TRANSACTIONS ON DATA PRIVACY 18 (2025) 51–66

Differentially Private Verification of Survey-Weighted Estimates

Tong Lin*, **Jerome P. Reiter****

*South Hall 5607A, University of California Santa Barbara, Santa Barbara, 93106, USA.

**Box 90251, Duke University, Durham, NC 27708, USA.

E-mail: tong_lin@umail.ucsb.edu, jreiter@duke.edu

Received 27 January 2024; received in revised form 14 August 2024; accepted 19 August 2024

Abstract. Several official statistics agencies release synthetic data as public use microdata files. In practice, synthetic data do not admit accurate results for every analysis. Thus, it is beneficial for agencies to provide users with feedback on the quality of their analyses of the synthetic data. One approach is to couple synthetic data with a verification server that provides users with measures of the similarity of estimates computed with the synthetic and underlying confidential data. However, such measures leak information about the confidential records, so that agencies may wish to apply disclosure control methods to the released verification measures. We present a verification measure that satisfies differential privacy and can be used when the underlying confidential data are collected with a complex survey design. We illustrate the verification measure using repeated sampling simulations where the confidential data are sampled with a probability proportional to size design, and the analyst estimates a population total or mean with the synthetic data. The simulations suggest that the verification measures can provide useful information about the quality of synthetic data inferences.

Keywords. Complex, Server, Survey, Synthetic, Validation.

1 Introduction

Survey sampling is widely used in various fields to make inferences about finite population quantities like population totals and averages. Typically, survey data are collected using complex sampling designs, such as stratified, probability proportional to size, or cluster sampling. These designs create unequal probabilities that individuals will be selected into the sample. Data analysts need to adjust for the unequal selection probabilities to obtain unbiased estimates of population quantities, for example, by using survey-weighted estimators.

Many survey data sets are collected under pledges to protect the confidentiality of data subjects' identities and sensitive information. As such, agencies seeking to disseminate survey data to the public typically apply some redaction strategies to reduce the risks of unintended disclosures. One approach is to generate synthetic data (Rubin, 1993; Little, 1993; Reiter, 2003; Drechsler, 2011; Raghunathan, 2021; Reiter, 2023), an approach taken for example, by the U.S. Bureau of the Census to share data from the Survey of Income and Program Participation. In this approach, the agency simulates new values of confidential information using models estimated from the confidential data. These data are released as public use files, available for secondary data analysis.

Naturally, the quality of inferences from synthetic data depend critically on the quality of the models used to generate the synthetic data (Reiter, 2005). When the synthetic data models fail to accurately capture the distribution of the confidential data, secondary analysts of the synthetic data can obtain unreliable results. Thus, it is beneficial for agencies to provide means for secondary analysts to get feedback on the quality of their analyses of the synthetic data (Reiter and Drechsler, 2010).

To do so, one approach is to provide secondary analysts access to a verification server (Reiter *et al.*, 2009; McClure and Reiter, 2012; Barrientos *et al.*, 2018b). This is a computer system that has both the confidential and synthetic data. The secondary analyst submits a query to the server for a measure of the similarity of estimates based on the confidential and synthetic data, for example, how far apart are the point estimates computed with the synthetic and confidential data. The server reports back the verification measure to the analyst, who can decide if the synthetic data results are of adequate quality for their purposes.

Verification measures leak information about the confidential data. For example, Reiter *et al.* (2009) and McClure and Reiter (2012) illustrate how attackers could learn confidential information from targeted queries for verifications of results from synthetic data. Thus, it can be beneficial to apply disclosure treatment to verification measures before release. In particular, several researchers (e.g., Amitai and Reiter, 2018; Yu and Reiter, 2018; Barrientos *et al.*, 2018b) have developed verification measures that satisfy differential privacy (Dwork *et al.*, 2006; Dwork, 2006). To date, however, researchers have not developed verification measures that satisfy differential privacy for survey-weighted analyses.

In this article, we propose such measures. The basic idea is to leverage the sub-sample and aggregate algorithm from the differential privacy literature (Nissim *et al.*, 2007). We split the confidential data into disjoint subsets, estimate a survey-weighted analysis in each subset, determine the fraction of these estimates falling within an analyst-specified distance of the synthetic data estimate, and add noise to this fraction using a Laplace Mechanism (Dwork *et al.*, 2006). We investigate the performance of this approach using simulations of probability proportional to size sampling and a survey-weighted estimate of a population total or mean. We consider settings where the synthetic data are representative of the population distribution and where they are not. The simulation results suggest that the methods can provide useful feedback on the quality of synthetic data estimates of population totals when the underlying confidential data are from a complex sample design.

The remainder of this article is organized as follows. In Section 2, we review survey-weighted estimates and differential privacy. In Section 3, we describe our strategy for verification based on the sub-sample and aggregate algorithm. We also discuss the settings of key parameters that affect the properties of the verification measures. In Section 4, we use simulation experiments to illustrate the performance of the verification measures. Finally, in Section 5, we briefly summarize the main findings.

2 Review of Survey-Weighted Estimation and Differential Privacy

In this section, we provide background useful for understanding the verification measures presented in Section 3.

2.1 Survey-weighted Estimation

Let P be a finite population with N elements, each with an index $i = 1, \dots, N$. Let $X = (x_1, \dots, x_N)$ be the population values of some variable. To motivate the methodology, we suppose that the analyst seeks inferences for the population total, $\tau = \sum_{i=1}^{N} x_i$. We extend to estimation of population means in Section 4.3. Let D be a subset of P comprising n elements randomly drawn from P. We define the indicator $I_i = 1$ if element i is in the sample D, and $I_i = 0$ otherwise. The vector $I = (I_1, \dots, I_N)$ represents the elements in D, and $n = \sum_{i=1}^{N} I_i$ is the sample size.

To determine I and hence D, in this article we consider probability proportional to size (PPS) sampling as an illustrative complex sampling design. Let $Z = (z_1, \dots, z_N)$ be a numerical variable known for all N units in P. We sample elements in P with unequal probabilities proportional to Z. For each unit $i = 1, \dots, N$, let $\pi_i = Pr(I_i = 1)$ be its first-order inclusion probability. 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 sum of the z_i in P excluding the cases sampled with certainty.

For any probability sampling design including PPS sampling, a common approach to estimate τ is the Horvitz and Thompson (1952) estimator. We weight each sampled element by the inverse of its inclusion probability, and sum over all units in D. More precisely, for $i = 1, \dots, N$, let $w_i = 1/\pi_i$. We estimate τ using

$$\hat{\tau} = \sum_{i \in D} x_i/\pi_i = \sum_{i \in D} w_i x_i. \tag{1}$$

The estimator in (1) is unbiased for τ for any sampling design, provided $\pi_i > 0$ for $i = 1, \dots, N$.

2.2 Differential Privacy

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

> **Definition 1 (ϵ-differential privacy)**: An algorithm $\mathcal{A}$ satisfies ϵ-differential privacy, abbreviated ϵ-DP, if for any neighboring data sets D and D^*, and any output $o \in Range(\mathcal{A})$, the
> $$Pr(\mathcal{A}(D) = o) \le exp(\epsilon) Pr(\mathcal{A}(D^*) = o). \tag{2}$$

The ϵ is known as the privacy budget. It quantifies the similarity between the outputs of $\mathcal{A}$ being implemented over D and D^*. Intuitively, smaller ϵ makes it more difficult for users to distinguish the data record that differs between D and D^*, 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 $\mathcal{A}_2(D)$ satisfies $(\epsilon_1 + \epsilon_2)$-DP. This is called the sequential composition property. Second, for any two data sets D and E measured on disjoint sets of individuals, releasing the outputs of $\mathcal{A}_1(D)$ and $\mathcal{A}_2(E)$ guarantees max(ϵ_1, ϵ_2)-DP. This is called the parallel composition property. Third, for any algorithm $\mathcal{A}_3$ that does not depend on D, releasing the output $\mathcal{A}_3(\mathcal{A}_1(D))$ satisfies ϵ_1-DP. This is called the post-processing property.

One approach to achieve ϵ-DP is the Laplace Mechanism (Dwork *et al.*, 2006). Let f be a function on $D \to \mathbb{R}^d$; for example, f could sum the elements of one column (i.e., one variable) of D. The global sensitivity is defined as $\Delta(f) = max_{(D,D^*)}\|f(D) - f(D^*)\|_1$ over all neighboring data sets D and D^*. The Laplace Mechanism perturbs $f(D)$ by adding noise drawn from a Laplace distribution, i.e., we compute $f(D)+\eta$, where $\eta \sim Laplace(0,\, \Delta(f)/\epsilon)$.

For some f, $\Delta(f)$ can be large, resulting in a high probability of adding large noise to $f(D)$. In such cases, we may want to satisfy ϵ-DP using an algorithm other than the Laplace Mechanism. One such mechanism, proposed by Nissim *et al.* (2007), is the sub-sample and aggregate algorithm. The basic idea is to randomly partition D into M disjoint subsets, $D' = \{D_1, \ldots, D_M\}$. For each D_k, we determine $f(D_k)$ and then $f_{avg}(D') = \sum_{k=1}^{M} f(D_k)/M$. For many f, including our verification measures, changing D by only one record changes the value of at most one $f(D_k)$. Thus, $\Delta(f_{avg}) = \Delta(f)/M$. We can apply a Laplace Mechanism to this $f_{avg}(D')$ using $\eta_{new} \sim Laplace(0, \Delta(f)/\epsilon M)$. Thus, we have reduced the variance of the noise significantly. We use the sub-sample and aggregate method to develop the differentially private, survey-weighted verification measures, as we now describe.

3 Differentially Private, Survey-weighted Verification

Suppose D is a confidential data set comprising $i = 1, \ldots, n$ individuals measured on p variables. Thus, for any individual i, we have $D_i = (x_{i1}, \ldots, x_{ip})$. We also have a survey weight, $w_i = 1/\pi_i$, where π_i is the first-order inclusion probability of individual i. As a public use file, the agency generates a synthetic data set D_0 comprising n_0 simulated individuals with values of the same p variables in D. We assume that the agency generates D_0 following the approach in Raghunathan *et al.* (2003), in which it (i) simulates values for the $N - n$ records not in D to create a completed population P' and then (ii) takes a simple random sample of size n_0 from P' that is released as D_0. Thus, all $j = 1, \ldots, n_0$ synthetic individuals in D_0 have the simple random sample weights N/n_0. The agency also might replace values for the records in D when making P'; this does not affect our methodology. For simplicity, we also assume that the agency releases only one synthetic data set, which is the case, for example, for the synthetic Longitudinal Business Database; see Kinney *et al.* (2011) and Kinney *et al.* (2014). Our verification measures also can be applied when multiple implicates are released. We simply use the estimates from the synthetic data combining rules (Raghunathan *et al.*, 2003) instead of the estimates from the one D_0.

Suppose that the synthetic data analyst intends to estimate the population total of one of the variables, say X, based on D_0. For example, x_i could be an indicator of whether individual i speaks a certain language, so that $\tau = \sum_{i=1}^{N} x_i$ is the total number of people who speak that language in the population. Let $\hat{\tau}_0 = N\bar{x}_0 = N\sum_{j \in D_0} x_j/n_0$ be the synthetic data analyst's estimate of τ computed from D_0. Let the synthetic data analyst's estimated variance of $\hat{\tau}_0$ computed with D_0 be $\hat{\sigma}^2(\hat{\tau}_0) = N^2((1 - n_0)/N)s_0^2/n_0$, where $s_0^2 = \sum_{j \in D_0}(x_j - \bar{x}_0)^2/(n_0 - 1)$.

3.1 Description of the Algorithm

To construct verification measures, we extend the approach introduced by Barrientos *et al.* (2018b) and Yang and Reiter (2024) to survey-weighted estimates. Let $\hat{\tau}$ be a survey-weighted 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 to motivate the verification algorithm. Let $\hat{d} = |\hat{\tau}_0 - \hat{\tau}|$ be the absolute difference between $\hat{\tau}_0$ and $\hat{\tau}$. When $\hat{d}$ is small, where small is defined by the synthetic data analyst, it suggests that $\hat{\tau}_0$ is sufficiently accurate for the analyst's purposes. We operationalize this by using a tolerance interval centered around $\hat{\tau}_0$, which we refer as $T(\hat{\tau}_0, \alpha)$. Here, α is a parameter that determines the width of the tolerance interval. To illustrate, suppose the synthetic data analyst views D_0 of adequate quality for their purposes if $\hat{\tau}_0$ is within three synthetic-data standard deviations of $\hat{\tau}$. This analyst can set $T = [\hat{\tau}_0 - 3\hat{\sigma}(\hat{\tau}_0), \hat{\tau}_0 + 3\hat{\sigma}(\hat{\tau}_0)]$. As another example, the analyst may decide that $\hat{\tau}_0$ is accurate enough as long as $\hat{\tau}$ is within some percentage of $\hat{\tau}_0$. This analyst can set $T(\hat{\tau}_0, \alpha) = [\hat{\tau}_0 \pm \alpha|\hat{\tau}_0|]$. The analyst then seeks to know whether $\hat{\tau} \in T(\hat{\tau}_0, \alpha)$.

To satisfy ϵ-DP, however, the agency cannot directly release an indicator of whether $\hat{\tau} \in T(\hat{\tau}_0, \alpha)$. The agency should not use a Laplace Mechanism to perturb this indicator, as its global sensitivity equals one, making the Laplace distribution too high variance to return useful information. Further, generally it is not feasible to release a version of $\hat{\tau}$ that satisfies ϵ-DP. As noted in Reiter (2019) and Drechsler (2023), to date there do not exist differentially private algorithms for releasing $\hat{\tau}$ from complex surveys that have low errors for reasonable privacy guarantees.

Instead, we use the sub-sample and aggregate method. The verification server randomly partitions the confidential D into M disjoint subsets, with each partition denoted $D_k \in \{D_1, \ldots, D_M\}$. The sample size of each D_k is $n_k = \lfloor n/M \rfloor$. When n is not divisible by M, some partitions have one more or one less unit than others. In each D_k, the server computes a survey-weighted population estimate of τ using only the data in D_k. To do so, the server inflates each w_i by a multiplicative factor of n/n_k. In particular, for $k = 1, \ldots, M$, the server computes

$$\hat{\tau}_k = \sum_{i \in D_k} w_i (n/n_k) x_i. \tag{3}$$

In each D_k, the synthetic data analyst specifies a tolerance interval $C(\hat{\tau}_0, \alpha, \gamma)$. This interval is not necessarily the same as $T(\hat{\tau}_0, \alpha)$. We discuss ways of setting $C(\hat{\tau}_0, \alpha, \gamma)$ in Section 3.3. For $k = 1, \ldots, M$, let $A_k = 1$ when $\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma)$, and $A_k = 0$ otherwise. Let $S = \sum_{k=1}^{M} A_k$ be the number of partitions where $A_k = 1$. Then, S/M is an estimate of the probability that, for an arbitrary D_k, the $\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma)$. Values of S/M near 1 indicate that the confidential-data estimates in the partitions frequently fall inside the tolerance intervals, which suggests that the estimates from the confidential data are similar to the estimate from the synthetic data. Values of S/M near 0 indicate that estimates from confidential data are dissimilar to the estimate from the synthetic data, suggesting the synthetic data estimate is not sufficiently accurate for the analyst's purposes.

To meet the ϵ-DP requirement, the verification server needs to add noise to S. We do so via the Laplace Mechanism. The server randomly draws a sample $\eta \sim Laplace(0, 1/\epsilon)$ and sets $S^R = S + \eta$. This Laplace Mechanism presumes a $\Delta(f) = 1$, that is, changing one record in D only affects at most one A_k. At the end of this section, we discuss the privacy properties of S^R in more detail.

Because S^R/M can be outside $[0, 1]$, we apply post-processing to enhance the interpretability of the reported verification measure. Specifically, we assume that each $A_k \sim Bernoulli(r)$, where r is the probability that any randomly generated $\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma)$. Hence, we assume that $S|r \sim Binomial(M, r)$. We suppose a uniform prior distribution for r, which equivalently is $r \sim Beta(1, 1)$ where $Beta$ represents a Beta distribution. Thus, the model for post-processing S^R is

$$S^R|S \sim Laplace(S, 1/\epsilon) \quad S|r \sim Binomial(M, r) \quad r \sim Beta(1, 1). \tag{4}$$

We obtain the posterior distribution $p(r|S^R)$ via a Gibbs sampler. The sampler does not use the true value of S, which is unavailable to the algorithm to maintain ϵ-DP. Rather, we average over plausible values of S. The full conditional for r is

$$p(r|S,\, S^R) \propto Pr(S|r)Pr(r) \propto r^S(1-r)^{M-S} \tag{5}$$

which is the kernel of a $Beta(S+1,\, M-S+1)$ distribution. The full conditional for S is

$$Pr(S|r,\, S^R) \propto Pr(S^R|S)Pr(S|r) \propto e^{-\frac{|S^R-S|}{1/\epsilon}} \frac{1}{\Gamma(S+1)\Gamma(M-S+1)} r^S(1-r)^{M-S}. \tag{6}$$

The verification server releases draws from $p(r|S^R)$, including the posterior median.

When $p(r|S^R)$ is concentrated near 1, the analyst can conclude that the synthetic and confidential data offer similar estimates of τ. When $p(r|S^R)$ is concentrated near 0, the analyst can conclude that the synthetic and confidential data estimates of τ are too dissimilar for $\hat{\tau}_0$ to be considered sufficiently accurate. Values of r near 0.5 suggest that the evidence is unclear.

3.2 Privacy Protection

In this section we discuss the privacy protection properties of these verification measures. First, because of the post-processing property of ϵ-DP mentioned in Section 2, releasing $p(r|S^R)$ does not affect the privacy guarantee endowed by generating S^R. The Bayesian modeling only uses S^R; it never uses other results from the confidential data. Second, presuming $\Delta(f) = 1$ for the verification measures implicitly presumes that changing one individual in D does not change the data, including the survey weights, for any other individuals in D. This could be violated, for example, when the agency adjusts survey weights for nonresponse or does data editing by using information from multiple records (Drechsler and Bailie, 2024). We leave accounting for this possibility to future research. Third, we note that D_0 may not satisfy ϵ-DP; indeed, most implementations of synthetic data to date do not. As a result, we cannot rely on the sequential composition property of ϵ-DP to quantify the privacy loss from releasing both S^R (or $p(r|S^R)$) and D_0. Of course, if D_0 (or more precisely $\hat{\tau}_0$) is differentially private (as in, e.g., Bowen and Liu, 2020; Liu, 2022), then the sequential composition property applies. Thus, agencies and analysts can interpret the privacy protection afforded by the verification measures as a bound on the additional privacy leakage due to releasing the verification measure over the leakage from releasing D_0 itself.

Even without the benefit of sequential composition, using a differentially private verification measure has some potentially appealing features. To illustrate, consider instead an agency that releases $d = \hat{\tau}_0 - \hat{\tau}$ exactly, i.e., with no privacy protection, as a verification measure. In this case, an adversary could learn $\hat{\tau}$ by subtraction. Suppose the adversary can construct a set of estimates that, in combination, can be used to isolate some targeted individual $D_i \in D$. For example, the adversary may request verifications where $\hat{\tau}$ is the survey-weighted, estimated total of some sensitive variable, say X_1, for two sets of individuals, one comprising all records in D and the other all records in $D - D_i$. Using these two data sets, the adversary also could ask for verifications where $\hat{\tau}$ is the survey-weighted estimate of the population size. The adversary can learn $w_i x_{i1}$ from the first set of verifications and w_i from the second set of verifications, which then reveals the sensitive x_{i1}.

The situation is improved when the verification measure is the indicator for $\hat{\tau} \in T(\hat{\tau}_0, \alpha)$. Now, the adversary only can learn bounds on w_i and x_{i1}. For example, suppose $\hat{\tau} = 1$ for all four verification queries described previously. Using trial and error, the adversary may be able to identify a set of possible values of the target's w_i that results in a $\hat{\tau}$ inside the tolerance interval for the population size. Similarly, for any feasible w_i identified, the adversary may be able to identify values of x_{i1} that result in a $\hat{\tau}$ inside the tolerance interval for the total of X_1, thereby creating a plausible range for x_{i1} over all feasible w_i. Whether this plausible range for x_{i1} is sufficiently wide to provide adequate protection is, of course, specific to the values in D and D_i. Generally, it is difficult for agencies to identify all such potential attacks before putting a verification server online, as well as to keep track of all possible sets of verification queries that might lead to isolating targeted individuals.

As with any interactive system, each time the agency releases a verification result, it leaks information about the underlying confidential data. Differentially private measures like those presented here offer a way for agencies to quantify the cumulative information leakage, which facilitates agency assessments of the trade offs in disclosure risk and data usefulness inherent from releasing verifications. Verification measures that do not satisfy formal privacy generally do not come with quantifiable metrics for agencies to assess the trade off. To be sure, non-formally-private verification measures may offer adequate protection; however, agencies applying them have to rely on intuition as opposed to mathematics to characterize the privacy protection.

Of course, using differentially private verification measures does not eliminate disclosure risks. The algorithm at best provides a way to bound the additional information leakage in the verification measures, along with a quantifiable metric for that information leakage. We discuss the issue of cumulative privacy loss further in Section 5.

3.3 Specifying the Tolerance Interval

The synthetic data analyst needs to specify $C(\hat{\tau}_0, \alpha, \gamma)$. Here, γ plays the role of an inflation factor that may be used to go from $T(\hat{\tau}_0, \alpha)$, which is based on a sample size of n, to $C(\hat{\tau}_0, \alpha, \gamma)$, which is based on a sample size of approximately n/M. Following Yang and Reiter (2024), we consider two classes of tolerance intervals. First, the analyst may set $C(\hat{\tau}_0, \alpha, \gamma) = T(\hat{\tau}_0, \alpha)$; we call this a fixed tolerance interval. To illustrate, suppose $\hat{\tau}_0 = 100000$ and $\hat{\sigma}(\hat{\tau}_0) = 1000$. The analyst wants to know if $\hat{\tau}$ falls within 10% of $\hat{\tau}_0$, i.e., within 10000. For a fixed tolerance interval, we have $T(\hat{\tau}_0, \alpha) = C(\hat{\tau}_0, \alpha, \gamma) = [90000,\ 110000]$.

Alternatively, the analyst may set $C(\hat{\tau}_0, \alpha, \gamma) \neq T(\hat{\tau}_0, \alpha)$; we call this an adjusted tolerance interval. The main motivation for adjusted tolerance intervals is that the smaller sample size in any D_k increases the variance associated with $\hat{\tau}_k$ compared to the variance of $\hat{\tau}$ from D. If we use a fixed tolerance interval with $C(\hat{\tau}_0, \alpha, \gamma) = T(\hat{\tau}_0, \alpha)$, any $\hat{\tau}_k$ has increased probability of falling outside $C(\hat{\tau}_0, \alpha, \gamma)$ even when $\hat{\tau} \in T(\hat{\tau}_0, \alpha)$. Thus, we use the parameter γ to inflate the tolerance intervals within the partitions.

To do so, we follow the strategy used by Barrientos *et al.* (2018a), which we explain using an illustrative example. Suppose the analyst has in mind $T(\hat{\tau}_0, \alpha) = [\hat{\tau}_0 \pm 3\hat{\sigma}(\hat{\tau}_0)]$. Here, we set $\alpha = 3$, although analysts could choose other values, e.g., $\alpha = 10$ for a tolerance of ± 10000 when $\hat{\sigma}(\hat{\tau}_0) = 1000$. Suppose we have $M = 25$ disjoint partitions, $(D_1, \ldots, D_{25})$. In this case, it can be reasonable to approximate $\hat{\sigma}(\hat{\tau}_k)$ with $\sqrt{n/n_k}\hat{\sigma}(\hat{\tau})$, that is, we inflate the variance to recognize the change in sample size going from D to D_k. We use the inflated standard error when constructing the adjusted tolerance interval, so that $C(\hat{\tau}_0, \alpha, \gamma) = [\hat{\tau}_0 \pm (5)3 \cdot \hat{\sigma}(\hat{\tau}_0)]$. Here, $\gamma = \sqrt{25} = 5$. As a default, we recommend setting $\gamma = \sqrt{M}$ for adjusted intervals. We note that $\gamma = 1$ in the fixed tolerance intervals.

3.4 Choosing M

In this section, we discuss the choice of the number of partitions M. We consider the effect of changing M on S/M itself and on the noise from the Laplace Mechanism. This discussion closely follows that in Yang and Reiter (2024).

By design, S and hence S/M can be one of $M + 1$ values. For instance, when $M = 5$, we have $S/M \in \{0, 0.2, 0.4, \ldots, 1\}$. In this case, S/M might not be granular enough for the analyst to make clear interpretations of the quality of D_0. In addition, with a small M, the perturbation from the Laplace Mechanism will have a greater proportional impact on S, potentially making it more difficult to interpret S^R. On the other hand, for a given D, fewer partitions means larger sample sizes in each D_k. Larger values of n_k reduce the variance of $\hat{\tau}_k$ in each partition, which can result in more reliable inferences about the differences in $\hat{\tau}$ and $\hat{\tau}_0$. Finally, a small M can increase the variance of S/M over the random partitions.

Analysts need to balance these trade offs in selecting M. Overall, the goal is to choose an M so that the verification results are consistent with the results that could be obtained using the full confidential data. In other words, we want $Pr(\hat{\tau} \in T(\hat{\tau}_0, \alpha))$ to be close to $Pr(\hat{\tau}_k \in C(\hat{\tau}_0, \alpha, \gamma))$. Specifically, if $\hat{\tau} \in T(\hat{\tau}_0, \alpha)$, the probability density of S^R/M should have most mass near 1. When $\hat{\tau}$ is outside $T(\hat{\tau}_0, \alpha)$, the probability density of S^R/M should have most mass near 0. In Section 4, we present simulation studies with different M to help inform this decision.

4 Simulation Studies

In this section, we conduct simulation studies to illustrate the properties of the verification measures. We first generate a population P comprising $N = 10000000$ individuals. For each individual i, we generate two variables (z_i, x_i) sampled from $z_i \sim Uniform(0, 10)$ and $x_i | z_i \sim N(z_i + 5, 2)$. For each unit i in P, we assign an inclusion probability proportional to z_i, so that $\pi_i = n z_i / \sum_{i=1}^{N} z_i$ where n is the sample size. Using π_i, we take a PPS sample from P to make the confidential data D.

We generate synthetic data from D using two strategies. The first method involves generating a D_0 that is representative of P. To do so, we need to account for the complex design when synthesizing. Failure to do so can result in synthetic data that do not look like P. Since our goal is to evaluate the verification measures rather than implement a synthesizer that handles survey weights (e.g., Kim *et al.*, 2020; Hu *et al.*, 2021; Mathur *et al.*, 2024), we simply take a simple random sample of size n_0 from P to create D_0. Of course, this is not possible in genuine applications; agencies need to account for the complex design using D to make D_0. However, given that we know D_0 is an accurate representation of P, ideally the verification measures indicate that the synthetic data provide accurate estimates.

In the second method, we generate D_0 directly from D but ignore the sampling design. Specifically, we randomly draw n_0 samples from $\mathcal{N}(\bar{x}, s_x^2)$, where $\bar{x}$ and s_x^2 are the sample mean and variance of the variable X in D. This synthesizer should lead to inaccurate estimates since D_0 is not representative of P. Thus, it allows us to examine the performance of the verification measure when D_0 offers unreliable estimates.

We focus on factors that could affect the performance of the verification algorithm, namely M, n_k, and the tolerance intervals. We consider $n_k \in \{500, 20000, 50000\}$ and $M \in \{25, 50, 90\}$ partitions. For each combination of n_k and M, we draw $n = n_k M$ samples from P using PPS sampling to make D. We set $n_0 = n$. We repeat the steps for generating (D, D_0) for 200 times for each of the two synthetic generation methods. We set $\epsilon = 1$ for all measures. For the tolerance intervals, we use a fixed tolerance interval of $T(\hat{\tau}_0, \alpha) = [\hat{\tau}_0 - \alpha\hat{\sigma}(\hat{\tau}_0), \hat{\tau}_0 + \alpha\hat{\sigma}(\hat{\tau}_0)]$. For the adjusted interval, we set $\gamma = \sqrt{M}$ and $C(\hat{\tau}_0, \alpha, \gamma) = [\hat{\tau}_0 - \alpha\sqrt{M}\hat{\sigma}(\hat{\tau}_0),\, \hat{\tau}_0 + \alpha\sqrt{M}\hat{\sigma}(\hat{\tau}_0)]$. We consider $\alpha \in \{1,\, 3,\, 5\}$.

For each pair of D and D_0, we compute two quantities. First, we define a binary variable Q, which is an indicator that takes value of 1 when $\hat{\tau}$ is inside the original tolerance interval, i.e., $Q = \mathbb{I}(\hat{\tau} \in T(\hat{\tau}_0, \alpha))$. For the 200 pairs of (D_0, D), we get $Q_1, \ldots, Q_{200}$. We then calculate $r_{full} = \sum_{i=1}^{200} Q_i/200$, which is an approximate estimate of $Pr(\hat{\tau} \in T(\hat{\tau}_0, \alpha))$. Of course, the synthetic data analyst does not get Q or r_{full}, as they have only the differentially private results. Nonetheless, we can use r_{full} to evaluate the differentially private measures. Second, with each (D, D_0), we implement the differentially private verification measure to compute the posterior distribution of r. We store the posterior medians of r. Ideally, within any simulation setting, the posterior medians of r are similar to r_{full}, indicating that the differentially private verification measure tends to result in similar conclusions as using the original interval.

4.1 Results for Synthesis Based on SRS of P

Figure 1 summarizes the results for the fixed tolerance intervals when the synthesizer faithfully represents P. There is obvious discrepancy between the values 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}, for all α considered. Quite simply, the verification measure with a fixed tolerance interval does not have acceptable performance.

We next turn to the results for the adjusted tolerance interval, displayed in Figure 2. In most instances, the posterior medians of r are close enough to the values of r_{full} that analysts likely would reach similar conclusions about the verification using either r or r_{full}. When $\alpha = 1$, r_{full} and the posterior medians of r are typically around 0.3. When $\alpha = 3$, the value of r_{full} increases to between 0.5 and 0.75. The majority of the posterior medians of r tend to be larger than r_{full}, suggesting some over-optimism in the verification decision. When $\alpha = 5$, r_{full} and the posterior medians of r tend to be above 0.8.

Holding constant M and α, we see that smaller values of n_k correspond to larger values of both r_{full} and medians of r. Evidently, in these simulations, decreasing n_k increases the probability that $\hat{\tau}$ and $\hat{\tau}_k$ are within the analyst's tolerance interval. However, r_{full} and the posterior medians of r tend to track each other for all n_k considered. Of course, the trend may not hold if n_k gets very small, as the variance of $\hat{\tau}_k$ may become so large as to make S/M go toward zero, particularly when the tolerance interval is tight around $\hat{\tau}_0$ compared to the variance of $\hat{\tau}_k$.

Holding constant n_k and α, the changes in M have little effect on the average value of the posterior medians of r in this simulation. Nonetheless, the variance of the posterior medians of r decreases as M grows larger. This is expected: increasing M reduces the impact of the noise from the Laplace Mechanism on S/M, and thus reduces the variance in S^R/M. As a default, we recommend setting $M = 20$ or $M = 25$ to ensure a fine enough grid while ideally keeping reasonably large sample sizes within the partitions.

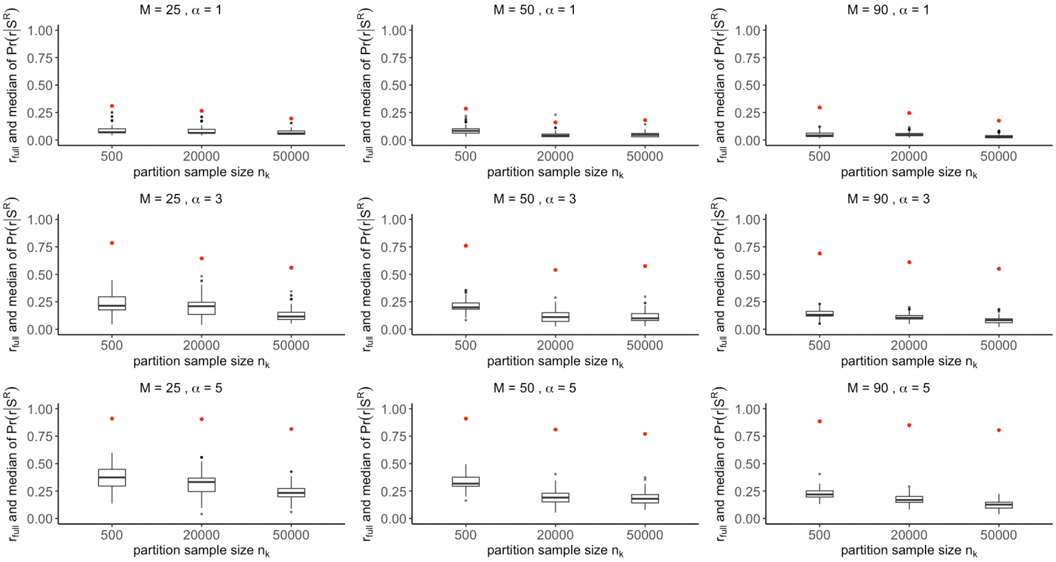

Figure 1: r_{full} (red points) and posterior medians of r (box plots) using fixed tolerance intervals for the population total. Synthetic data are a SRS from P.

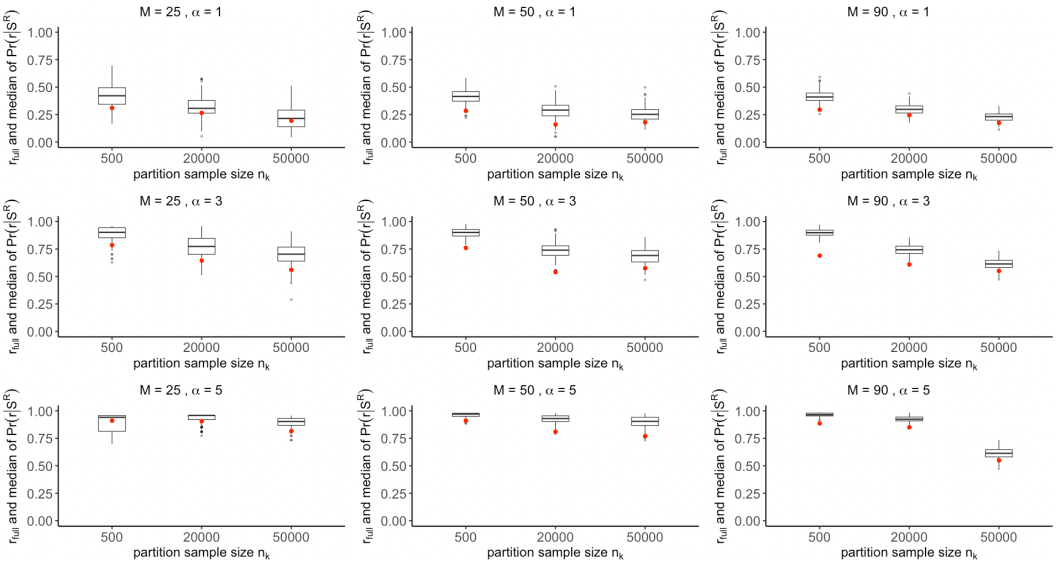

Figure 2: r_{full} (red points) and posterior medians of r (box plots) using adjusted tolerance intervals for the population total. Synthetic data are a SRS from P.

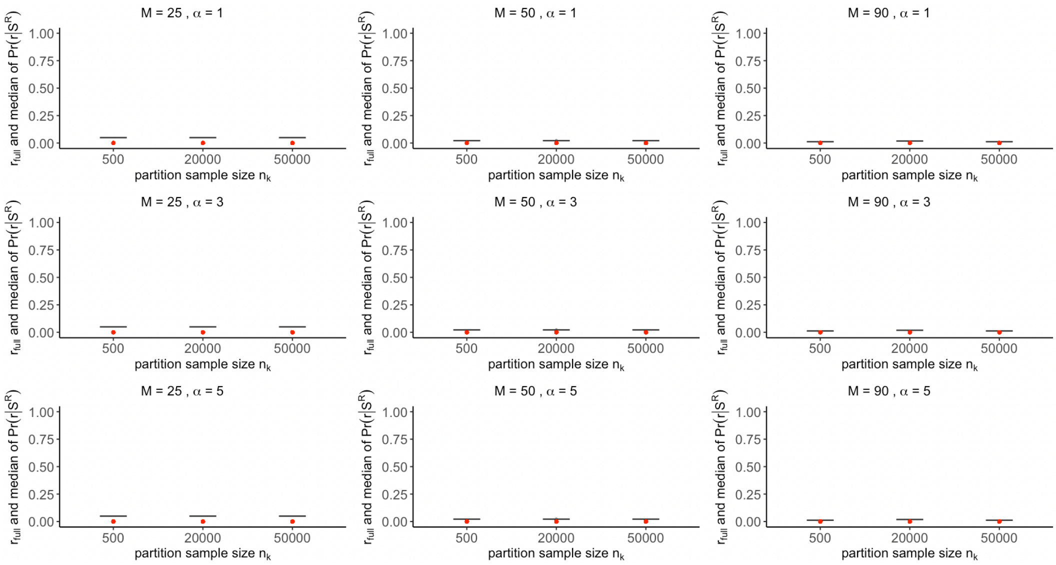

Figure 3: r_{full} (red points) and posterior medians of r (box plots) using adjusted tolerance intervals for the population total. Synthetic data are a biased sample.

4.2 Results for Biased Synthesis

We now turn to the results from the simulation where the agency disregards the sampling design when generating D_0. We expect these synthetic data to be low quality for estimating τ and desire the verification measures to indicate as such.

We first provide evidence that accounting for the survey design is important in this simulation. For each generated D, we estimate τ using both the Horvitz and Thompson (1952) estimator and an unweighted estimator $N\bar{x}$. The true value is $\tau = 99984562$. While the Horvitz and Thompson (1952) estimator is unbiased, the averages of $N\bar{x}$ across the simulation settings tend to be around 117000000, which is much larger than τ.

We implement the verification procedure using the biased synthetic data. Because the fixed tolerance intervals performed poorly in Section 4.1, we only display the results for the adjusted tolerance intervals, shown in Figure 3. Regardless of the value we set for n_k, M, and α, r_{full} and the posterior medians of r are close to 0. The poorly generated synthetic data lead to a biased estimate of τ, so that the $\hat{\tau}_k$ tend not to lie within the tolerance interval. Evidently, the verification measures appropriately clue the analyst that the synthetic data are unreliable for estimating τ accurately.

4.3 Simulation Studies with a Population Average

As an additional set of studies, we repeat the simulations from Section 4.1 and 4.2 using the population average, $\bar{X} = \sum_{i=1}^{N} x_i/N$. For the synthetic data, the estimate is simply $\bar{x}_0 = \sum_{j \in D_0} x_j/n_0$, with estimated variance $\hat{\sigma}_0 = (1 - n_0/N)s_0^2/n_0$. For the confidential data D and each partition D_k, we estimate $\bar{X}$ using survey-weighted ratio estimators. Reusing $\hat{\tau}$ and $\hat{\tau}_k$ for convenience, we have

$$\hat{\tau} = \sum_{i \in D} w_i x_i / \sum_{i \in D} w_i \tag{7}$$

$$\hat{\tau}_k = \sum_{i \in D_k} w_i (n/n_k) x_i / \sum_{i \in D_k} w_i (n/n_k). \tag{8}$$

These are the usual estimators of $\bar{X}$ for PPS samples as well as other common designs.

The patterns in the simulations using the population average mimic those for the population total. In particular, the fixed tolerance interval does not perform well, displaying properties similar to those in Figure 1; we do not display these results here. The adjusted tolerance interval performs reasonably well, especially when $M = 25$, as evident in Figure 4 for the design where the synthetic data come from a good-fitting model and in Figure 5 when the synthetic data come from a biased model. Overall, the results suggest the verification measures can be useful for population averages as well as totals.

5 Final Remarks

In this article, we address the gap in existing verification measures for synthetic data when the underlying confidential data come from a complex survey design. Our findings in the simulation experiments suggest that adjusted tolerance intervals tend to yield more reliable verifications than fixed tolerance intervals. Hence, we recommend using adjusted tolerance intervals as a general practice. Of course, as with all simulation studies, these findings are specific to the simulation design presented here.

The performance of the differentially private verification measures depends on the features of the synthetic data, the confidential data, the desired tolerance interval, and the privacy budget. Thus, it is beneficial for agencies to help analysts assess the usefulness of the verification measures for their particular setting. One approach is for the agency to provide a tool—perhaps as a component of the verification server—for analysts to simulate approximate sampling distributions of the verification measures for parameters of their choosing. The simulations could be based on previously released data (e.g., from a prior year run of the survey) that mimic the sampling design and population characteristics of the confidential data. Alternatively, the simulation tool could follow the strategy outlined by Yang and Reiter (2024). In our context, this strategy entails the following.

1. The analyst specifies a grid of values of $\hat{\tau}$ of scientific interest.

2. The analyst specifies a grid of possible values of the standard error of $\hat{\tau}$, say $\hat{\sigma}$, using the sampling variance from the synthetic data as an anchor.

3. For a given $(\hat{\tau}, \hat{\sigma})$, the analyst generates M plausible values of the estimated totals within partitions. These are sampled from normal distributions centered at $\hat{\tau}$ with standard deviation $\sqrt{M}\hat{\sigma}$. Using these M draws, the analyst computes the verification measure for the tolerance interval(s) of interest.

4. The analyst repeats step 3 a large number of times to get an approximate sampling distribution of S^R/M for that $(\hat{\tau}, \hat{\sigma})$.

5. The analyst repeats steps 3 and 4 for each $(\hat{\tau}, \hat{\sigma})$ set in step 1 and 2.

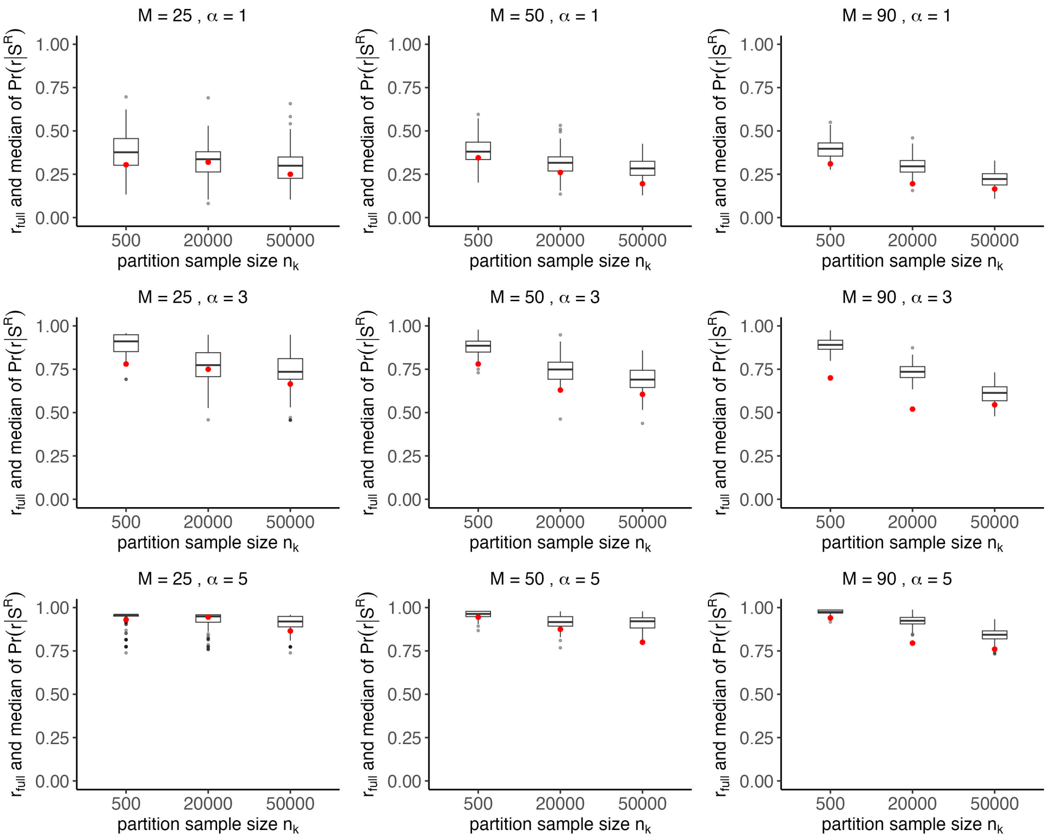

Figure 4: r_{full} (red points) and posterior medians of r (box plots) using adjusted tolerance intervals for the population average. Synthetic data are a SRS from P.

For any (M, α) and plausible set of confidential data results, the analyst can assess whether or not the verification measure is likely to return a correct verification. If the simulated performance is unsatisfactory at all reasonable levels of M, the analyst may have to consider verifications based on other tolerance intervals, e.g., using different α than initially proposed, to get a reliable result.

Finally, we note that some analysts of synthetic data are likely to request many verifications, which if granted can accumulate information leakage. Exactly how much leakage to permit is a policy decision (Yang and Reiter, 2024). For example, as suggested in Barrientos *et al.* (2018b), agencies may decide to give each user a privacy budget. This strategy is under consideration for the synthetic tax file plus validation system being developed by the U. S. Internal Revenue Service (Burman *et al.*, 2024). See Barrientos *et al.* (2018b) for additional discussion of this approach. Alternatively, the agency may decide to enforce a total privacy budget. Such a budget could be exhausted, particularly for data where many verifications are granted. Of course, the agency always can make the policy decision to allow for more verifications, i.e., to increase the total privacy budget.

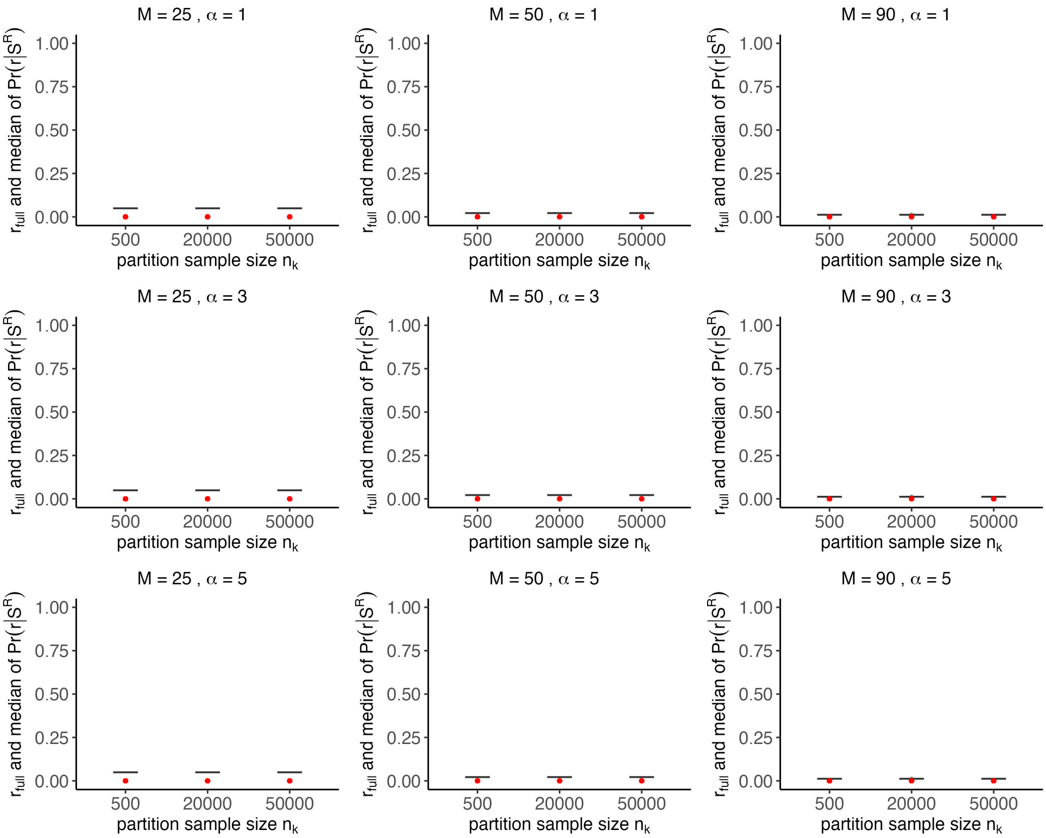

Figure 5: r_{full} (red points) and posterior medians of r (box plots) using adjusted tolerance intervals for the population average. Synthetic data are a biased sample.

Acknowledgements

This research was supported by the U. S. National Science Foundation grant SES-2217456.

References

Amitai, G. and Reiter, J. P. (2018). Differentially private posterior summaries for linear regression coefficients. *Journal of Privacy and Confidentiality*, **8**, Article 3.

Barrientos, A. F., Reiter, J. P., Machanavajjhala, A., and Chen, Y. (2018a). Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, **28**, 440 – 453.

Barrientos, A. F., Bolton, A., Balmat, T., Reiter, J. P., Figueiredo, J. M. D., Machanavajjhala, A., Chen, Y., Kneifel, C., and DeLong, M. (2018b). 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**, 1124 – 1156.

Bowen, C. M. and Liu, F. (2020). Comparative study of differentially private data synthesis methods. *Statistical Science*, **35**, 280 – 307.

Burman, L., Johnson, B., Bryant, V., MacDonald, G., and McClelland, R. (2024). Protecting privacy and expanding access in a modern administrative tax data system. Presentation. Slides accessed on July 30, 2024.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*. Berlin: Springer-Verlag.

Drechsler, J. (2023). Differential privacy for government agencies—are we there yet? *Journal of the American Statistical Association*, **118**, 761 – 773.

Drechsler, J. and Bailie, J. (2024). Whose data is it anyway? Flavors of differential privacy for survey sampling. In *Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences*, page forthcoming. University of Chicago Press.

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

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

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

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

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

Kinney, S. K., Reiter, J. P., and Miranda, J. (2014). Synlbd 2.0: Improving the synthetic Longitudinal Business Database. *Statistical Journal of the International Association for Official Statistics*, **30**, 129 – 135.

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

Liu, F. (2022). Model-based differentially private data synthesis and statistical inference in multiple synthetic datasets. *Transactions on Data Privacy*, **15**, 141 – 175.

Mathur, S., Reiter, J. P., and Si, Y. (2024). Fully synthetic data for complex surveys. *Survey Methodology*, **50**, forthcoming.

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

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

Raghunathan, T. E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, **8**, 129 – 140.

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

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

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

Reiter, J. P. (2019). Differential privacy and federal data releases. *Annual Review of Statistics and Its Application*, **6**, 85 – 101.

Reiter, J. P. (2023). Synthetic data: A look back and a look forward. *Transactions on Data Privacy*, **16**, 15 – 24.

Reiter, J. P. and Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, **20**, 405 – 422.

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

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

Yang, C. and Reiter, J. P. (2024). Differentially private methods for releasing results of stability analyses. *The American Statistician*, **78**, 180 – 191.

Yu, H. and Reiter, J. P. (2018). Differentially private verification of regression predictions from synthetic data. *Transactions on Data Privacy*, **11**, 279 – 297.