_i^{-1} \right\rfloor$ for each i^{th} element. Consequently, the pseudo-population is constituted as $\mathcal{P}^{(m)} = \mathcal{P}^f \cup \mathcal{P}^{(m)c}$, encapsulated by the set $\{(\check{y}_i, \check{\pi}_i)\}$ for every i within $\mathcal{P}^{(m)}$. In this context, each $\{(\check{y}_i, \check{\pi}_i)\}$ pair represents an element of the pseudo-population, corresponding to a value from the sample and its associated selection probability, adhering to the original sampling design.

3. **Draw Simple Random Samples from each Pseudo Population:**

   For $m = 1, \ldots, M$ we draw a simple random sample $\mathcal{D}_{srs}^{(m)}$ of size n from $\mathcal{P}^{(m)}$.

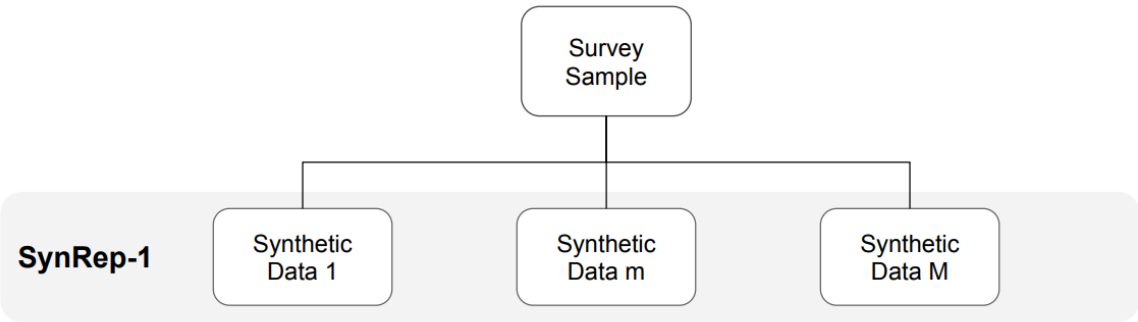

FIGURE 2.2: Process for generating fully synthetic datasets by Bayesian Pseudo Posterior.

4. **Generating Synthetic Data:**

For $m = 1, \ldots, M$, we estimate a synthetic model by using $\mathcal{D}_{srs}$ and draw from the predictive distributions to form synthetic data replicates. For $m = 1, \ldots, M$, we draw one synthetic data sample $S^{(m)}$ of size n from each $\mathcal{D}_{srs}^{(m)}$. Then, we release $\mathcal{S} = \left\{ S^{(m)} : m = 1, ..., M \right\}$

2.2.2 Inferences

We use the same inferential approach as for the WFPBB method described in Section 2.1

2.3 Bayesian Pseudo Posterior Method

The generation of record-level synthetic databases employs the Bayesian pseudo-posterior mechanism, as detailed in Savitsky et al. (2022) and Kim et al. (2020). Figure 2.2 illustrates the general procedure of the pseudo-likelihood method.

2.3.1 Method

The pseudo-likelihood is formed by exponentiating the likelihood contribution for each observed unit in the sample by its sampling weight, w_i, in,

$$pseudo(\theta; Y, w) = \prod_{i=1}^{n} p(y_i|x_i, \theta)^{w_i}, \tag{2.20}$$

$$pseudo(\theta; X, w) = \prod_{i=1}^{n} p(x_i|\theta)^{w_i}, \tag{2.21}$$

where $w_i \propto 1/\pi_i,\ i = 1, \ldots, n$, is the standardized sampling weight such that $\sum w_i = n$, where n denotes the observed sample size in $\mathcal{D}$. This is required to regulate the amount of estimated posterior uncertainty. Let $\mathbf{w} = (w_1, \ldots, w_n)$ denote the vector of these unit-indexed sampling weights.

The Bayesian pseudo-posterior is constructed through a multi-step process, which is shown in Savitsky et al. (2022) and Kim et al. (2020). Initially, the probability space, denoted as (Θ, β_Θ), is established and equipped with a prior distribution $\xi(\theta)$. Upon observing a database sequence, $Y = (Y_1, \ldots, Y_n)$, where $Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} P_{\theta_0}$ for a certain $\theta_0 \sim \Theta$, we proceed to formulate the pseudo-likelihood. Given these, by employing a Markov chain Monte Carlo (MCMC) method, the parameters in Θ can be updated individually using their full conditional distributions, respectively (Kim et al., 2020).

In practice, after t burn-in period, we can apply the MCMC algorithm for T iterations to draw synthetic data, $\mathcal{S}$, from the pseudo-posterior distribution,

(a) Utilize the series θ, derived from the weighted pseudo-posterior distribution as estimated previously, to randomly select $m = 1, \ldots, M$ parameter values.

(b) For each selected parameter value m and for each database record i in $1, \ldots, n$, generate a synthetic data value. In other words, this procedure effectively accomplishes a draw from the pseudo-posterior predictive distribution, given the posterior parameters.

(c) Subsequently, release the synthetic datasets $\mathcal{S} = (S^{(1)}, \ldots, S^{(M)})$.

2.3.2 Inferences

For inferences, we use the methods of Raghunathan et al. (2003), which we review here. We still denote the population quantity as Q, and $Q = Q(X, Y)$ means it depends on both X and Y. We use the q and v to represent the posterior mean and variance, respectively, of Q, based on the data sample $\mathcal{D}$.

Let $(q^{(m)}, v^{(m)})$: $m = 1, \ldots, M$ be the values of the q and v computed on the m synthetic dataset, and we consider $(q^{(m)}, v^{(m)} : m = 1, \ldots, M)$ as sufficient summaries of the synthetic databases $\mathcal{S}$ for the purpose of inference. According to Raghunathan et al. (2003), the

estimate of Q is,

$$\bar{q}_M = \sum_m \frac{q^{(m)}}{M}, \tag{2.22}$$

and, we write

$$T_M = (1 + M^{-1})b_{syn} - \bar{v}_M. \tag{2.23}$$

where $\bar{v}_M = \sum_m \frac{v^{(m)}}{M}$ and $b_{syn} = \sum_m \frac{(q^{(m)} - \bar{q}_M)}{M-1}$ as the approximate posterior variance. Reiter (2002) shows that if $T_M > 0$, and both n and M are large, the inferences for scalar Q can be based on normal distributions. For moderate M, inferences can be based on t-distributions with degrees of freedom that can be $M - 1$ approximately. Therefore, here, the $(1 - \alpha)\%$ confidence interval for Q can be expressed as,

$$\bar{q}_M \pm t_{M-1}(\alpha/2)\sqrt{T_M}. \tag{2.24}$$

If there are negative values in T_M, it is required to make the estimator of $Var(\bar{q}_m)$ to be positive and eligible to compute. Thus, according to Reiter (2002)'s instruction, (2.20) can be modified as below,

$$T_M^* = max(0, T_M) + \delta * (\frac{n_{syn}}{n}\bar{v}_M). \tag{2.25}$$

where $\delta = 1$ if $T_M < 0$, and $\delta = 0$ if $T_M \geqslant 0$. n_{syn} is the sample size of the synthetic data, and generally, $n_{syn} = n$.

3. Simulation

This section presents the results of simulation studies to illustrate the repeated sampling performances of the three methods.

3.1 Description

We create a population $\mathcal{P}$ of size $N = 1,000,000$ by first generating 1 million auxiliary variables, which we denote as X. All of these are generated from a Poisson distribution with a parameter of $\lambda = 5$, to which 1 is added to avoid 0 value. We then use a linear regression model to generate the survey variable, here we denote as Y, where

$$Y = \beta_0 + \beta_1 \times X + \varepsilon. \tag{3.1}$$

with $\beta_0 = 1$, $\beta_1 = 2$, and $\varepsilon \sim N(0,1)$. The population mean of X and Y are 6.001 and 13.001, respectively.

We use a probability proportional to size (PPS) sampling without replacement (ppswor) method to create the sample data $\mathcal{D}$ with size n, where we implement two different sizes, 1000 and 5000. The inclusion probabilities are proportional to the size variable, which is computed by the *inclusionprobabilities* function. Brus (2022) clearly presents the process of doing the ppswor method in chapter 8, and the R packages we utilized are the *sampling* and *survey* (Brus, 2022).

We take $K = 2000$ independent samples from the population, $\{\mathcal{D}_k : k = 1,\ldots,2000\}$. For each $\mathcal{D}_k$, we leverage three different methods presented above respectively to generate M synthetic datasets $\mathcal{S} = (S^{(1)},\ldots,S^{(M)})$. We select $M = 20$.

In any simulation run, the first two methods, Weighted Finite Population Bayesian Bootstrap (WFPBB) and Weighted Pseudo Population Bootstrap (WPPB) methods, require us to create pseudo-populations, $\mathcal{P}_{pseudo} = (\mathcal{P}^{(1)},\ldots,\mathcal{P}^{(M)})$ from $\mathcal{D}_k$. From each $\mathcal{P}_{pseudo}$, we create pseudo-samples $\mathcal{D}_{srs} = (\mathcal{D}_{srs}^{(1)},\ldots,\mathcal{D}_{srs}^{(M)})$ by taking a simple random sample directly from each $\mathcal{P}^{(m)}$, including X and Y in the sample, with a sample size of $n =$ 1000 or 5000 to match the sample size of $\mathcal{D}_k$. To make synthetic data for each $\mathcal{D}_{srs}^{(m)}$, we

begin with synthetic values of X. To do so, we begin by subtracting 1 from all x_i values and calculate the mean of the resulting values to obtain the estimated parameter $\hat{\lambda}$ for λ. We use this estimated $\hat{\lambda}$ as the parameter for a Poisson distribution, from which we draw n values. Finally, we add 1 to each Poisson draw to make the synthetic X values. To make the synthetic Y, we fit a linear regression of Y on X using $\mathcal{D}_{srs}^{(m)}$ to obtain maximum likelihood estimates, $\hat{\beta}_0^{(m)}$ and $\hat{\beta}_1^{(m)}$. In addition, we estimate

$$\sigma^{(m)} = \sqrt{\frac{\text{SSE}^{(m)}}{n-(1+p)}}. \tag{3.2}$$

where p is the number of predictor variables in the model and $\text{SSE}^{(m)} = \sum_{i=1}^{n}(y_i^{(m)} - \hat{y}_i^{(m)})^2$. We generate synthetic values of Y using the regression model with the estimated parameters and the synthetic X values.

The third method, Bayesian Pseudo Posterior, can allow us to generate the synthetic data $\mathcal{S} = (S^{(1)}, \ldots, S^{(M)})$ directly by having computed the estimated posterior model parameter θ. We use $X-1$ to compute the posterior λ like the above process. The Bayesian model implemented in Stan employs some prior distributions for the model parameters, specifically:

- The prior distributions for the linear regression coefficients β_1 and β_2 are assumed to be normal with mean 0 and variance 100, denoted as $\beta_1,\ \beta_2 \sim \mathcal{N}(0, 100)$.
- The standard deviation of the normal distribution, σ, is assumed to follow a Gamma distribution with shape parameter 2 and rate parameter 1, denoted as $\sigma \sim \text{Gamma}(2, 1)$.
- The parameter of the Poisson distribution, λ, is assumed to follow a Uniform distribution ranging from 0.1 to 100, denoted as $\lambda \sim \text{Uniform}(0.1, 100)$, because we assume λ is larger than 0.

For the parameter σ, we also try a uniform prior, $U(0.1,10)$. However, the result does not change much, since we have a large dataset, which dominates the posterior distribution of the parameter.

The model incorporates weighted log probability density functions for a normal dis-

tribution and log probability mass functions for a Poisson distribution to account for the observational weights:

$$\text{Lp} = \sum_{i=1}^{n} w_i \cdot \log \mathcal{N}(y_i | \beta_1 + \beta_2 x_i,\ \sigma), \tag{3.3}$$

$$\text{Lp}_p = \sum_{i=1}^{n} w_i \cdot \log \text{Poisson}(x_i - 1 | \lambda). \tag{3.4}$$

The objective function of the model, aimed to be maximized during the estimation process, consists of the sum of the weighted log probabilities of both the normal and Poisson distributions, thereby adjusting the model for the survey design factors and accommodating the potentially non-independent and identically distributed (non-i.i.d.) nature of survey data. The burn-in period is 2000, and then we run 6000 MCMC iterations to get the estimates of parameters.

3.2 Simulation Results

Figure 3.1 summarizes the results of the simulations for samples of sizes $n = 1000$ and $n = 5000$. The estimations $\bar{q}_M$ for the population mean across all three methodologies exhibit an absence of bias, thereby accurately reflecting the true population mean.

Figures 3.2 and 3.3 display the averages of the variance estimates and their empirical variances across the 2000 simulations. We hope that these two values are as close as possible. While each method yields similar empirical variances, a marked deviation becomes apparent when comparing the empirical variance with the averages of the variance estimates derived from the Weighted Pseudo Bayesian Bootstrap and Bayesian Pseudo Posterior methods. This discrepancy is notably consistent irrespective of the sample sizes, whether 1000 or 5000.

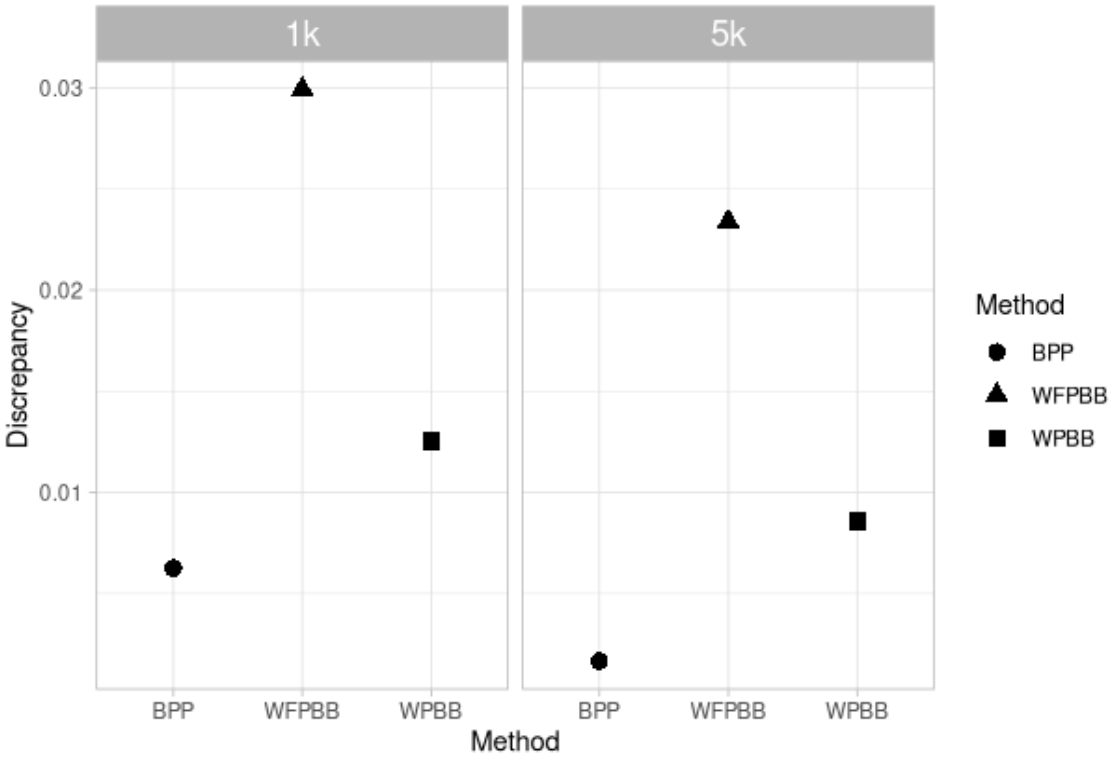

FIGURE 3.1: The discrepancy between the estimated mean and the true population mean across three methods. Results based on simulating from the true model for the data.

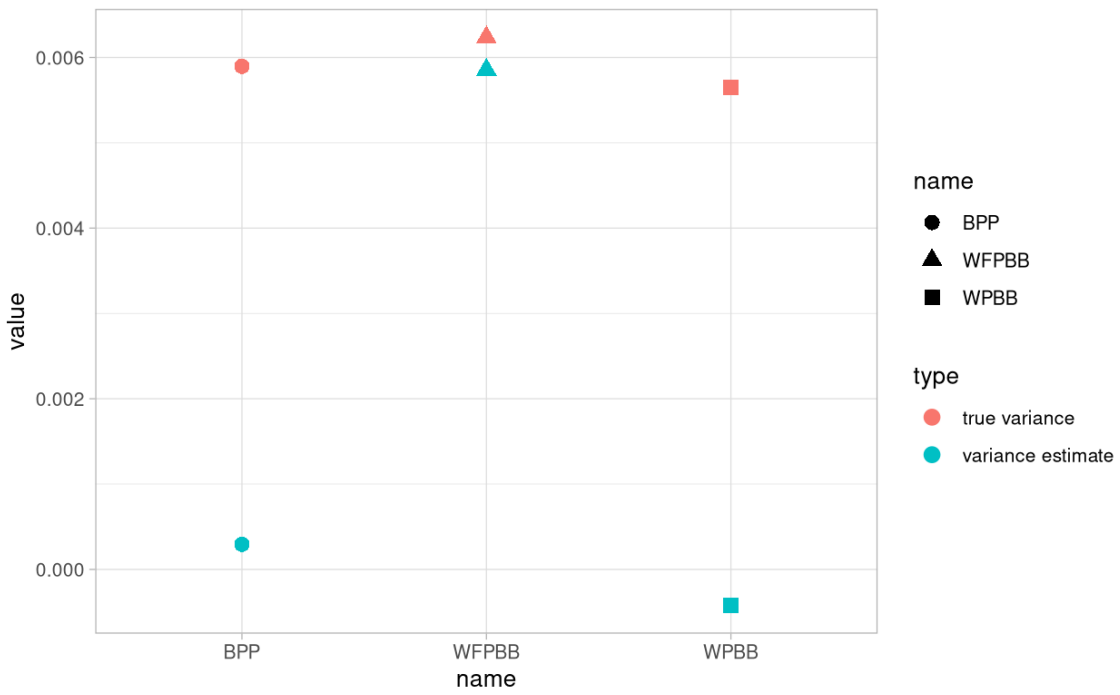

FIGURE 3.2: The discrepancy between the true variance and the estimated variance among three methods when sample size is 5000.

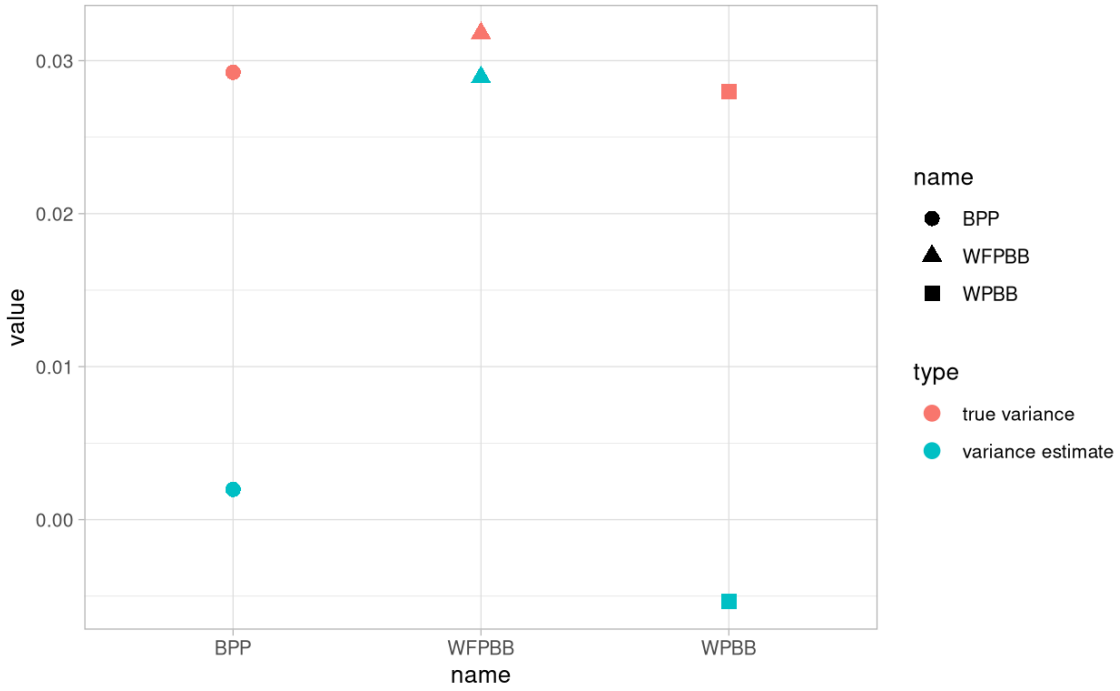

FIGURE 3.3: The discrepancy between the true variance and the estimated variance among three methods when sample size is 1000.

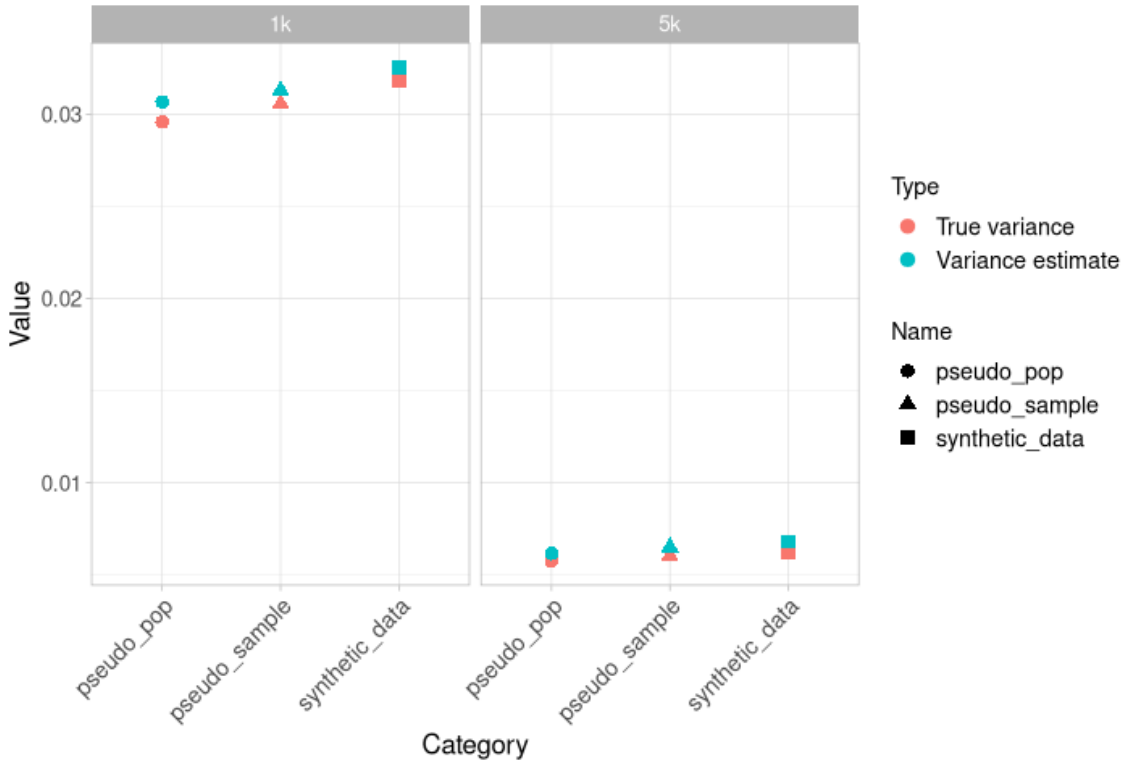

FIGURE 3.4: The discrepancy between the true variance and the estimated variance of WFPBB for the whole underlying process, with sample size of both 1000 and 5000.

The bad estimate for the variance of WPBB results from the process of creating the pseudo-population. Since it repeats the sample units in $\mathcal{D}$ for $\left\lfloor \pi_i^{-1} \right\rfloor$ times to generate the fixed part of the pseudo-population, which constructs the majority part of the pseudo-population and is the same across M pseudo-populations for each single survey sample, $\mathcal{D}$, the difference among the $\mathcal{P}^{(m)}$ is non-trivial, leading to low variability. The Bayesian Pseudo Posterior method, although it can have a similar estimate of the true model parameters, does not guarantee the estimate for variance will be accurate as given fixed posterior model parameters, the generated synthetic datasets in $\mathcal{S} = \left\{ S^{(1)}, \ldots, S^{(M)} \right\}$ will have insufficient variability.

Upon examining the results of the WFPBB for all procedures, we observe that the estimated variance aligns closely with the underlying true variance for each procedure, based

Table 3.1: Coverage comparison for three methods with a sample size of both 5000 and 1000. The proportion of negative variance estimates is in the parentheses.

Size \ Method	WFPBB	WPBB	BPP
5000	91.0 (6.4)	83.5 (51.0)	80.55 (49.2)
1000	90.7 (6.2)	83.6 (55.4)	79.95 (48.5)

on the Figure 3.4, which presents the discrepancy between the true variance and the estimated variance of WFPBB for the whole underlying process, and we can see the difference is ignorable. This indicates that the pseudo population $\mathcal{P}_{pseudo} = \left\{\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(M)}\right\}$, generated via WFPBB, exhibits sufficient variability across different pseudo population datasets. This level of variability is comparable to that found in the pseudo-simple random sample $\mathcal{D}_{srs} = \left\{\mathcal{D}_{srs}^{(1)}, \ldots, \mathcal{D}_{srs}^{(M)}\right\}$ and $\mathcal{S} = \left\{S^{(1)}, \ldots, S^{(M)}\right\}$. Additionally, as pseudo-populations and synthetic datasets are generated, the variance of the data samples increases due to the introduction of randomness with each new sampling. It is also evident that an increase in sample size corresponds to a decrease in variance.

As a check on the analysis results, we pretend we are data users who do not know the underlying true process for generating the data. Thus, we just use the X directly that was given to us to compute the estimate of the λ, which we denote as $\hat{\lambda}_{raw}$, and then we use this as the parameter of the Poisson distribution and draw synthetic X from it, without adding 1. The patterns closely match what is observed when using the actual distribution of X. It's evident that while all three methods can accurately estimate the true population mean, there is a clear distinction when estimating variance. In this aspect, the WFPBB method stands out, consistently delivering more satisfactory outcomes.

We next check the coverage rate of 95% multiple imputation confidence intervals. We force any negative variance estimates to be positive using (2.19) and (2.25). As evident in Table 3.1, all three methods have demonstrated coverage mostly at or above 80%. Nevertheless, after adjusting all the negative values for these three methods, WFPBB shows a

higher coverage rate compared to the other two, reaffirming its preference as a method.

4. Conclusion

In this work, we have discussed three different methods of generating synthetic data while incorporating the complex survey variable. The simulation study has illustrated that although all three methods can give an unbiased result of the population mean estimate, the weighted pseudo-population bootstrap method and the Bayesian pseudo-posterior method may not have a good estimate of the variance. Thus, we recommend that analysts utilize the WFPBB to generate the synthetic data when in a complex survey pattern. However, there is a limitation to our work, since we did not analyze the disclosure risk of these three methods.

Bibliography

Abowd, J., Stinson, M., & Benedetto, G. (2006). *Final report to the social security administration on the sipp/ssa/irs public use file project* (Technical Report). U.S. Census Bureau Longitudinal Employer-Household Dynamics Program.

Bhattacharya, A., Pati, D., & Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics*, *47*(1), 39–66.

Brus, D. (2022). *Spatial Sampling with R* (1st ed.). Chapman; Hall/CRC.

Chauvet, G. (2007). Méthodes de Bootstrap en population finie.

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Holmberg, A. (1998). A bootstrap approach to probability proportional to size sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 378–383.

Hu, J., Savitsky, T. D., & Williams, M. R. (2021). Risk-efficient Bayesian data synthesis for privacy protection. *Journal of Survey Statistics and Methodology*, *10*(5), 1370–1399.

Kim, H. J., Drechsler, J., & Thompson, K. J. (2020). Synthetic microdata for establishment surveys under informative sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(1), 255–281.

Kim, H., Reiter, J., & Karr, A. (2018). Simultaneous edit-imputation and disclosure limitation for business establishment data. *Journal of Applied Statistics*, *45*, 63–82.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, *79*, 363–384.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *9*, 407–426.

Mashreghi, Z., Haziza, D., & Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, *10*, 1–52.

Mathur, S., Si, Y., & Reiter, J. P. (2023). Fully synthetic data for complex surveys.

Murthy, M. N. (1970). *Sampling Theory and Methods*. Statistical Publishing Society.

Raab, G. M., Nowok, B., & Dibben, C. (2018). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, *7*(3), 67–97.

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

Reiter, J. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, *29*(2), 181–188.

Reiter, J. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *168*(1), 185–205.

Reiter, J. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, *18*, 531–544.

Reiter, J. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *International Statistical Review*, *77*, 179–195.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley Sons.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.

Savitsky, T. D., Williams, M. R., & Hu, J. (2022). Bayesian pseudo posterior mechanism under asymptotic differential privacy. *The Journal of Machine Learning Research*, *23*(1), 2484–2520.

United Nations Economic Commission for Europe. (2022). Synthetic data for national statistical organizations [Accessed: 2022-01-12].