Differentially Private Verification with Survey Weights

by

Tong Lin

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
David L. Banks

_____
Amy H. Herring

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

2023

ABSTRACT

Differentially Private Verification with Survey Weights

by

Tong Lin

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
David L. Banks

_____
Amy H. Herring

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

2023

Abstract

Survey sampling is a popular technique used in various fields for making inferences about populations from samples. However, the release of survey data can lead to confidentiality concerns due to the presence of sensitive information about individuals. To mitigate this issue, data stewards generate synthetic data that reflects the statistical features of confidential data to obscure sensitive variables. Synthetic data can be released for public use as a substitute of confidential data. However, the quality of synthetic data may impact the accuracy of inferences drawn from it. Therefore, assessing the quality of inferences derived from synthetic data is essential. Researchers have proposed a verification procedure that allows analysts to submit queries regarding their inferences and evaluate their accuracy by comparing results from synthetic data with those from confidential data. This approach enables the protection of individual privacy while facilitating the public use of confidential data.

This thesis proposes a differentially private verification measure for synthetic data in the context of complex survey designs. To ensure differential privacy, we use the sub-sample and aggregate method. We partition the confidential data into disjoint partitions and compute survey-weighted estimates of the statistics of interest. Analysts can set a tolerance interval reflecting their desired level of estimate accuracy from synthetic data. Since smaller partitions have higher variance in estimates, we suggest to use a wider tolerance interval for partitions. We refer to a tolerance interval that does not account for such higher variance as a fixed tolerance interval, while a tolerance interval with inflation as a varying one. We define an indicator to signify whether estimates from the partitions fall within the tolerance interval, and compute the sum of indicators from all partitions. To satisfy differential privacy, we add a noise from the Laplace Mechanism to this metric. Bayesian post-processing is

then applied to improve interpretability, and the summary statistics of the posterior distribution of the metric is released.

The proposed measure generalized the application of privacy-preserving techniques and enables analysts to validate the quality of their inferences based on synthetic data in the context of complex survey data sets.

Contents

List of Figures

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Jerome Reiter, whose guidance and support have been invaluable to me. He provided me with guidance on the research direction and technical details during my research, and taught me how to approach problems in a scientific manner. He also helped me a lot during my application. His kindness, patience, and insight have been instrumental in helping me grow as a researcher.

I would like to express my appreciation to my independent study advisor, Dr. David Banks, whose expertise, kindness, and detailed instruction made the project enjoyable and productive. He provided me with invaluable advice and support during the application process.

I would like to extend my sincere thanks to my advisor, Dr. Amy Herring, when I was admitted as a Master's student for enthusiastically answering my questions and helping me settle into the learning environment. Special thanks to Dr. Fan Li for guiding me in the field of spatial and causal inference, which I plan to continue exploring in my future studies.

I am also grateful to all the faculty members in the Department of Statistical Science for their enthusiastic and professional teaching. I would also like to thank all the staff in the department for creating a warm and supportive community, making my time at Duke University an unforgettable experience. Many thanks to Duke University for providing me with resources that have enabled me to spend two wonderful years in pursuit of my academic goals.

Finally, I want to express my appreciation to my family and friends for their unwavering support and encouragement throughout my studies, especially during the challenging times when I switched majors.

Chapter 1

Introduction

Survey sampling is widely used in various fields, especially social science, to make inferences about a population from a sample. By selecting a representative sample of the population, researchers can draw conclusions that apply to the larger population. Survey sampling is often used to estimate the summary statistics of a population, such as total, mean, variance, etc. Often, surveys are sampled with complex designs, such as stratified, probability proportional to size, or cluster sampling. These designs create unequal probabilities of selection of individuals into the sample. In order to obtain unbiased estimates of population parameters, researchers use sampling weights in survey to adjust for differences in the probability of selection of individuals in the sample.

However, publishing the survey data may raise confidentiality concerns. We refer the survey data as the confidential data, which include the actual information of individuals. In practice, data stewards cannot directly release confidential data since they often contain sensitive information. To protect data privacy, data stewards typically use measures such as aggregation, data swapping, and adding random noise. Nonetheless, these methods may not be effective, because they are still releasing confidential data which may lead to disclosure risk (Reiter, Oganian, and Karr, 2009; Matthews and Harel, 2011; Barrientos et al., 2018). Also, the statistics from the confidential data can threaten the privacy of someone not in the data set (Dwork, 2006).

As a means of obscuring sensitive information in data sets, Rubin (1993) pro-

posed the release of synthetic data intended for public use. This approach involves generating data that reflect the statistical characteristics of the confidential data. We define such generated data as the synthetic data. In practice, data stewards can release the synthetic data as a substitute for the confidential data to prevent the disclosure of sensitive information to unauthorized individuals. Although researchers can still conduct analysis on synthetic data, their inferences may be influenced by the synthetic generation process. For instance, if the synthetic data fails to accurately reflect the structure of the confidential data, it can lead to biased results (Reiter, 2005). Therefore, it is crucial to evaluate the accuracy of inferences from synthetic data to ensure its consistency with confidential data. For example, Reiter, Oganian, and Karr (2009) put forward a verification procedure that compares inferences from the synthetic data with those obtained from the confidential data, and determine whether there are significant differences between the two sets of inferences.

Although existing verification methods include procedures for confidential data obtained from simple random sampling or census, they do not account for the case where confidential data are obtained from a complex sampling design (Reiter, Oganian, and Karr, 2009; McClure and Reiter, 2012; Barrientos et al. 2018). In this thesis, we propose a differentially private verification measure for synthetic data when the confidential data come from a complex survey design. The proposed measure aims to generalize the application of privacy-preserving techniques for verification servers to complex survey data sets.

The remainder of this thesis is organized as follows. In Chapter 2, we review the complex sampling design, the generation of synthetic data, and concepts of differential privacy. We also review survey weights in the context of making inference about the population. In Chapter 3, we illustrate the procedure of the sub-sample and aggregate algorithm when using weighted estimators. We mainly discuss the context of estimating the population total, and provide examples and implementation details.

We also discuss the settings of several important parameters that may affect the output of the algorithm. In Chapter 4, we use simulation experiments to demonstrate the application of the algorithm in estimation of the population total. We present the analysis of the results and sanity check for survey weights in the simulation attempts. Finally, in Chapter 5, we summarize the main findings and propose directions for future research.

Chapter 2

Review of Synthetic data and Differential Privacy

This section provides background useful for understanding the verification measures that are developed in this thesis.

2.1 Review of sampling design

Selecting a representative sample of the population of interest is an essential aspect of survey design. To obtain accurate inference for the population parameters, the researchers need to employ an appropriate sampling method. Therefore, it is crucial to learn about the difference between various sampling techniques and sampling design.

In general, sampling methods can be classified into two categories, namely probability sampling and non-probability sampling (Omair, 2014; Tyrer and Heyman, 2016). Probability sampling takes the randomization strategy when choosing a sample from a finite population, wherein the selection is purely based on random selection, including simple random sampling, systematic sampling, stratified sampling, cluster sampling, and multi-stage sampling. Non-probability sampling techniques are characterized by an unknown likelihood of being selected (Acharya et al., 2013), including quota sampling, snowball sampling, and convenience sampling.

Here we focus on illustrating two probability sampling techniques. Let P be a finite population with N elements, with the index of elements denoted as $i = 1, 2, ..., N$. S is a subset of P which is comprised of n elements randomly drawn from

P. Suppose X is a survey variable in P. We are interested in making inference for the population total of X based on S, which we denote as $\tau = \sum_{i=1}^{N} x_i$. We define the indicator $I_i = 1$ if element $i \in S$, and $I_i = 0$ otherwise. Then we get a vector $I = (I_1, ..., I_N)$ which represents the elements in S, and $n = \sum_{i=1}^{N} I_i$.

In simple random sampling (SRS), each element in P has equal probability to be sampled. With $Pr(I_i = 1) = n/N$, we get an unbiased estimate of the population total $\hat{\tau} = \sum_{i \in S}(N/n)x_i$. The estimated variance of $\hat{\tau}$ is given by $N^2(1 - n/N)s^2/n$, where $s^2 = \sum_{i \in S}(x_i - \bar{x})^2/(n-1)$.

In practice, an alternative strategy that researchers may use is probability-proportional-to-size (PPS) sampling, where elements in P are sampled with unequal probabilities based on their sizes, where the size is defined by a numerical variable Z known for all units in P. In this case, we need to assign corresponding unequal weights when conducting analysis (Lumley, 2004). The weighted estimators, which take into account sample weights, are typically unbiased when estimating their corresponding population quantities (Korn and Graubard, 1995). Given that various PPS sampling methods will result in different joint selection probabilities for pairs of units, here we consider fixed sample size design (Hanif and Brewer, 1980; Zheng and Little, 2005). Let $\pi_i = Pr(I_i = 1)$ be the first-order inclusion probability of the i-th element in P. In PPS sampling of n units, we have $\pi_i = nz_i / \sum_{i=1}^{N} z_i$. For any record i where this quantity exceeds 1, we set that record's $\pi_i = 1$. For the remaining records, we recompute the π_i based on the population sizes excluding the cases sampled with certainty.

For any probability sampling design, a common approach to estimate the population total is weighting the sample elements by the inverse of the inclusion probability, which was proposed by Horvitz and Thompson (1952). We have

$$\hat{\tau} = \sum_{i \in S} \frac{y_i}{\pi_i} = \sum_{i \in S} w_i y_i,$$

where w_i represents the survey weight of the i-th element, which is defined as $w_i = (\pi_i)^{-1}$.

2.2 Review of synthetic data

One approach to reduce the data disclosure risk is to generate synthetic data which reflect the correlation structure of the variables in the confidential data for public use. In practice, however, the use of synthetic data to protect sensitive information may pose a trade-off with the ability to accurately reflect the features of confidential data. This trade-off must be carefully considered in the synthetic process to ensure that we can still draw meaningful inferences from the synthetic data. Overall, there are two approaches to generating the synthetic data, namely fully and partial (Reiter and Raghunathan, 2007). Both approaches strike the balance between obscuring the sensitive information and maintaining the actual features of the confidential data (Raghunathan, 2021).

Rubin (1993) proposed to create synthetic microdata constructed using multiple imputation. The goal is to mitigate the risk of information disclosure and enables users to obtain valid statistical inference as well. The fully synthetic approach involves randomly sampling units from P for each synthetic data set, treating nonsampled units as missing data and handle this using multiple imputation (Rubin, 1987), and releasing multiple synthetic data sets (Rubin, 1993). Raghunathan et al. (2003) evaluated the use of this multiple imputation framework and put forward the idea of obtaining valid inferences by combining the point and variance estimates from multiple data sets. Reiter (2002) illustrated the validity of inferences obtained from the synthetic data sets in different sampling design. He also discussed the specification of the number and size of synthetic data sets.

Although conceptually the fully synthetic approach can lower the disclosure risk

by obscuring the actual information and benefits the users by ensuring valid inference, it can be difficult to generate the full synthetic data set, especially when the survey contains complex variables. Little (1993) raised a method which partially masks the high risk information. The masked data could be a subset of the rows or columns of the data matrix. The replacement of only selected with synthetic data simplifies the generation of synthetic data. Reiter (2003) put forward techniques for making inferences on partially synthetic data using the multiple imputation framework, but with varying rules for combining point and variance estimates. Abowd and Woodcock (2001) adopted the method of synthesizing only sensitive variables, and using actual survey data for non-sensitive variables. They applied the approach to longtitudinal linked data (Abowd and Woodcock, 2004). Furthermore, Little et al. (2004) raised the idea of synthesizing only key variables for sensitive units plus another subset of nonsensitive units. Reiter (2005) demonstrated the application of CART to generate partially synthetic data for sensitive variables and identifiers. Furthermore, he proposes strategies for altering survey weights in the context of simulating identifiers using partially synthetic data (Mitra and Reiter, 2006).

Abowd and Woodcock (2001, 2004) and Kinney et al. (2011) applied the sequential regression modeling to generate synthetic data. Suppose D is a confidential data set with size of $n \times p$. Let X_{ij} be the value of the j-th variable of the i-th individual, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$. We specify the joint distribution as

$$f(X_{i1}, X_{i2}, ..., X_{ip}) = f_1(X_{i1}) f_2(X_{i2}|X_{i1}) ... f_p(X_{ip}|X_{i1}, X_{i2}, ..., X_{i\,(p-1)}).$$

f_j represents the conditional distribution of X_j given $X_1, X_2..., X_{j-1}$. We specify each conditional distribution as a regression, then generate the synthetic data by sampling from the corresponding sequence of predictive distributions.

As synthetic data gains popularity as a means of limiting the disclosure risk, researchers have raised concerns about the potential impact of the generation method

on the accuracy of their inferences made merely from the synthetic data. Researchers then proposed an integrated system that enables researchers to validate their analysis results (Barrientos et al., 2018). The system involves three components - a synthetic data set, a verification server (Reiter, Oganian, and Karr, 2009), and means for approved users to access the confidential data (Barrientos et al., 2018). The analysts can assess the quality of their inference by submitting a query about their estimates to the verification server, then the verification procedure will give results. The verification process should be differentially private. In this thesis, we build on the verification measure proposed by Barrientos et al. (2018) and Yang (2022), which we review in the context of developing the new verification measures in Chapter 3.

2.3 Review of differential privacy

The concept of differential privacy (DP) was proposed by Dwork (2006). In general, DP provides a mathematical measure of privacy risk. It offers a strong guarantee of privacy protection by limiting the amount of information that can be inferred about an individual in a data set.

Assume that $\mathcal{A}$ is an algorithm which takes a data set D_1 as input. We denote the output of $\mathcal{A}$ as $\mathcal{A}(D_1) = o$. We then define a neighboring data set D_2, which has the same data size as D_1. D_1 and D_2 differ in one row with all other rows identical. In accordance with the description provided by Barrientos et al. (2018), we present the definition of ϵ-DP as follows.

Definition 1 ($\epsilon - DP$): An algorithm $\mathcal{A}$ gives ϵ-DP if for any neighboring data sets D_1 and D_2, and any output $o \in Range(\mathcal{A})$, it satisfies

$$Pr(\mathcal{A}(D_1) = o) \leq exp(\epsilon) Pr(\mathcal{A}(D_2) = o).$$

$\mathcal{A}$ satisfying ϵ-DP ensures that any user who obtains the result of $\mathcal{A}(D_1)$ cannot

guess the presence of an individual in D_1 with more than $|1 - 1/exp(\epsilon)|$ of confidence (Bkakria, A. et al., 2019). ϵ is defined as the privacy budget. It quantifies the similarity between the outputs of $\mathcal{A}$ being implemented over D_1 and D_2. Intuitively, smaller ϵ makes it more difficult for users to distinguish the data record that differs between D_1 and D_2, and thus guarantees a higher privacy level.

DP has three important properties. Suppose that $\mathcal{A}_1$ and $\mathcal{A}_2$ are algorithms that satisfy ϵ_1-DP and ϵ_2-DP, respectively. First, for any data set D, releasing the outputs $\mathcal{A}_1(D)$ and $\mathcal{A}_2(D)$ guarantees $(\epsilon_1 + \epsilon_2)$-DP. Second, for any D_1 and D_2 where $D_1 \cap D_2 = \varnothing$, releasing the outputs of $\mathcal{A}_1(D)$ and $\mathcal{A}_2(D)$ guarantees $\max(\epsilon_1,\ \epsilon_2)$-DP. Third, for any algorithm $\mathcal{A}_3$, releasing the output $\mathcal{A}_3(\mathcal{A}_1(D))$ ensures ϵ_1-DP (Barrientos et al., 2018).

A frequently utilized approach to achieve ϵ-DP is the Laplace Mechanism (Dwork, 2006). Laplace Mechanism ensures ϵ-DP by adding a random perturbation. For any function $f : D \to \mathbb{R}^d$, the global sensitivity is defined as $\Delta(f) = max_{(D_1,D_2)} \| f(D_1) - f(D_2) \|_1$ for neighboring data sets D_1 and D_2. Applying the Laplace Mechanism, we have LM(D) = f(D) + η, where $\eta \sim Laplace(0,\ \Delta(f)/\epsilon)$.

However, sometimes the sensitivity of function f can be very high, resulting in a significant perturbation being added to the output. This may lead to reduced accuracy. This issue can be addressed by reducing the sensitivity of f. Dwork et al. (2006) proved that if f can be effectively approximated from random samples on all inputs, then the global sensitivity of f will be low. Then f can be published with a small amount of noise. Similarly, Nissim et al. (2007) put forward the sub-sample and aggregate framework which uses a 'smoothed' version of f to reduce the Laplace noise. The basic idea is to randomly partition data set D into M disjoint partitions $D^{'} = \{D_1, ..., D_M\}$. For each partition D_k, we calculate $f(D_k)$. Then $f_{avg}(D^{'}) = \sum_{k=1}^{M} f(D_k)/M$. In this case, for any neighboring data set which differs from D by only one record, it can change the value of at most one $f(D_k)$. We

then conclude that $\Delta(f_{avg}) = \Delta(f)/M$. If we apply the Laplace Mechanism to this smoothed $f_{avg}(D^{'})$, the noise satisfies $\eta_{new} \sim Lapalce(0, \Delta(f)/\epsilon M)$. This framework reduces the variance of the noise significantly (Yang, 2022). We use the sub-sample and aggregate method to develop the differentially private verification measures.

Chapter 3

Framework of the Differentially Private Verification Algorithm with Survey Weights

Suppose P is a finite population with N individuals. Suppose that the data agency takes a probability sample with unequal weights of n individuals from P and has labeled it as D. Hence, D is a confidential data set consisting of measurements on p variables $X_1, ..., X_p$ and a survey weight variable W. The survey weight of the i-th individual in D is defined as $w_i = 1/\pi_i$, where π_i is the first-order inclusion probability of individual i. To reduce the risk of data disclosure, the agency generates a synthetic data set for public use. The synthetic data set D_0 with size of n_0 involves measurements of p variables $X_1, ..., X_p$. We assume that D_0 is treated as a simple random sample from P, so that individuals in D_0 have the equal weight of N/n_0. We refer the researchers conducting analysis of the synthetic data as the synthetic data analysts.

Let X_1 be the variable of interest. The synthetic data analyst intends to estimate the population total of X_1 based on D_0. For example, X_1 could be a measure of the annual income of individuals from P. We define the true population total of X_1 in P as τ. $\hat{\tau}_0$ and $\hat{\sigma}(\hat{\tau}_0)$ represent the estimated value of τ and its standard error derived from D_0, respectively. We then introduce the framework of the verification algorithm.

3.1 Framework of the algorithm

Let $\hat{\tau}$ be an unbiased estimate of τ computed with the confidential data D. The synthetic data analyst cannot compute $\hat{\tau}$, since they do not have access to D. However, we define it for use in the verification algorithm. Basically, we assess the quality of the analyst's inference by approximating the distance between $\hat{\tau}_0$ and $\hat{\tau}$. When the distance between these two estimates falls within a range acceptable to the synthetic data analyst, $\hat{\tau}_0$ is considered to accurately reflect the features of the population. Otherwise, the estimate $\hat{\tau}_0$ fails to pass the verification, and D_0 is deemed not high enough quality for purposes of estimating τ.

We define the difference between $\hat{\tau}_0$ and $\hat{\tau}$ by the absolute distance $\hat{d} = |\hat{\tau}_0 - \hat{\tau}|$. If $\hat{d}$ falls within a specific interval, it suggests that $\hat{\tau}_0$ is reasonably accurate. The synthetic analyst could construct a tolerance interval based on $\hat{\tau}_0$ which we refer as $T(\hat{\tau}_0, \alpha)$. α is the parameter that determines the width of the tolerance interval. For example, $T = [\hat{\tau}_0 - 2\hat{\sigma}(\hat{\tau}_0),\ \hat{\tau}_0 + 2\hat{\sigma}(\hat{\tau}_0)]$ implies tolerance of two standard deviations, where the standard deviation derives from D_0. To satisfy DP, we cannot directly release the indicator of whether $\hat{\tau}$ falls in $T(\hat{\tau}_0, \alpha)$. Instead, we follow the sub-sample and aggregate method proposed by Nissim et al. (2007).

We randomly partition D into M disjoint partitions, with each partition denoted as $D_k \in \{D_1, ..., D_M\}$. The data size of D_k is $n_k = \lfloor n/M \rfloor$. Given that n might not be divisible by M, some partitions might have one more or one less unit than others. We need to take into account the impact of changes in sample size on the population estimate. Therefore, we adjust for such impact by inflating n/n_k times the weight of each unit in D_k. The inflation factor is approximately equal to M. For $k \in \{1, 2, ..., M\}$, we calculate the value of the Horvitz-Thompson estimator of the population total with data in D_k and the inflated weights and refer it as $\hat{\tau}_k$.

We define the tolerance interval used for the partitions as $C(\hat{\tau}_0, \alpha, \gamma)$. $C(\hat{\tau}_0, \alpha, \gamma)$

is not necessarily the same as $T(\hat{\tau}_0, \alpha)$. Let A_k be an indicator of the event that $\hat{\tau}_k$ falls in $C(\hat{\tau}_0, \alpha, \gamma)$. In other words, $A_k = \mathbb{I}(\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma))$. Iterating over all M partitions, we get M binary values $A_1, ..., A_M$. We discuss defining the tolerance intervals in Chapter 3.1.

S is the frequency of estimates falling within the tolerance interval in the partitions. Hence, $S = \sum_{k=1}^{M} A_k$ is a discrete variable that can take values in the set $\{0, 1, 2, ..., M\}$ with strictly positive probability. This makes S/M a variable taking discrete values on support of $[0, 1]$, representing the estimated probability that $\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma)$. Values of S near M indicate that the observed data estimates in the partitions frequently fall inside the tolerance intervals, which suggests the synthetic data results are similar to those from the confidential data as determined by the analyst's tolerance level. Values of S near 0 indicate that the synthetic data results are dissimilar from the confidential data results, suggesting the synthetic data results are not sufficiently accurate for the analyst's purposes.

To meet the ϵ-DP requirement, we need to add a perturbation to S. Here we apply the Laplace Mechanism. We randomly draw a sample $\eta \sim \mathit{Laplace}(0, 1/\epsilon)$ as the noise term. We assume that the change in one data record will only affect one specific weight and the value of at most one A_k. This guarantees a global sensitivity of 1. Accordingly, $S^R = S + \eta$ is the perturbed frequency variable.

Because S^R/M can be outside $[0, 1]$, we apply post-processing to improve the interpretability of the reported verification measure. Specifically, we apply a Bayesian approach to S^R/M to provide the synthetic analyst with the posterior distribution of S^R/M. A_k is a random variable which follows Bernoulli distribution. We assume that $A_1, A_2, ..., A_M \overset{\text{iid}}{\sim} \mathit{Bernoulli}(r)$, where r signifies the probability of any randomly generated $\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma)$. Hence, S is a random sample drawn from $S|r \sim \mathit{Binomial}(M, r)$. Let ψ be the prior that we specify for r. We have

$$S^R|S \sim Laplace(S,\, \frac{1}{\epsilon})$$

$$S|r \sim Binomial(M,\, r)$$

$$r \sim \psi.$$

In the simulation experiments discussed in Chapter 5, we assign a $Beta(1, 1)$ distribution as the prior of r.

We obtain the posterior distribution $Pr(r|S^R)$, i.e., $Pr(\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma)|S^R)$, by applying Gibbs sampler. The joint distribution is

$$Pr(r,\, S,\, S^R) \propto Pr(S^R|S,\, r)Pr(S|r)Pr(r).$$

Then we derive the full conditionals

$$\begin{aligned} Pr(r|S,\, S^R) &\propto Pr(S|r)Pr(r) \\ &\propto r^S(1-r)^{M-S} \\ &\propto Beta(S+1,\, M-S+1) \end{aligned}$$

$$\begin{aligned} Pr(S|r,\, S^R) &\propto Pr(S^R|S)Pr(S|r) \\ &\propto e^{-\frac{|S^R-S|}{1/\epsilon}} \frac{1}{\Gamma(S+1)\Gamma(M-S+1)} r^S(1-r)^{M-S}. \end{aligned}$$

The data agency can release the summary statistics, for instance, the posterior median or posterior mean of r, to provide an interpretable assessment about the quality of the synthetic analyst's inference.

The algorithm format of the verification procedure is in Algorithm 1.

We now turn to discuss the factors that may affect the outcome of the sub-sample and aggregate procedure, including the tolerance interval, total number of partitions of confidential data, and privacy budget. We assume that the choice of privacy budget (ϵ) is based on preference. We only discuss the choice of other aspects.

Algorithm 1: Subsample and aggregate for data with weights

Data: D, D_0

D represents the confidential data, with data size denoted by n, and the weights are stored in variable W. We have values $(x_1, \ldots, x_n)$ in D.

Result: $r_{Gibbs,noise} \leftarrow Pr(\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma)|S^R)$

$\tau \leftarrow \hat{\tau}$; `/* Estimated population total from` D `*/`

$\tau_0 \leftarrow \hat{\tau}_0$; `/* Estimated population total from` D_0 `*/`

$C(\hat{\tau}_0, \alpha, \gamma)$; `/* Tolerance interval */`

Randomly partition D into M disjoint subgroups $(D_1, \ldots, D_M)$.

for $k \leftarrow 1$ *to* M **do**

$D_k \leftarrow$ The k-th partition of D of size $n_k = \lfloor \frac{n}{M} \rfloor$;

$W_k \leftarrow W \times \frac{n}{n_k}$;

$\tau_k \leftarrow \hat{\tau}_k = \sum_{i \in D_k} w_i x_i$; `/* Horvitz-Thompson estimator */`

if $\tau_k \in C(\hat{\tau}_0, \alpha, \gamma)$ **then**

$A_k \leftarrow 1$;

else

$A_k \leftarrow 0$;

end

$S \leftarrow \sum_{k=1}^{M} A_k$;

$S^R \leftarrow S + Laplace(0, \frac{1}{\epsilon})$;

$r_{noise} \leftarrow \frac{S^R}{M}$;

$r_{Gibbs,noise} \leftarrow Pr(\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma)|S^R)$;

end

3.2 Defining the tolerance interval

The synthetic data analyst is required to specify a tolerance interval which reflects their tolerance for the accuracy of the inference. They are free to choose a threshold

as the boundary of the tolerance region. For example, suppose the variable of interest takes binary values. The synthetic data analyst would like to estimate the total of this variable. They may determine that $\hat{\tau}_0$ is accurate enough as long as $\hat{\tau}$ is within some percentage of $\hat{\tau}_0$, i.e., $T(\hat{\tau}_0, \alpha) = [\hat{\tau}_0 \pm \alpha|\hat{\tau}_0|]$. The analyst also need to specify a tolerance interval for the sub-sample and aggregate method, which we denote as $C(\hat{\tau}_0, \alpha, \gamma)$. Here, γ plays the role of an inflation factor which may be used to go from $T(\hat{\tau}_0, \alpha)$ to $C(\hat{\tau}_0, \alpha, \gamma)$. We consider the case $T(\hat{\tau}_0, \alpha) = C(\hat{\tau}_0, \alpha, \gamma)$ to represent a fixed tolerance interval, while $T(\hat{\tau}_0, \alpha) \neq C(\hat{\tau}_0, \alpha, \gamma)$ signifies a varying tolerance interval.

We use an example to illustrate the choice of tolerance interval. Suppose $\hat{\tau}_0 = 100000$ and $\hat{\sigma}(\hat{\tau}_0) = 1000$. The analyst wants to know if $\hat{\tau}$ falls within 10% of $\hat{\tau}_0$. We know that $10\%\hat{\tau}_0 = 10000$. For a fixed tolerance interval, we have $T(\hat{\tau}_0, \alpha) = C(\hat{\tau}_0, \alpha, \gamma) = [90000, 110000]$. If the analyst decides to use a varying tolerance interval, we suggest to inflate $C(\hat{\tau}_0, \alpha, \gamma)$. The reason is that a smaller sample size will enlarge the variance of the estimate of the population total. In this case, the estimates in the partitions are less likely to fall into the unadjusted $C(\hat{\tau}_0, \alpha, \gamma)$ even when $\hat{\tau}$ is inside $T(\hat{\tau}_0, \alpha)$. For ease of interpretation, we construct the tolerance interval based on the estimate $\hat{\tau}_0$ and its standard error $\hat{\sigma}(\hat{\tau}_0)$. In our setting, $10\%\hat{\tau}_0 = 10\hat{\sigma}(\hat{\tau}_0)$. Therefore, for $T(\hat{\tau}_0, \alpha)$, instead of $[\hat{\tau}_0 \pm 10\%|\hat{\tau}_0|]$, we suggest to define it as $T(\hat{\tau}_0, \alpha) = [\hat{\tau}_0 \pm 10\hat{\sigma}(\hat{\tau}_0)]$. Suppose we have $M = 25$ disjoint partitions. As is proposed in Barrientos et al. (2018), we approximate $\hat{\sigma}(\hat{\tau}_k)$ by $\hat{\sigma}(\hat{\tau}_k) = \sqrt{n/n_k}\hat{\sigma}(\hat{\tau}_0)$. We use this inflated standard error when constructing $C(\hat{\tau}_0, \alpha, \gamma)$. In other words, we inflate the width of the tolerance interval to reflect the increased standard error in the estimates computed with partitioned data. Thus, $C(\hat{\tau}_0, \alpha, \gamma) = [\hat{\tau}_0 \pm 10 \cdot 5 \cdot \hat{\sigma}(\hat{\tau}_0)]$. 10 and 5 are specific values of α and γ, respectively.

Basically, α and γ are parameters that decide the width of the interval. The analysts could choose α based on their preference and tolerance for accuracy. We

set $\gamma = 1$ in the fixed tolerance intervals. While for the varying intervals, we set $\gamma = \sqrt{M}$ by default.

3.3 Choosing M

In the sub-sample and aggregate algorithm, we apply the Laplace Mechanism to S. We have $S/M + \eta/M = S^R/M$, where $\eta \sim Laplace(0, 1/\epsilon)$. In this section, we discuss the choice of number of partitions M. We mainly consider the effect of the variation in M on S/M itself and the noise from the Laplace Mechanism.

We know that $S/M \in [0, 1]$, and S can take $M + 1$ values. If M is small, S/M will only take a limited grid of values. For instance, when $M = 10$, we have $S/M \in \{0, 0.1, 0.2, ..., 1\}$. In this case, S/M provides a rough estimate of r that might not be detailed enough for the analyst. In addition, with a small M, the perturbation from the Laplace Mechanism will have a greater impact on S/M. From the perspective of sample size, for a certain D, less partitions means larger sample size of each partition. Large value of n_k reduces the variance of the estimate for each partition. Meanwhile, a small M will lead to increase in the variance of S/M. On the contrary, large value of M reduces the uncertainty in S/M and mitigates the impact of the perturbation.

According to the above discussion, the synthetic data analyst needs to consider such trade-off when choosing the value of M. Basically, we tend to choose an M, such that the estimation results obtained in the partitions are consistent with the results estimated using the full confidential data set, which requires $Pr(\hat{\tau} \in T(\hat{\tau}_0, \alpha))$ to be close to $Pr(\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma))$. Specifically, if $\hat{\tau}$ is inside $T(\hat{\tau}_0, \alpha)$, the maximum value of the probability density of S^R/M should be close to 1. While when $\hat{\tau}$ is outside $T(\hat{\tau}_0, \alpha)$, the maximum value of the probability density of S^R/M should be around 0. We hope to be able to select M by comparing the results obtained under different

M values.

Chapter 4

Simulation Study

In this section, we conduct simulation experiments to illustrate the implementation of the verification algorithm.

We first generate a population data set P which contains $N = 10000000$ data records. We generate two variables X and Y sampled from the following distribution

$$X \sim Uniform(0,\, 10)$$

$$Y|X \sim \mathcal{N}(X + 5,\, 2).$$

Here 2 represents the variance of $Y|X$. For each unit in P, we assign an inclusion probability which is proportional to the size based on X. Suppose the confidential data D comprises n individuals from P, then the i-th individual has a probability of $\pi_i = nx_i/\sum_{i=1}^{N} x_i$ to be sampled. We take a probability-proportional-to-size sample from P to make the confidential data D. The survey weight of the i-th individual in D is generated by the inverse of the inclusion probability, that is, $w_i = 1/\pi_i$.

To generate the synthetic data, we employ two distinct methods to provide a comprehensive illustration of the role of weights in our settings.

The first method involves generating a representative synthetic data set. To do so, we need to account for the complex design when synthesizing from D. Failure to do so can result in synthetic data that do not look like P. However, our goal is to evaluate the verification measures rather than implement a synthesizer that turns a survey-weighted sample into a representative sample. Therefore, to simplify the synthesis task, we take advantage of the fact that we are in a simulation setting and simply take a simple random sample of size n_0 from P to create D_0. Ideally, the

verification measure should reveal that the synthetic data provide accurate estimates, since we know D_0 is representative of P. Of course, this is not possible in genuine applications; data stewards need to account for the complex design when using D to make D_0.

In the second method, we generate D_0 by ignoring the weights. Specifically, we randomly draw n_0 samples from the Normal distribution $\mathcal{N}(\bar{y}_c, s_c^2)$, where $\bar{y}_c$ and s_c^2 are the mean and standard deviation of the variable Y in D. We use this synthesizer to examine the performance of the verification measure when the synthetic data are not an accurate representation of P. Ideally, the verification measure should reveal this inaccuracy.

In the implementation of the sub-sample and aggregate algorithm, we focus on factors that could affect the result of the algorithm, which are data size of the partitions, number of partitions, and parameters that define the tolerance interval. The sample size of each partition is $n_k \in \{500, 20000, 50000\}$, and we have $M \in \{25, 50, 90\}$ partitions. For each combination of n_k and M, we draw $n_k \times M$ samples from P following the methods discussed above to make D. Here we set $n_0 = n$. We repeat this step for 200 times and generate 200 pairs of D_0 and D for each of the two synthetic generation methods.

Next we define the tolerance interval. In Chapter 3.1, we construct the interval based on the estimated population total $\hat{\tau}_0$ and its standard error $\hat{\sigma}(\hat{\tau}_0)$ from D_0. We specify a tolerance interval for the full confidential data set and another for the partitions, namely $T(\hat{\tau}_0, \alpha)$ and $C(\hat{\tau}_0, \alpha, \gamma)$. We assume that $\alpha \in \{1, 3, 5\}$. We consider both fixed tolerance interval and the varying tolerance interval. For the fixed interval, we set $\gamma = 1$, which indicates $T(\hat{\tau}_0, \alpha) = C(\hat{\tau}_0, \alpha, \gamma) = [\hat{\tau}_0 - \alpha\hat{\sigma}(\hat{\tau}_0),\ \hat{\tau}_0 + \alpha\hat{\sigma}(\hat{\tau}_0)]$. For the varying interval, we set $\gamma = \sqrt{M}$ for $C(\hat{\tau}_0, \alpha, \gamma)$. Therefore, we have $C(\hat{\tau}_0, \alpha, \gamma) = [\hat{\tau}_0 - \alpha\sqrt{M}\hat{\sigma}(\hat{\tau}_0),\ \hat{\tau}_0 + \alpha\sqrt{M}\hat{\sigma}(\hat{\tau}_0)]$.

For each pair of D_0 and D, we use two metrics to assess $\hat{\tau}_0$: utilizing the full

confidential data set D, and utilizing the partitions $\{D_1, ..., D_M\}$. Notably, only the differentially private result obtained from the partitions will be released to the analyst. We use the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) to calculate the population total.

When using the full data set D, we define a binary variable Q, which is an indicator that takes value of 1 when $\hat{\tau}$ is inside the tolerance interval, i.e., $Q = \mathbb{I}(\hat{\tau} \in T(\hat{\tau}_0, \alpha))$. For the 200 pairs of D_0 and D, we get $Q_1, ..., Q_{200}$. We then calculate $r_{full} = \sum_{i=1}^{200} Q_i/200$, which is an approximate estimate of $Pr(\hat{\tau} \in T(\hat{\tau}_0, \alpha))$.

When using the partitions, we implement the sub-sample and aggregate procedure (Nissim et al., 2007). We inflate the survey weights of individuals in the partitions by M. Hence, the i-th individual in D_k has a weight of $w_i \times M$, where w_i is its original weight in D. In each D_k, we then estimate $\hat{\tau}_k$ using these adjusted weights, and we compute A_k, i.e., whether or not the $\hat{\tau}_k$ is inside $C(\hat{\tau}_0, \alpha, \gamma)$. We use Gibbs sampler to sample the posterior distribution of r. For each Gibbs sampler, we run 1200 iterations, with the first 200 as burn-in.

4.1 Results for synthesis based on SRS of P

We display the results for the utilization of fixed tolerance intervals in Figure 4.1. We notice obvious discrepancy between the value of r_{full} and posterior medians of r, which indicates inconsistency between the conclusions drawn from using the full data set and the partitions. The posterior medians of r are always much smaller than their corresponding r_{full}. When $\alpha = 1$, the value of r_{full} is about 0.2, while the majority of medians of $Pr(r|S^R)$ are lower than 0.1. When $\alpha = 3$, the value of r_{full} fluctuates around 0.6, while most of the posterior medians of r are smaller than 0.25. The most notable discrepancy is observed when $\alpha = 5$, where r_{full} exceeds 0.75, while the posterior medians of r are at least 0.5 lower than their corresponding r_{full}

values.

As expected, the increase of α leads to a higher probability that the estimate from D falls within the tolerance interval. With the same value of α, as M gets larger, the posterior medians of r become smaller, implying that the estimates are less likely to be inside the tolerance interval. Meanwhile, such decreasing trend turns more significant as α gets larger. When $\alpha = 1$, the posterior medians of r are slightly higher at $M = 25$ than $M = 90$. While when $\alpha = 5$, we can observe a more evident change between $M = 25$ and $M = 90$. Given both M and α fixed, smaller partition sample size corresponds to larger value of both r_{full} and medians of $Pr(r|S^R)$, indicating higher probability that $\hat{\tau}$ and $\hat{\tau}_k$ are within the analyst's tolerance. When n_k is small, we expect greater uncertainty in the estimate of population total in D_k. Accordingly, the larger standard error of the estimate makes a wider tolerance interval. Therefore, $\hat{\tau}$ and $\hat{\tau}_k$ are more likely to be within the interval. As α grows larger, we notice a more apparent difference between the result at $n_k = 500$ and at $n_k = 50000$. This is because the increase of α will amplify the change in the probability that $\hat{\tau}$ and $\hat{\tau}_k$ being inside the tolerance interval.

We then run simulations for the varying tolerance interval. The simulation results are displayed in Figure 4.2. We can observe an overlap between r_{full} and the posterior medians of r. In most instances, the values of r_{full} are within the range of the posterior medians of r. By inflating the tolerance interval for the sub-sample and aggregate version, the posterior medians of r are close to their corresponding r_{full}. When $\alpha = 1$, both r_{full} and posterior medians of r are around 0.3. When $\alpha = 3$, the value of r_{full} increases to between 0.5 and 0.9. The majority of the posterior medians of r are distributed within this scope as well. When $\alpha = 5$, r_{full} and posterior medians of r reach above 0.8.

Given all other parameters fixed, the increase of α corresponds to a higher probability that the estimate from the confidential data falls within the tolerance interval.

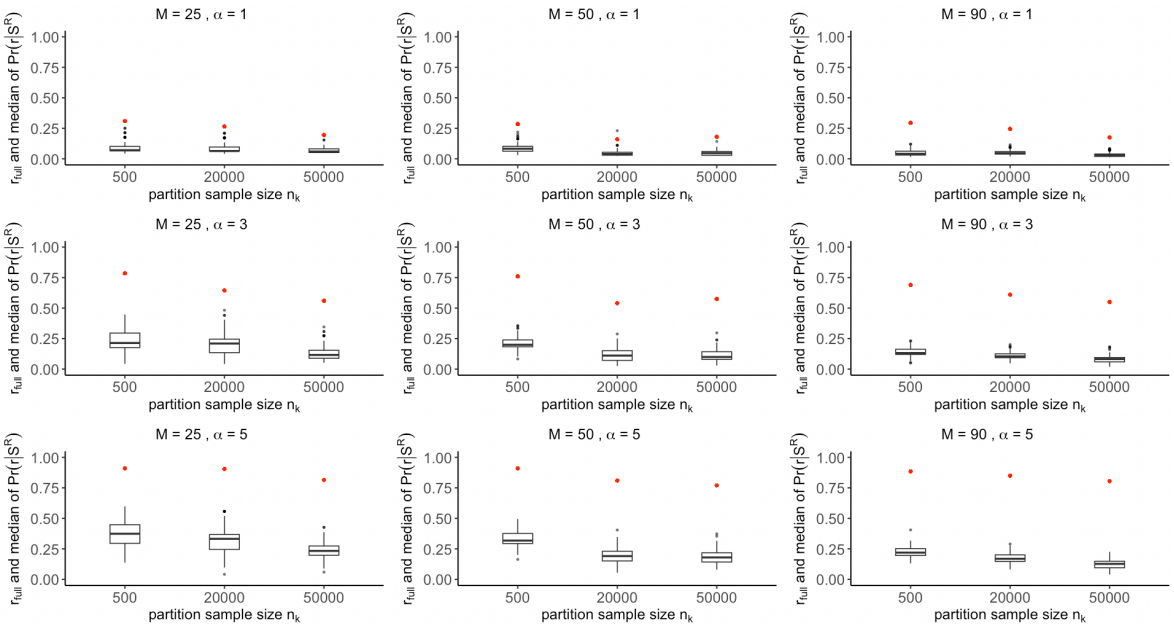

Figure 4.1: r_{full} (red points) and posterior medians of r (boxplots) for the fixed tolerance interval. Synthetic data are a SRS from P.

With the same n_k and α, changes in M have no appreciable effect on the average value of posterior medians of r. Nonetheless, we can observe an obvious decrease in the variance of the posterior medians of r as M grows larger. Larger M brings down the impact of the noise from the Laplace Mechanism on S/M, and thus reduces the variance in S^R/M and the uncertainty in the posterior. Similar to the case in fixed tolerance interval, given both M and α fixed, smaller partition sample sizes correspond to larger values of both r_{full} and medians of $Pr(r|S^R)$, indicating higher probability that $\hat{\tau}$ and $\hat{\tau}_k$ are within the analyst's tolerance interval.

In comparison, when using a fixed tolerance interval, the posterior medians of r maintain low values. Even if we set $\alpha = 5$, the posterior medians are lower than 0.5, while the values of r_{full} are higher than 0.8. With fixed tolerance intervals, the measurements from the sub-sample and aggregate method do not objectively reflect the estimate from the full confidential data set. Whereas, the varying tolerance

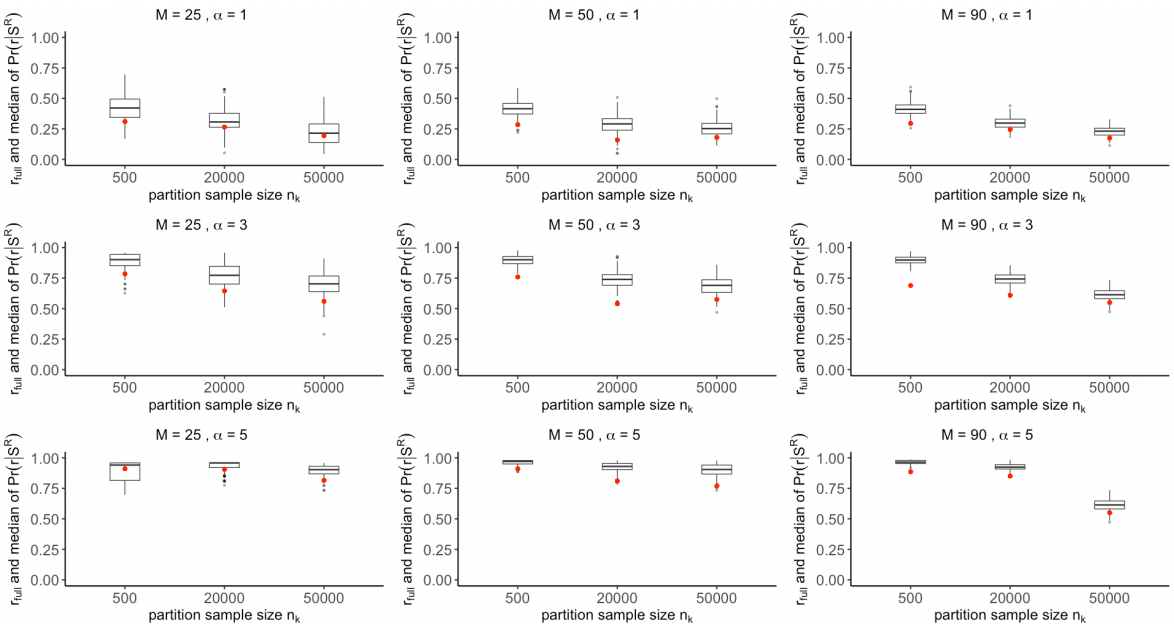

Figure 4.2: r_{full} (red points) and posterior medians of r (boxplots) for the varying tolerance interval. Synthetic data are a SRS from P.

intervals perform better in this verification procedure. The value of the posterior medians of r are consistent with their corresponding r_{full} when we inflate the intervals used for the partitions.

Figure 4.3 and Figure 4.4 demonstrate the difference between median of S/M and its corresponding median of $Pr(r|S^R)$. In general, given all other parameters fixed, larger M mitigates the impact of the perturbation from the Laplace Mechanism. Hence, $S/M - median(Pr(r|S^R))$ becomes less unstable and approaches 0. Overall, using the varying tolerance interval makes this value approaches 0. We expect the inflated interval to lead to a higher probability that the estimates fall into its region. Thereupon, larger value of S/M can obscure the noise from Laplace Mechanism, which lowers the variation in $S/M - median(Pr(r|S^R))$.

These results suggest that analysts choose a varying tolerance interval when submitting the query.

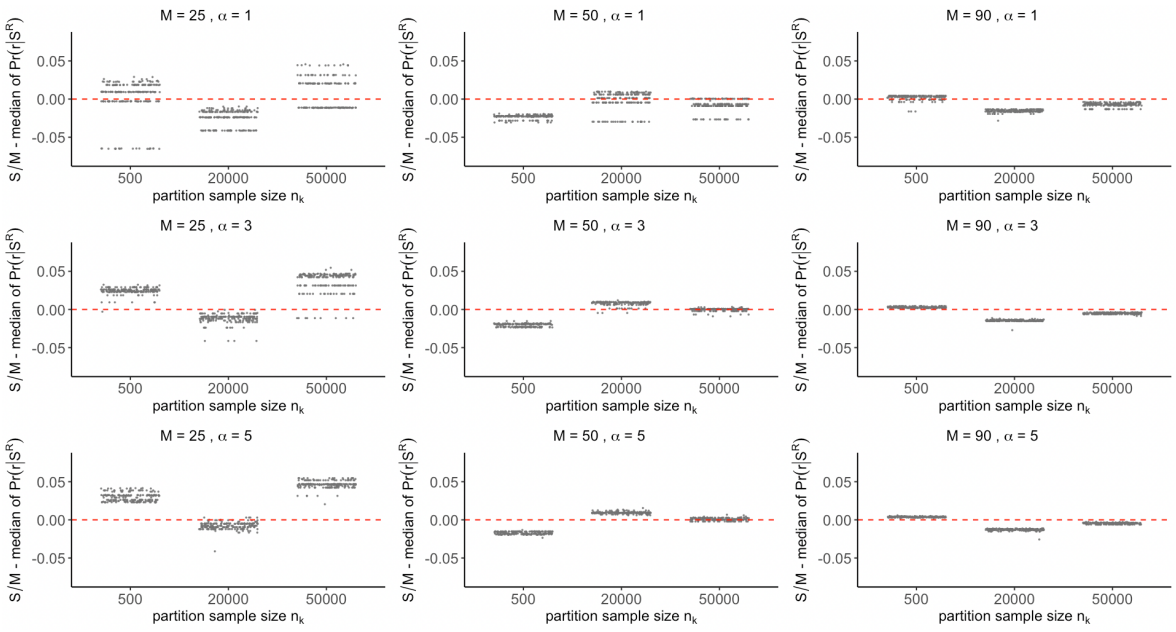

Figure 4.3: S/M - posterior median of r for fixed tolerance intervals. Synthetic data are a SRS from P.

4.2 Results for biased synthesis

We now turn to the results from the simulation where the data steward disregards the sampling design when generating D_0. As discussed previously, we expect these synthetic data to be low quality and desire the verification measures to indicate as such.

Before turning to the verification measures, we first provide evidence that accounting for the survey design is important in our simulation. Let $\tau = \sum_{i=1}^{N} y_i$ be the population total. To illustrate the influence of the survey weights, for each generated D, we estimate τ using both the Horvitz-Thompson estimator and an unweighted estimator. Specifically, we denote the estimated value of the Horvitz-Thompson estimator as $\hat{\tau} = \sum_{i \in D} w_i y_i$. Furthermore, we compute the unweighted estimator as $\hat{\tau}_{unwt} = N\bar{y}_c$, where $\bar{y}_c$ represents the mean of variable Y in D.

Figure 4.5 displays the comparison of the estimated population total between the

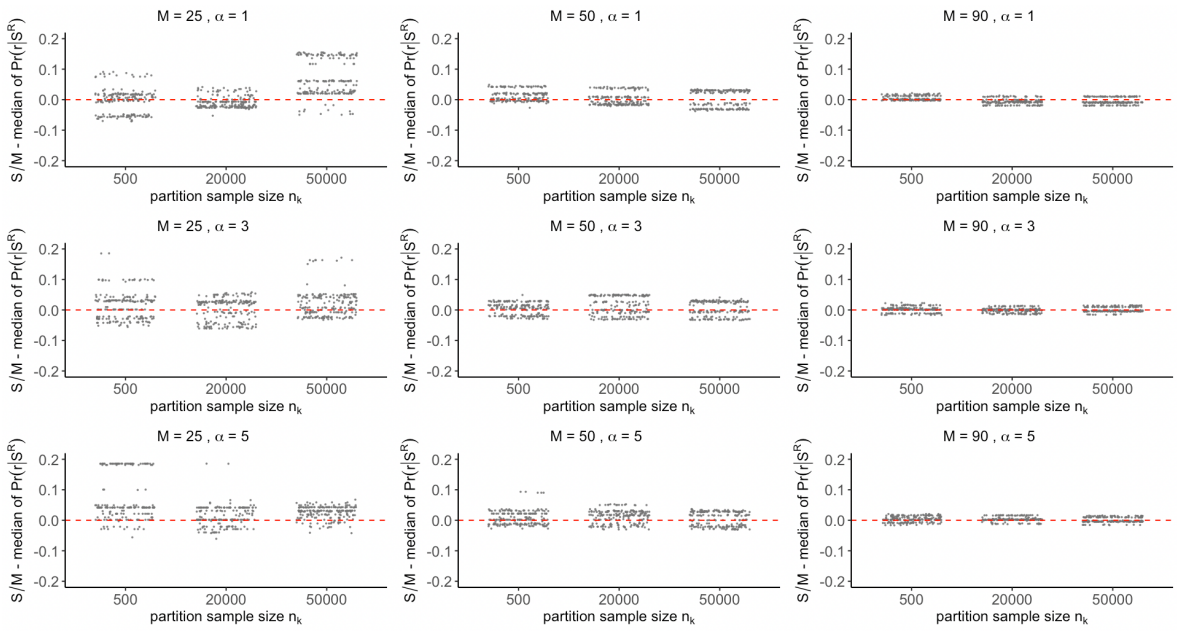

Figure 4.4: S/M - posterior median of r for varying tolerance intervals. Synthetic data are a SRS from P.

weighted and unweighted estimators when $M = 50$. The red dashed line represents the actual value of τ, which is 99984562. The black dots represent the population total estimated using the Horvitz-Thompson estimator ($\hat{\tau}$). Evidently, the weighted estimates cluster around the true value. The blue dots represent the results of the unweighted estimator $\hat{\tau}_{unwt}$. The unweighted estimates exhibit a significant deviation from the true value. Therefore, the Horvitz-Thompson method provide an unbiased estimate of the population total while the unweighted estimator that discards the PPS sampling does not.

Next, we proceed to implement the verification procedure using the biased synthetic data. Figure 4.6 and Figure 4.7 show the results for fixed tolerance intervals and varying tolerance intervals in this setting. Regardless of the value we set for n_k, M, and α, r_{full} and the posterior medians of r are close to 0. This observation is as expected, as the poorly generated synthetic data lead to extremely biased re-

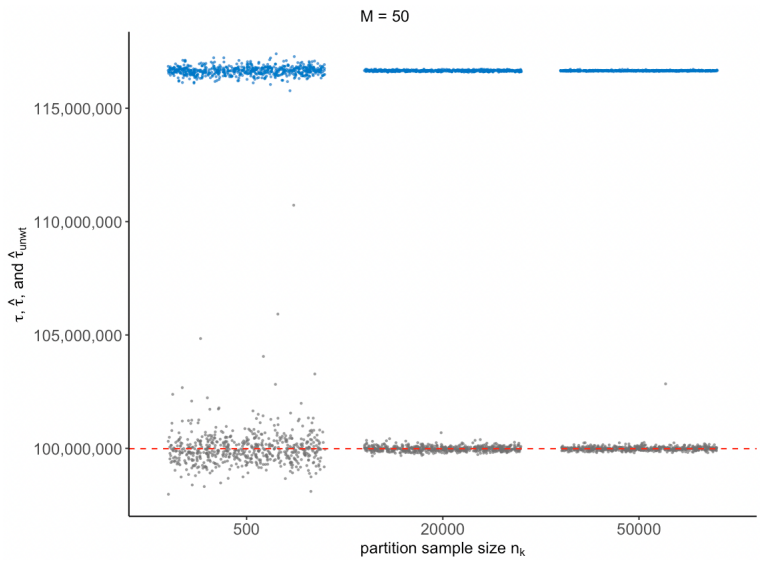

Figure 4.5: Estimated population total from confidential data. Horvitz-Thompson estimator (black points), and unweighted estimator (blue points).

sults, making it difficult for the estimates obtained from D to lie within the tolerance interval. Evidently, the verification measures readily reveal to the analyst that the synthetic data are highly inaccurate for estimating τ.

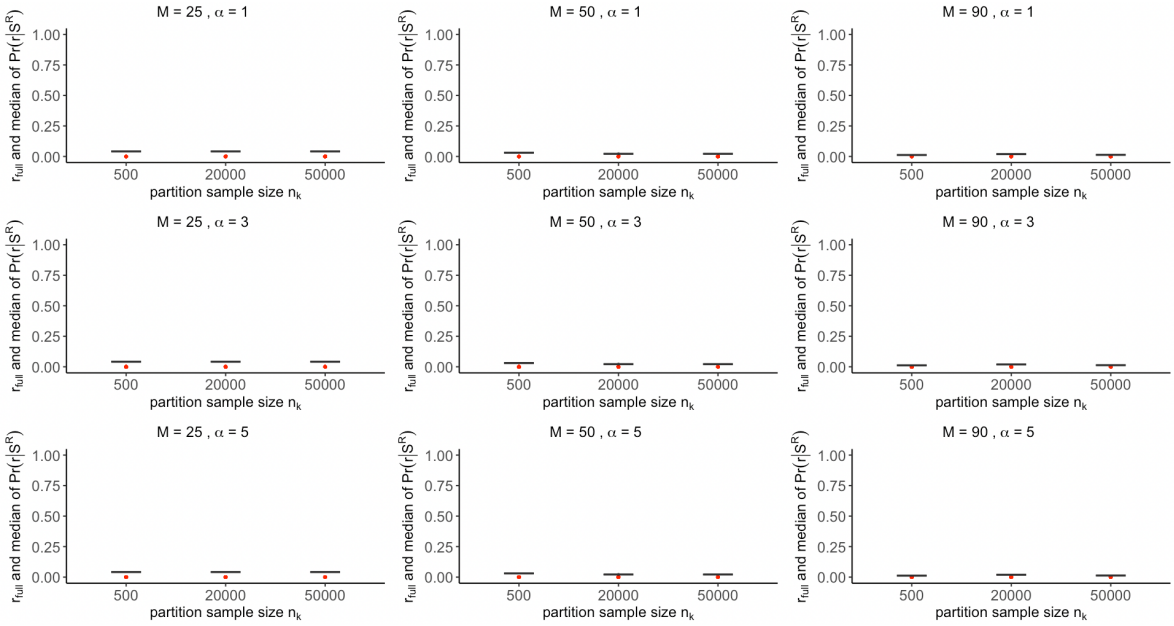

Figure 4.6: r_{full} (red points) and posterior medians of r (boxplots) for the fixed tolerance interval. Synthetic data are biased.

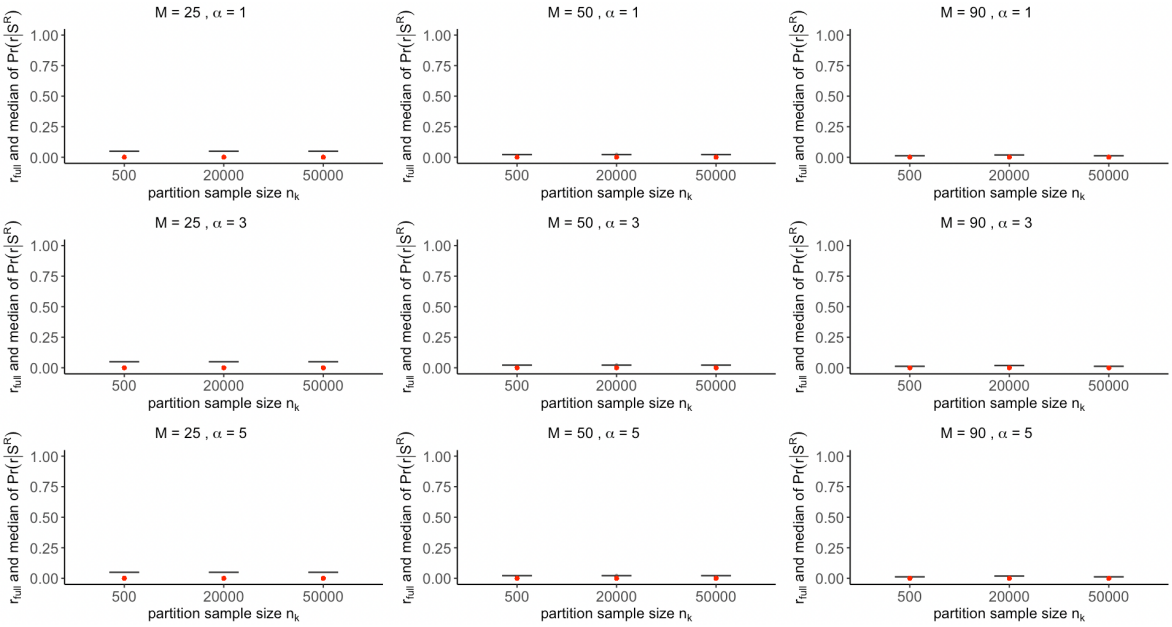

Figure 4.7: r_{full} (red points) and posterior medians of r (boxplots) for the varying tolerance interval. Synthetic data are biased.

Chapter 5

Conclusions

In this thesis, we address the gap in existing verification measures for synthetic data when the corresponding confidential data come from complex survey designs. Our approach employs the sub-sample and aggregate method in conjunction with the Laplace Mechanism, which ensures differential privacy of the output.

Our findings through the simulation experiments indicate that the varying tolerance interval yields more accurate outcomes. Hence, we recommend that users adopt the varying tolerance interval when submitting their verification query to ensure the sub-sample and aggregate algorithm produces result consistent with that derived from the full data set.

Our thesis is based on the assumption that survey weights are fixed when making change to one single data record, yielding a global sensitivity of 1. One potential direction for future research could be exploring how changes in global sensitivity impact algorithm results when the assumption of fixed survey weights is relaxed.

Bibliography

[1] J. M. Abowd and S. D. Woodcock. Disclosure limitation in longitudinal linked data. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 215277:63, 2001.

[2] J. M. Abowd and S. D. Woodcock. Multiply-imputing confidential characteristics and file links in longitudinal linked data. *Privacy in Statistical Databases: CASC Project Final Conference, PSD 2004, Barcelona, Spain, June 9-11. Springer Berlin Heidelberg.*, pages 290–297, 2004.

[3] A. S. Acharya, A. Prakash, P. Saxena, and A. Nigam. Sampling: Why and how of it? *Indian Journal of Medical Specilaities.*, 4(2):330–333, 2013.

[4] A. F. Barrientos, A. Bolton, T. Balmat, J. P. Reiter, J. M. De Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, and M. DeLong. Providing access to confidential research data through synthesis and verification: An application to data on employees of the u.s. federal government. *The Annals of Applied Statistics*, 12(2):1124–1156, 2018.

[5] A. Bkakria, A. Tasidou, N. Cuppens-Boulahia, F. Cuppens, F. Bouattour, and F. Ben Fredj. Optimal distribution of privacy budget in differential privacy. *Risks and Security of Internet and Systems: 13th International Conference, CRiSIS 2018, Arcachon, France, October 16–18, 2018, Revised Selected Papers 13. Springer International Publishing.*, pages 222–236, 2019.

[6] C. Dwork. Differential privacy. *Automata, Languages and Programming. Part II, Lecture Notes in Computer Science. Springer, Berlin*, 4052:1–12, 2006.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity

in private data analysis. *Theory of Cryptography Conference*, pages 265–284, Springer, 2006.

[8] R. M. Groves. *Survey Errors and Survey Costs*. John Wiley & Sons Inc., 1989.

[9] M. Hanif and K. R. W. Brewer. Sampling with unequal probabilities without replacement: A review. *International Statistical Review*, 48:317–335, 1980.

[10] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663 – 685, 1952.

[11] S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, 79(3):362–384, 2011.

[12] E. L. Korn and B. I. Graubard. Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49:291–295, 1995.

[13] R. J. A. Little. Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–426, 1993.

[14] R. J. A. Little, F. Liu, and T. E. Raghunathan. Statistical disclosure techniques based on multiple imputation. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, pages 141–152, 2004.

[15] T. Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9:1–19, 2004.

[16] Gregory J. Matthews and O. Harel. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5:1–29, 2011.

[17] D. R. McClure and J. P. Reiter. Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality*, 4(1), 2012.

[18] R. Mitra and J. P. Reiter. Adjusting survey weights when altering identifying design variables via synthetic data. *Privacy in Statistical Databases 2006, Lecture Notes in Computer Science, New York: Springer-Verlag*, pages 177–188, 2006.

[19] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*, pages 75–84, 2007.

[20] A. Omair. Sample size estimation and sampling techniques for selecting a representative sample. *Journal of Health Specialties*, 2(4):142, 2014.

[21] T. E. Raghunathan. Synthetic data. *Annual Review of Statistics and Its Application*, 8:129–140, 2021.

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

[23] J. P. Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18:531–544, 2002.

[24] J. P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–188, 2003.

[25] J. P. Reiter. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100(472):1103–1112, 2005.

[26] J. P. Reiter. Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21:441 – 462, 2005.

[27] J. P. Reiter, A. Oganian, and A. F. Karr. Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, 53(4):1475–1482, 2009.

[28] D. B. Rubin. *Multiple Imputation for Survey Nonresponse*. John Wiley & Sons Inc., 1987.

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

[30] H. Taherdoost. Sampling methods in research methodology; how to choose a sampling technique for research. *International Journal of Academic Research in Management*, 5(2):18–27, 2016.

[31] S. Tyrer and B. Heyman. Sampling in epidemiological research: Issues, hazards and pitfalls. *BJPsych Bulletin*, 40(2):57–60, 2016.

[32] C. Yang. A differentially private bayesian approach to replication analysis. *Duke University*, 2022.

[33] H. Zheng and R. J. A. Little. Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21(1):1–20, 2005.