Bootstrap-based, General-purpose Statistical Inference from Differential Private Releases

Marcel Neunhoeffer[*] Daniel Sheldon[†] Adam Smith[*]

Working Paper—May 2, 2023

Abstract

Statistical inference with differential privacy is essential and often depends on bespoke solutions. The combination of sampling and privacy noise for proper inference is not trivial, especially when sampling and privacy noise come from different distributions. We propose a general-purpose method combining the bootstrap with differentially private non-parametric distribution estimation. Our method applies non-private estimators (e.g., MLE for logistic regression) to differentially private synthetic data or distribution estimates. The advantage of our approach is that the bootstrap is pure post-processing of a differentially private mechanism—it does not access the sensitive data multiple times and does not increase the privacy budget. The joint sampling and privacy distribution of statistical estimators is approximated through statistical simulation. We present the results of a series of Monte Carlo experiments and show that our method produces valid inferences for a wide range of data sets (univariate data, multivariate data) and statistical problems (i.e., linear and non-linear queries). Furthermore, we show that our method produces valid confidence intervals that are narrower than confidence intervals produced by bespoke methods.

[*] Department of Computer Science, Boston University. {marceln,ads22}@bu.edu.

[†] Department of Computer Science, University of Massachussets, Amherst.

1 Introduction

Disclosure avoidance technology for statistical data is becoming increasingly sophisticated, with a wide range of complex algorithms—for example, the TopDown algorithm employed by the US Census Bureau to create data products from the 2020 decennial census—applied to the data to generate public-facing releases. This complexity stems from an improved understanding of possible attacks on statistical releases (e.g., Dick et al. (2023)). These new algorithms can be analyzed via rigorous frameworks such as differential privacy Dwork & Rothblum (2016*a*). A substantial resulting benefit is that the details of the algorithm—including noise variances, swap rates, and similar parameters—can be made public.

Even though disclosure limitation necessarily restricts what is revealed about a data set, transparent public descriptions of the algorithms allow, in theory, for principled statistical inference that incorporates both the uncertainty due to sampling—the traditional concern of statistical inference—and the additional randomness and distortion introduced for disclosure limitation. For example, confidence intervals for a parameter estimate, and tests of significance, measures of goodness of fit should all be adjusted to account for the extra processing of the data. The research community has started to develop the methodology necessary—there is a growing body of work on carefully designing differentially private algorithms to include information for uncertainty quantification (confidence intervals, or p-values, for example), usually in ways that are tailored to a particular inference task. This type of bespoke analysis is challenging, however, for complex algorithms like those used to generate differentially private synthetic data.

In this paper, we extend a recent parametric approach of Ferrando et al. (2022) to propose a general-purpose method combining the bootstrap with differentially private nonparametric distribution estimation. Our method applies to any non-private estimator (e.g., MLE for logistic regression) that is run on differentially private synthetic data or distribution estimates. The advantage of our approach is that the bootstrap is pure post-processing of a differentially private mechanism—the method does need to access the sensitive data multiple times and, thus, the approach does not increase the privacy budget and applies even when the disclosure methodology is designed and run by another party (e.g., when an outside researcher seeks to use Census-generated data products). The joint sampling and privacy distribution of statistical estimators is approximated through statistical simulation.

We present the results of a series of Monte Carlo experiments with univariate and multivariate datasets. Our method produces valid inferences for various statistical problems (i.e., linear and non-linear queries). Furthermore, we show that our method produces valid confidence intervals that are narrower than those produced by bespoke methods.

1.1 Related Work

The problem of statistical inference with differential privacy has been studied for some time. A first line of work focuses on developing bespoke methods for statistical tests and confidence intervals for specific estimators. For example, Vu & Slavkovic (2009) develop statistical hypothesis tests and contingency tables under ϵ-differential privacy. Wang et al. (2015) develop differentially private likelihood ratio and chi-squared tests. Gaboardi et al. (2016) develop a differentially private chi-squared test. Karwa & Vadhan (2017) present a method to estimate confidence intervals of the mean of a normal population. Awan & Slavković (2018) derive the uniform most powerful hypothesis test for binary data. Canonne et al. (2019) present optimal hypothesis tests for simple hypotheses under differential privacy and apply them to change-point detection. Drechsler et al. (2022) develop non-parametric confidence intervals for the median (and other quantiles) based on a private estimate of the cumulative distribution function.

What these papers have in common is that developing a differentially private statistical test requires careful analysis of the sampling and privacy noise. This means that each test requires a novel analysis, and even if an analysis exists, practical implementations for applied researchers often still need to be developed.

That is why more general methods based on the bootstrap have also found favor in the literature of statistical inference with differential privacy. Brawner & Honaker (2018) present a bootstrap method to estimate the mean and standard deviation of a distribution by splitting up the data into disjoint subsets and applying a private estimator to each of the subsets. Wang et al. (2022) present a tighter analysis of this bootstrap idea on disjoint subsets. A limitation of this idea is that the privacy budget of the private estimator needs to be distributed among the disjoint subsets. Covington et al. (2021) and Evans et al. (2020) present methods that are based on the bag of little bootstraps. Covington et al. (2021) still split the privacy budget across the different subsets. The method in Evans et al. (2020) is limited to scalar estimands.

A limitation that these approaches have in common is that they require a lot of data. After all, each of the disjoint subsets must contain enough information. For example, Covington et al. (2021) present their results based on experiments where each subset of the data had at least 200 observations (and the full dataset had at least 100000 observations).

A more recent line of work looks at post-processing differentially private distribution estimates or synthetic data for statistical inference. Räisä et al. (2023) offer a Bayesian noise-aware method combined with techniques from multiple imputation to produce valid inferences from differentially private synthetic data. Ferrando et al. (2022) is the closest relative to the method we propose in this paper. They propose a parametric bootstrap based on privately learned parameters of the assumed underlying data-generating process.

2 Preliminaries

2.1 Differential Privacy

Differential privacy Dwork et al. (2016) is a formal definition of privacy for statistical data analyses, which limits the amount of information that can leak about any individual. Given a space of datasets $\mathcal{X}^n$, we say that two datasets X, X' are *neighboring* if they differ in one individual's information. There are a few variants of differential privacy, but they all aim to formalize the same intuition, namely:

> A randomized algorithm $\mathcal{M}$ is differentially private if every pair of neighboring datasets $X, X' \in \mathcal{X}^n$, the random variables $\mathcal{M}(X)$ and $\mathcal{M}(X')$ are similarly distributed.

The different variants of DP use different measures of "similarly distributed". They are typically parametrized by a positive real number, the *privacy parameter*, which determines how similar the distributions should be. They share a similar interpretation: because the change of one record does not significantly affect the output distribution, little is revealed about any one individual. Consider an outside observer who is trying to learn about a particular data subject (say, Alice). Then, no matter what the observer knows ahead of time about Alice, they will draw similar conclusions *whether or not Alice's true data were used to compute* $\mathcal{M}(X)$. Kasiviswanathan & Smith (2008) elaborate a Bayesian formulation of this idea.

There are many different ways to design differentially private algorithms. A typical approach is to add Gaussian noise to some statistic computed from the data, where the variance of the noise is chosen based on the privacy parameter and the "sensitivity" of the function (see the Matrix Mechanism below).

Zero-concentrated differential privacy (zCDP) (Dwork & Rothblum 2016*b*, Bun & Steinke 2016) is a particular variant of differential privacy that quantifies the closeness of distributions via Renyi divergences and is especially useful to analyze the Gaussian mechanism.

Definition 2.1 (Zero-Concentrated Differential Privacy (zCDP) Bun & Steinke (2016))**.** A randomized algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{R}$ is *ρ-zCDP* if for every pair of neighboring datasets

$X, X' \in \mathcal{X}^n$, and for all $\alpha \in (1, \infty)$,

$$\mathrm{D}_\alpha(\mathcal{M}(X)||\mathcal{M}(X')) \leq \rho\alpha,$$

where D_α denotes the Rényi divergence of order α.

To a first approximation, zCDP provides a meaningful privacy guarantee when $\rho \leq 1$, with the guarantee getting stronger as ρ goes to 0. A good way to understand this guarantee is by considering how it applies when the outputs of the algorithm are distributed as Gaussians. The condition imposed by zCDP on $\mathcal{M}(X)$ and $\mathcal{M}(X')$ is, very roughly, that they are no more distinguishable than two univariate Gaussian distributions of variance 1 with means of 0 and $\sqrt{2\rho}$, respectively.

The most widely used variant of differential privacy, often dubbed *approximate DP*, is implied by zCDP. Specifically, Bun & Steinke (2016) show that if $\mathcal{M}$ satisfies ρ-zCDP, then $\mathcal{M}$ satisfies $(\rho + 2\sqrt{\rho \log(\frac{1}{\delta})}, \delta)$-approximate differential privacy for any $\delta > 0$. (Readers unfamiliar with the parameters of approximate DP may ignore the specifics.)

Most standard notions of differential privacy, including zCDP, satisfy a sort of data-processing property, generally known as *closure under post-processing*. Specifically, if $\mathcal{M}$ is ρ-zCDP, then for every (possibly randomized) algorithm A the composed mechanism $A(\mathcal{M}(\cdot))$ is also ρ-zCDP.

The differentially private mechanisms we use in this paper are based on Gaussian noise. Suppose the universe of possible data records is partitioned into d disjoint bins. We may represent a data set X as a histogram h_X—a vector of d counts indicating how many data records fell into each bin. Suppose further that we wish to release an approximation to the vector Mh_X where M is a fixed, known matrix. For example, M could be the lower-triangular matrix describing *cumulative sum queries*:

$$M_{\text{sums}} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \\ 1 & \cdots & 1 \end{pmatrix}$$

(In this case, the ith entry of $M_{\text{sums}}h_X$ is the sum of the first i entries of h_X.)

Consider the mechanism which simply adds isotropic Gaussian noise to the desired output: $\mathcal{M}(X) = MX + Z$, where $Z \sim \mathcal{N}(0, \sigma^2\mathbb{I})$ and $\mathbb{I}$ is the identity matrix. To satisfy ρ-zCDP, it suffices to set $\sigma^2 = \frac{\|M\|_{1\to 2}^2}{2\rho}$, where $\|M\|_{1\to 2}$ is the maximum ℓ_2 norm of any column of M.

In many cases, one can introduce substantially less error by adding non-isotropic noise. Specifically, suppose we have a factorization $M = LR$, where L, R are arbitrary matrices. Then we can satisfy ρ-zCDP by releasing $Mh_X + LZ$ where $Z \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$ and $\sigma^2 = \frac{\|R\|^2_{1\to 2}}{2\rho}$. The mean squared error (i.e. expected squared ℓ_2 norm of the noise) is $\frac{1}{2\rho}\operatorname{tr}(L^T L)\|R\|^2_{1\to 2}$. The idea of the matrix mechanism (Li et al. 2015) is to optimize over L and R to minimize the mean squared error. Our algorithm for approximating a CDF differentially privately uses this approach to release cumulative sums via a recent, explicit factorization M_{sums} of Fichtenberger et al. (2022).

2.2 Bootstrap Estimation of Confidence Intervals

Suppose we have a data set sampled i.i.d from an unknown probability distribution P (the *population*):

$$x_i \overset{\text{iid}}{\sim} P \text{ for } i = 1, 2, \ldots, n.$$

In order to estimate some real-valued population-level quantity $f(P)$, we compute an estimate $\hat{\theta} = f(X)$ by evaluating the functional f on $X = (x_1, ..., x_n)$. [1] For example, if f is the median, then $f(X)$ and $f(P)$ are the empirical and true population medians, respectively. The sampling distribution of $\hat{\theta}$ at P is the distribution of $\hat{\theta} = s(X)$ when X is indeed sampled according to P.

$$P \overset{\text{iid}}{\to} X \overset{f}{\to} \hat{\theta}$$

Suppose we wish to quantify the variability of our estimate by deriving a confidence interval for $f(P)$. Under mild assumptions, a good strategy for building a confidence interval is to base it on the variability of the sampling distribution at P. Since P is unknown, a common and widely successful approach is to estimate that sampling distribution via the *bootstrap*: specifically, in the *non-parametric* boostrap, we consider the sampling distribution of f on data drawn from the empirical distribution $\hat{P}$ of X; in the *parametric* (a.k.a. *model-based*)bootstrap, we assume a specific form for the distribution P and look at the sampling distribution of f at some estimate $\hat{P}$ (e.g., derived from estimated parameters). Either way, we can then estimate the sampling distribution at $\hat{P}$ by repeated sampling.

$$\hat{P} \overset{\text{iid}}{\to} X^* \overset{f}{\to} \hat{\theta}^*$$

[1] For convenience, we use the same symbol f to denote the estimator evaluated on the data set and the statistical functional defined on distributions; by assumption, $f(\hat{P}) = f(X)$ when $\hat{P}$ is the empirical distribution on X.

2.3 Privacy and the Bootstrap

The bootstrap methodology does not directly lend itself to a setting where we insist on differentially private outputs. The problem is that we must account for the information leaked not only in the initial estimate $\hat{\theta}$, but also the information leaked by the description of the confidence interval. Running a standard nonparametric bootstrap involves using the data set many times which complicates the privacy analysis considerably.

The simplest way around this challenge is to use the composition properties of differential privacy to account for many data accesses that come with bootstrapping (as in Honaker (2015), Brawner & Honaker (2018)). Unfortunately, the upper bound on the privacy parameter ρ that one gets via composition increases quickly with the number of bootstrap samples; in most situations, this significantly decreases the accuracy at a given final parameter ρ.

Another approach is based on the subsample-and-aggregate framework of Nissim et al. (2007), in which the dataset is randomly split into k subsamples of smaller size $n' = n/k$, the statistic f is computed (without noise) on each subsample and the k resulting estimates are aggregated differentially privately. This approach fares well in the asymptotic limit (see, e.g., Smith (2011)) or, in practice, on large data sets Evans et al. (2020), Covington et al. (2021). However—as discussed in the Introduction—it generally performs poorly on modest samples sizes.

Finally, Ferrando et al. (2022) take a different approach, which we build on in this paper. They assume a parametric model for the true population distribution and consider a mechanism $\mathcal{M}$ that outputs a (noisy, differentially private) estimate $\tilde{\theta}$ of the full parametric description of the population. This can be interpreted as a (parametric) estimate of a synthetic population $\tilde{P} = P_{\tilde{\theta}}$. They propose sampling fresh data sets from $\tilde{P}$ and running the private estimator $\mathcal{M}$ on them to estimate the (quantiles of) the sampling distribution of $\tilde{\theta}$ at $\tilde{P}$. The hope—borne out by their experiments—is that the variance of the sampling distributions at the synthetic population $\tilde{P}$ will be similar to the variable at the true population P when the data size is sufficiently large to have $\tilde{P} \approx P$.

2.4 Bayesian Methods for Inference from Private Outputs

A very different approach to principled inference from noisy releases is a Bayesian one. Briefly: if we posit a prior distribution on the true population P, then we can meaningfully ask to construct the posterior distribution on P (or its parameters) given the output of the mechanism $\mathcal{M}(X)$. This approach has been used successfully for inference based on outputs generated by several specific mechanisms (generally ones based on the addition of unbiased noise), as in (e.g, Bernstein & Sheldon 2018, Kulkarni et al. 2021).

There are two drawbacks of the Bayesian approach: First, the formulation of the prior may influence the results of inference. This problem is not specific to privacy, however, and there are standard solutions (e.g., using minimally informative priors or comparing results obtained with several different priors). More fundamentally, the Bayesian approach requires a detailed understanding of the likelihood function for the differentially private mechanism $\mathcal{M}$, which might be difficult to write analytically or hard to work with computationally (for example, in the case of noisy gradient-based optimization methods, or the Census Bureau's TopDown algorithm).

We focus on bootstrap-based methods because of their simplicity, generality, and their familiarity among practictioners.

3 Description of the Methodology

The starting point for our paper is the observation that the parametric approach of Ferrando et al. (2022) can be adapted to nonparametric settings: The bootstrap methodology applies to any private mechanism whose output can be viewed as a synthetic population $\tilde{P}$. Crucially, confidence interval generation entails no changes to the mechanism itself and no additional privacy cost, making it ideally suited to settings where the mechanism is designed and run by another entity (a government agency or another research group, for example).

Suppose our differentially private mechanism $\mathcal{M}$, on input X, outputs a population estimate $\tilde{P}$. This population could take any form—for example, a list of parameters for a parametric model, a nonparametric density estimate, a CDF, or a synthetic data set. From $\tilde{P}$, we compute a statistic $\tilde{\theta} = f(\tilde{P})$, which we view as an estimate of a true population quantity $f(P)$.

Given $\tilde{\theta}$ and $\tilde{P}$, we simulate many runs of the entire process—sampling plus private estimation—to understand the sampling distribution of $\tilde{\theta}$ *on data drawn from* $\tilde{P}$. When $\tilde{P}$ is close to P, this gives us a good bound on the sampling distribution of $\tilde{\theta}$ *on the true population* P. Absent privacy constraints, we could take $\tilde{P}$ to be the empirical distribution $\hat{P}$ on X, in which case the method would be the same as the standard non-parametric bootstrap.

The method is described in Algorithm 1, and illustrated in Figure 1.

Lemma 3.1. *If $\mathcal{M}$ is ρ-zCDP, then for every choice of f, B, and α, Algorithm 1 is also*

Algorithm 1

```
Inputs: Dataset $X$ of size $n$, differentially private algorithm $\mathcal{M}$, functional $f$, number of bootstrap replicates $B$, significance level $\alpha$.
Output: $100 \cdot (1-\alpha)\%$ Confidence Interval for $\tilde{\theta}$: $CI^{1-\alpha}_{\tilde{\theta}}$
$\tilde{P} \leftarrow \mathcal{M}(X)$
$\tilde{\theta} \leftarrow f(\tilde{P})$
for $b = 1, 2, \ldots, B$ do
    Sample $\tilde{X}^{*b}$ of size $n$ from $\tilde{P}$
    $\tilde{P}^{*b} \leftarrow \mathcal{M}(\tilde{X}^{*b})$
    $\tilde{\theta}^{*b} \leftarrow f(\tilde{P}^{*b})$
end for
return $CI^{1-\alpha}_{\tilde{\theta}} \leftarrow \left(\text{quantileEstimate}^{\frac{\alpha}{2}}\left(\tilde{\theta}^{*1}, \ldots, \tilde{\theta}^{*B}\right), \text{quantileEstimate}^{1-\frac{\alpha}{2}}\left(\tilde{\theta}^{*1}, \ldots, \tilde{\theta}^{*B}\right)\right)$
```

$$\begin{tikzcd}
P \arrow[r, "\text{sample}"] \arrow[d, "f"] & X \arrow[r, "\mathcal{M}"] & \tilde{P} \arrow[r, "\text{sample}"] \arrow[d, "f"] & \tilde{X}^{*b} \arrow[r, "\mathcal{M}"] & \tilde{P}^{*b} \arrow[d, "f"] \\
\theta & & \tilde{\theta} & & \tilde{\theta}^{*b} \\
& & & b = 1, \ldots, B &
\end{tikzcd}$$

Figure 1: A diagram of our proposed methodology.

ρ-zCDP.

Proof. We only access the sensitive data X in the first line, when we obtain $\mathcal{M}(X)$. Every subsequent step in the algorithm is post-processing. □

The same output $\tilde{P}$ maybe be used for different functionals f and significance levels α (subject to the usual caveats about multiple hypothesis testing and adaptive data analysis).

This flexibility and generality has a few specific benefits:

- In many cases, the functional f we wish to understand is specified by an estimator s which is only defined on finite data sets, rather than on distributions. For example, s might be a particular estimator for logistic regression. In such cases, one can extend

s to a functional f_s defined on arbitrary distributions P: to evaluate $f_s(P)$, sample a sufficiently large data set $\tilde{Y}$ from P and return $s(\tilde{Y})$. The functional f_s is not always deterministic, but in most cases of interest, its value will be sufficiently concentrated to essentially be deterministic.

- In Algorithm 1, we summarize the list of $\tilde{\theta}^{*b}$ by applying a quantileEstimate function to get a confidence interval for θ. There are several choices for how to perform this estimation, just as in the classical bootstrap (see, e.g., Efron & Hastie (2021)). To construct a $100 \cdot (1-\alpha)$%-confidence interval, where α is the desired significance level, the quantileEstimate function could, for example, directly use the $(\frac{\alpha}{2}, 1-\frac{\alpha}{2})$ quantiles of the resulting bootstrap distribution, this is the so-called *percentile method* (Efron & Hastie 2021, p. 185). By comparing the initial estimate $\tilde{\theta}$ to the median of the bootstrap distribution, we can also construct bias-corrected percentile confidence intervals by shifting the quantiles of the percentile method to account for median bias (this is called the *bias-corrected percentile method* (BC) (Efron & Hastie 2021, p. 190)). Alternatively, we can use a normal approximation to construct a confidence interval using the mean and the standard deviation of the bootstrap distribution. In our applications, we compare the percentile method and the bias-corrected percentile method.

4 Confidence Intervals for Univariate Data

To show how our method works, we start with a simple case. We want to calculate confidence intervals for functionals f calculated on univariate data. I.e., the data X is a vector where each individual contributes one entry, and each entry is drawn i.i.d. from an underlying population distribution P. In our example, we use the median as our functional f, to compare our method to the bespoke method for differentially private confidence intervals for the median presented in Drechsler et al. (2022). However, as we release a full distribution estimate $\tilde{P}$ with differential privacy, we can calculate other statistics of interest, e.g., the mean. going beyond the median.

We further compare the confidence intervals generated by our method to non-private confidence intervals calculated on the same sample and a naive differentially private confidence interval that uses a synthetic data representation of $\tilde{P}$ as if it were the original data sample. We compare the confidence intervals on two dimensions—the length and the coverage rate. The best confidence interval is the shortest interval with at least nominal coverage of the true population value.

Table 1: The experimental conditions for the univariate Monte Carlo experiments.

Data-generating processes P	standard normal: $\mathcal{N}(\mu = 0, \sigma^2 = 1)$ lognormal: Lognormal$(\mu = 0, \sigma^2 = 1)$ bimodal adult age
Sample size n	10, 25, 50, 100, 500
Privacy parameter ρ	0.005, 0.01, 0.05, 0.1, 0.2, 0.5, 1
Noise addition mechanism	Tree-based mechanism as in Drechsler et al. (2022). Matrix mechansim with ρ-zCDP as in Fichtenberger et al. (2022).

To empirically evaluate the performance of the different methods, we set up several Monte Carlo simulations. First, we set a true population distribution P from which we draw 1000 samples each of size n. We consider both synthetic populations—specified by mixtures of Gaussians—and "real" populations corresponding to the empirical distribution on a large real data set. We then use the different methods to generate confidence intervals for each sample. With this, we can evaluate and compare the performance of the different methods. To make the results comparable across different population distributions, we calculate the confidence interval length relative to the non-private confidence interval (calculated on the same data X). This allows us also to understand the relative contributions of sampling and privacy noise to the length of the confidence intervals (as in Drechsler et al. (2022)). A relative confidence interval length between 1 and 2 means, roughly, that the sampling noise is at least as influential as variability due to privacy. A relative length larger than 2 indicates that the distortion due to privacy noise is more significant.

4.1 Experimental Setup

To show how our method works, we set up 280 experiments—we evaluate four different data-generating processes, seven values of the privacy parameter ρ, five sample sizes, and two noise addition mechanisms. Table 1 summarizes these experimental conditions, we run all 280 permutations of these conditions.

4.2 Results

In this section, we highlight some main observations about the CIs generated by the bootstrap-based approach from our experiments. We include additional results in the Supplementary Materials. The generality of our conclusions is, of course, limited by the experiments we performed. We hope they are representative of many basic inference tasks.

Shorter confidence intervals than bespoke methods. In Figure 2, we look at confidence intervals for the median of a standard normal distribution with a sample size of 100 across the different levels of the privacy parameter ρ. We show the confidence interval length relative to the non-private intervals for the same sample size. The different colors indicate the different methods (naive, bootstrap, bespoke) to get differentially private confidence intervals. The blue symbols indicate our proposed method, the purple symbols indicate the naive method, and the green symbols indicate the bespoke method for confidence intervals of the median described in Drechsler et al. (2022). The shapes indicate the two noise addition mechanisms, with circles indicating the matrix mechanism described in Fichtenberger et al. (2022) and triangles indicating the tree-based noise addition mechanism described in Drechsler et al. (2022).

First, both the bootstrap-based method and the bespoke method produce confidence intervals of at least nominal coverage across all values of ρ. Accounting for both privacy and sampling noise means that the length of the confidence intervals should depend on ρ. This is the case for both the bootstrap-based method and the bespoke method. The naive method neglects the additional privacy noise, therefore, the length of the confidence intervals is independent of the values of ρ, which leads to invalid confidence intervals.

Using the same noise addition mechanism, the confidence intervals produced by our bootstrap-based method are always shorter than the confidence intervals produced by the bespoke method.

In Figure 3, we look at confidence intervals for the median of a log-normal distribution with a fixed value of $\rho = 0.05$ across different sample sizes. Again, using the same noise addition mechanism, the confidence intervals produced by our bootstrap-based method are always shorter than the confidence intervals produced by the bespoke method. Furthermore, the confidence intervals produced by our bootstrap-based method have good coverage properties independent of the sample size.

Adaptability to different noise addition mechanisms. Both Figure 2 and Figure 3 show that the coverage properties of the bootstrap-based method are independent of

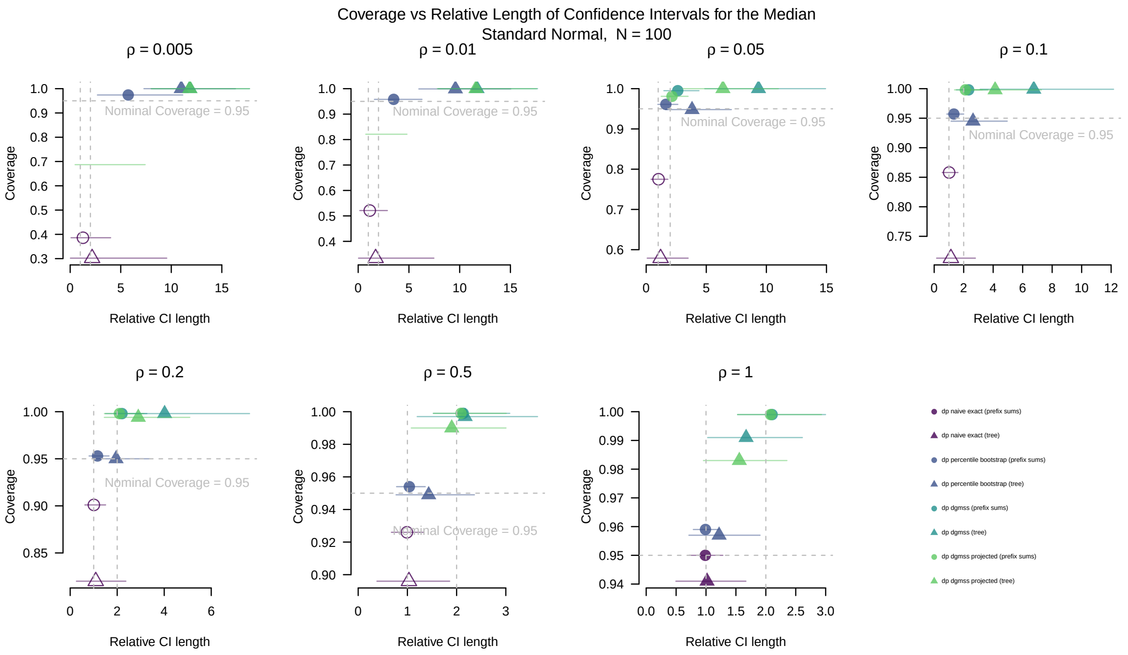

Figure 2: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with $N = 100$ drawn from a standard normal distribution. The symbols indicate the average relative length across 1000 repetitions in our Monte Carlo simulation. The horizontal bars indicate the range between the 2.5- and 97.5-percentile of relative length in these simulations. Each panel shows the results for a different value of the privacy parameter ρ. Hollow symbols indicate that the method did not achieve empirical coverage of at least 0.938 (the nominal confidence level is 0.95, accounting for Monte Carlo error. Any empirical coverage of less than 0.938 is statistically significantly less than 0.95).

a particular noise-addition mechanism $\mathcal{M}$. While the length of the confidence intervals depends on a particular noise-addition mechanism (as different amounts of noise have to be added for the different mechanisms), as long as the same mechanism is used on the original sample and for the bootstrap samples, the bootstrap method produces valid confidence intervals.

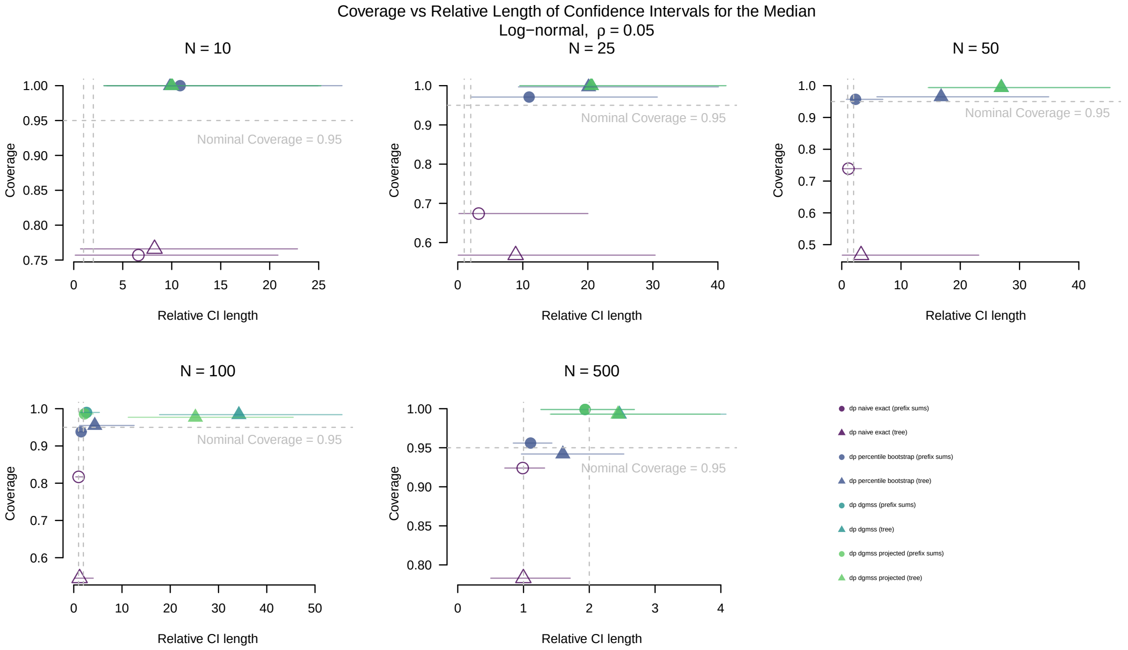

Figure 3: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different sample sizes drawn from a log-normal distribution. The symbols indicate the average relative length across 1000 repetitions in our Monte Carlo simulation. The horizontal bars indicate the range between the 2.5- and 97.5-percentile of relative length in these simulations. Each panel shows the results for a different sample size. Hollow symbols indicate that the method did not achieve empirical coverage of at least 0.938 (the nominal confidence level is 0.95, accounting for Monte Carlo error. Any empirical coverage of less than 0.938 is statistically significantly less than 0.95).

Appropriate coverage, independent of the privacy parameter and the significance level α. In Figure 4 we plot the significance level α against the observed empirical coverage. The different line types indicate different values of the privacy parameter ρ. The purple lines show our bootstrap-based method and the teal lines show the bespoke method. The diagonal indicates the perfect relationship between the significance level and empirical coverage. Our bootstrap-based method is close to the diagonal independent of the significance level α and independent of the level of privacy protection. In contrast, the bespoke method is underconfident, i.e., the empirical coverage is larger than would be expected by the significance level α.

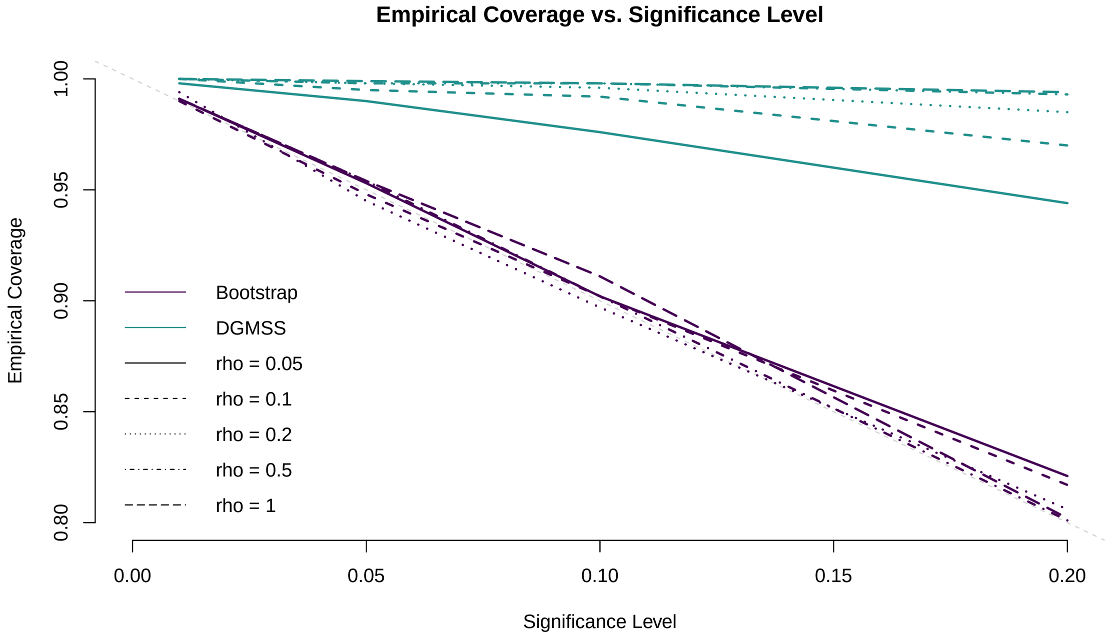

Figure 4: Significance level plotted against the empirical coverage rate. Comparing the bootstrap-based method and the bespoke method proposed by Drechsler et al. (2022) for different values of ρ.

5 Confidence Intervals for Multivariate Data

In this section, we show that our method can also be used with multivariate data, as long as a good differentially private mechanism $\mathcal{M}$ to estimate the multivariate data distribution $\tilde{P}$ exists. We start with a simple multivariate case where such a mechanism exists.

Consider a dataset with three binary variables X_1, X_2, Y, where X_1 and X_2 are independently drawn from a Bernoulli distribution and Y depends on X_1, X_2, and the true regression coefficients β. For our experiments, we set the true regression coefficients to $\beta_0 = -2$, $\beta_1 = 1$, and $\beta_2 = 1$.

The data-generating process can be described as follows:

$X_1 \sim \mathcal{B}(n, p = 0.5)$. $X_2 \sim \mathcal{B}(n, p = 0.5)$. $Y \sim \mathcal{B}(n, p = p_i)$, where $p_i = \frac{1}{1+e^{-X\beta}}$, with $X = \begin{pmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{1i} & X_{2i} \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$. The data matrix $D = \begin{pmatrix} Y_1 & X_{11} & X_{21} \\ Y_2 & X_{12} & X_{22} \\ \vdots & \vdots & \vdots \\ Y_i & X_{1i} & X_{2i} \end{pmatrix}$ has dimensions $n \times k$.

With three binary variables, the full data distribution can be summarized by a histogram with eight bins that describes a multinomial distribution[2].

To get a differentially private distribution estimate $\tilde{P}$, we produce a noisy histogram by adding independent zero-centered Gaussian noise to each histogram bin. The variance of the Gaussian distribution depends on the privacy parameter ρ such that $\sigma^2 = \frac{1}{2\rho}$, as the sensitivity of the histogram is 1[3].

5.1 Experimental Setup

Our goal is to produce valid confidence intervals for the coefficients of a logistic regression model. To evaluate the effect of the privacy noise we vary ρ and set it to the same values as in the univariate experiments in section 4.2 $(0.005, 0.01, 0.05, 0.1, 0.2, 0.5, 1)$. We also vary the sample size n and evaluate our method for samples with 100, 500, and 1000 observations. We compare our bootstrap-based confidence intervals to naive[4] differentially private and non-private confidence intervals for the same data.

5.2 Results

Our bootstrap-based method has good coverage properties for multivariate statistics of interest, independent of the privacy level ρ. In Figure 5 we show the empirical coverage for differentially private confidence intervals for the logistic regression

[2] Without privacy, this is $\hat{P}$

[3] As histogram bins can be negative after noise addition, we truncate any negative values to 0 to get a valid multinomial distribution. This introduces bias.

[4] For the naive method, we generate a synthetic dataset $\tilde{Y}$ deterministically from the noisy histogram and calculate the confidence intervals as if the differentially private synthetic data $\tilde{Y}$ were the original sample X.

coefficients across different values of the privacy parameter ρ. The significance level α is 0.05. Circles indicate our bootstrap-based confidence intervals, and triangles indicate naive differentially private confidence intervals. As in the univariate experiments in section 4.2, our bootstrap-based method produces confidence intervals with good empirical coverage (close to the nominal level) independent of ρ. The naive method does not produce valid confidence intervals for small values of ρ.

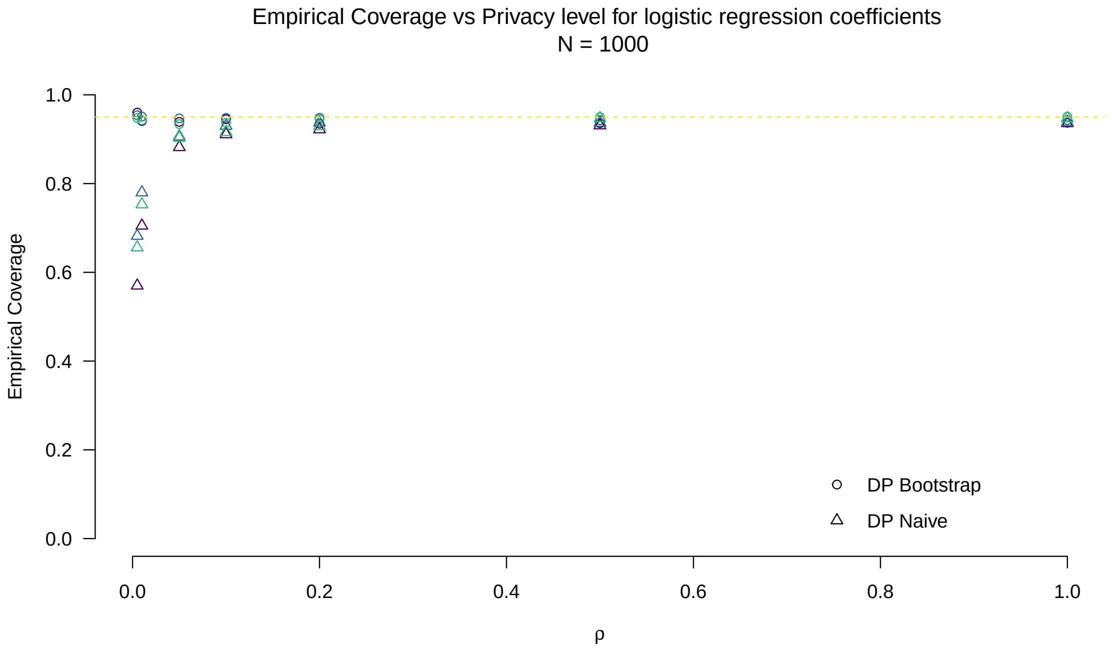

Figure 5: The privacy parameter ρ plotted against the empirical coverage of confidence intervals for logistic regression coefficients with a sample size of 1000. The colors indicate the three regression coefficients, and the shapes show two different methods to produce confidence intervals (circles our bootstrap-based method, and triangles a naive method).

Finally, in Figure 6, we evaluate the width of the confidence intervals of the bootstrap-based method relative to the non-private confidence interval for $n = 1000$.

As there is no bespoke method for confidence intervals of logistic regression coefficients, we compare the effect of the privacy noise to a decrease in the sample size in a non-private setting. For example, a $\rho = 0.5$ has a similar effect as having about 400 observations instead of 1000 observations in a non-private setting.

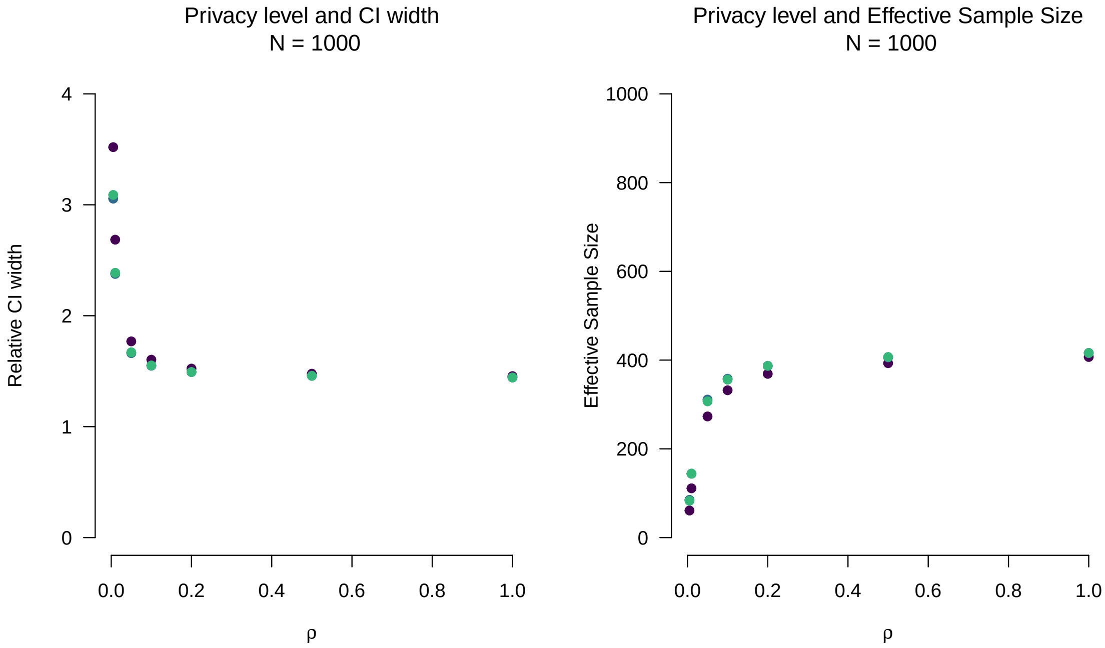

Figure 6: Left Panel: Confidence interval width relative to the width of non-private confidence intervals with $n = 1000$ for different values of the privacy parameter ρ. Right Panel: Effective sample size as a function of ρ. The colors indicate the three regression coefficients (and most of them overlap).

6 Conclusion

In this paper, we adapt a parametric bootstrap-based method to generate differentially private confidence intervals, showing that it applies in more general, nonparatmetric settings.

We show that our proposed method produces valid confidence intervals across a range of data-generating processes and settings of the privacy parameters ρ. Furthermore, comparing the confidence intervals from our method to a bespoke method for confidence intervals of the median with univariate data, we show that our method produces tighter confidence intervals at a given nominal coverage.

Our method assumes that the mechanism $\mathcal{M}$ produces an reasonable estimate $\tilde{P}$ of the underlying data-generating process P. While we show that such mechanisms exist for univariate data and simple multivariate data, we leave it to future research to evaluate the performance of more general differentially private distribution estimators (or synthesizers) such as graphical-model based approaches (McKenna et al. 2019).

References

Awan, J. & Slavković, A. (2018), 'Differentially private uniformly most powerful tests for binomial data', *Advances in Neural Information Processing Systems* **31**.

Bernstein, G. & Sheldon, D. R. (2018), Differentially private bayesian inference for exponential families, *in* S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 31, Curran Associates, Inc.

Brawner, T. & Honaker, J. (2018), 'Bootstrap inference and differential privacy: Standard errors for free', *Unpublished Manuscript* .

Bun, M. & Steinke, T. (2016), Concentrated differential privacy: Simplifications, extensions, and lower bounds, *in* 'Theory of Cryptography Conference', Springer, pp. 635–658.

Canonne, C. L., Kamath, G., McMillan, A., Smith, A. & Ullman, J. (2019), The structure of optimal private tests for simple hypotheses, *in* 'Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing', pp. 310–321.

Covington, C., He, X., Honaker, J. & Kamath, G. (2021), 'Unbiased statistical estimation and valid confidence intervals under differential privacy', *arXiv preprint arXiv:2110.14465* .

Dick, T., Dwork, C., Kearns, M., Liu, T., Roth, A., Vietri, G. & Wu, Z. S. (2023), 'Confidence-ranked reconstruction of census microdata from published statistics', *Proceedings of the National Academy of Sciences* **120**(8).

Drechsler, J., Globus-Harris, I., McMillan, A., Sarathy, J. & Smith, A. (2022), 'Nonparametric differentially private confidence intervals for the median', *Journal of Survey Statistics and Methodology* **10**(3), 804–829.

Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2016), ‘Calibrating noise to sensitivity in private data analysis’, *Journal of Privacy and Confidentiality* **7**(3). Originallay appeared in the proceedings of the 2006 *Theory of Cryptography Conference*.

Dwork, C. & Rothblum, G. N. (2016*a*), ‘Concentrated differential privacy’, *ArXiv* **abs/1603.01887**.

Dwork, C. & Rothblum, G. N. (2016*b*), ‘Concentrated differential privacy’, *arXiv preprint arXiv:1603.01887* .

Efron, B. & Hastie, T. (2021), *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, Vol. 6, Cambridge University Press.

Evans, G., King, G., Schwenzfeier, M. & Thakurta, A. (2020), ‘Statistically valid inferences from privacy protected data’, *URL: GaryKing. org/dp* .

Ferrando, C., Wang, S. & Sheldon, D. (2022), Parametric bootstrap for differentially private confidence intervals, *in* ‘International Conference on Artificial Intelligence and Statistics’, PMLR, pp. 1598–1618.

Fichtenberger, H., Henzinger, M. & Upadhyay, J. (2022), ‘Constant matters: Fine-grained complexity of differentially private continual observation’.
URL: *https://arxiv.org/abs/2202.11205*

Gaboardi, M., Lim, H., Rogers, R. & Vadhan, S. (2016), Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing, *in* ‘International conference on machine learning’, PMLR, pp. 2111–2120.

Honaker, J. (2015), ‘Efficient use of differentially private binary trees’.

Karwa, V. & Vadhan, S. (2017), ‘Finite sample differentially private confidence intervals’, *arXiv preprint arXiv:1711.03908* .

Kasiviswanathan, S. P. & Smith, A. D. (2008), ‘On the ‘semantics’ of differential privacy: A bayesian formulation’, *CoRR* **abs/0803.3946**.

Kulkarni, T., Jälkö, J., Koskela, A., Kaski, S. & Honkela, A. (2021), Differentially private bayesian inference for generalized linear models, *in* M. Meila & T. Zhang, eds, ‘Proceedings of the 38th International Conference on Machine Learning’, Vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 5838–5849.
URL: *https://proceedings.mlr.press/v139/kulkarni21a.html*

Li, C., Miklau, G., Hay, M., McGregor, A. & Rastogi, V. (2015), 'The matrix mechanism: optimizing linear counting queries under differential privacy', *The VLDB Journal* **24**(6), 757–781.
URL: *https://doi.org/10.1007/s00778-015-0398-x*

McKenna, R., Sheldon, D. & Miklau, G. (2019), Graphical-model based estimation and inference for differential privacy, *in* K. Chaudhuri & R. Salakhutdinov, eds, 'Proceedings of the 36th International Conference on Machine Learning', Vol. 97 of *Proceedings of Machine Learning Research*, PMLR, pp. 4435–4444.
URL: *https://proceedings.mlr.press/v97/mckenna19a.html*

Nissim, K., Raskhodnikova, S. & Smith, A. D. (2007), Smooth sensitivity and sampling in private data analysis, *in* 'Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA', pp. 75–84.

Räisä, O., Jälkö, J., Kaski, S. & Honkela, A. (2023), Noise-aware statistical inference with differentially private synthetic data, *in* F. Ruiz, J. Dy & J.-W. van de Meent, eds, 'Proceedings of The 26th International Conference on Artificial Intelligence and Statistics', Vol. 206 of *Proceedings of Machine Learning Research*, PMLR, pp. 3620–3643.
URL: *https://proceedings.mlr.press/v206/raisa23a.html*

Smith, A. (2011), Privacy-preserving statistical estimation with optimal convergence rates, *in* 'Proceedings of the forty-third annual ACM symposium on Theory of computing', pp. 813–822.

Vu, D. & Slavkovic, A. (2009), Differential privacy for clinical trial data: Preliminary evaluations, *in* '2009 IEEE International Conference on Data Mining Workshops', IEEE, pp. 138–143.

Wang, Y., Lee, J. & Kifer, D. (2015), 'Revisiting differentially private hypothesis tests for categorical data', *arXiv preprint arXiv:1511.03376* .

Wang, Z., Cheng, G. & Awan, J. (2022), 'Differentially private bootstrap: New privacy analysis and inference strategies', *arXiv preprint arXiv:2210.06140* .

A Additional Results: Univariate Data

A.1 Standard Normal DGP

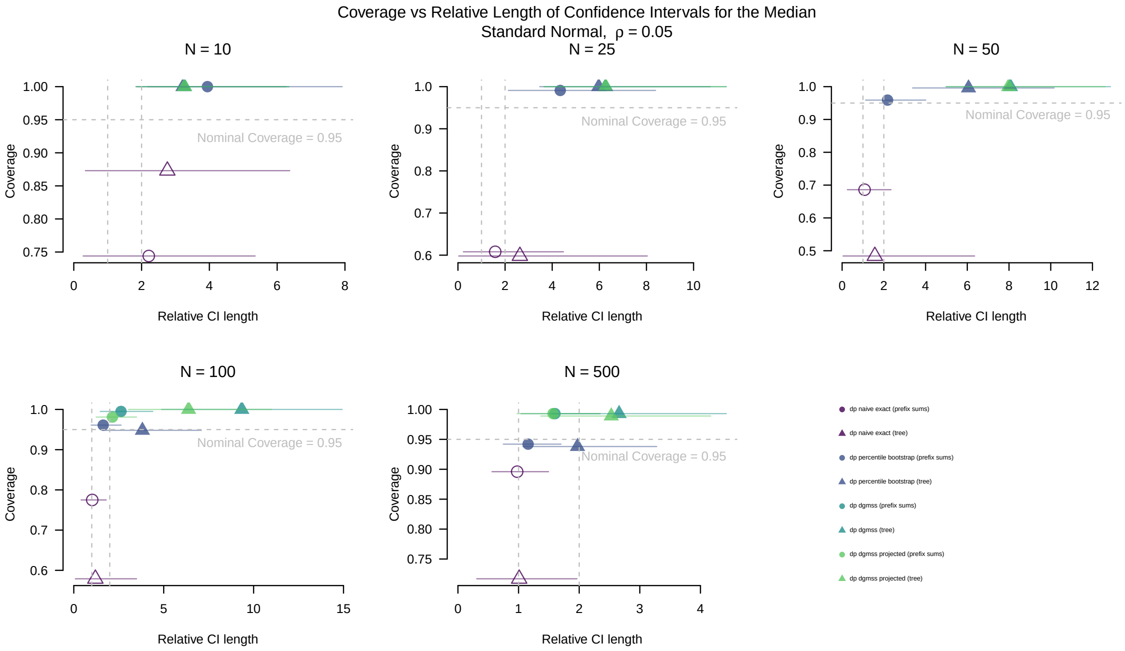

Figure 7: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different sample sizes drawn from a standard normal distribution with $\rho = 0.05$.

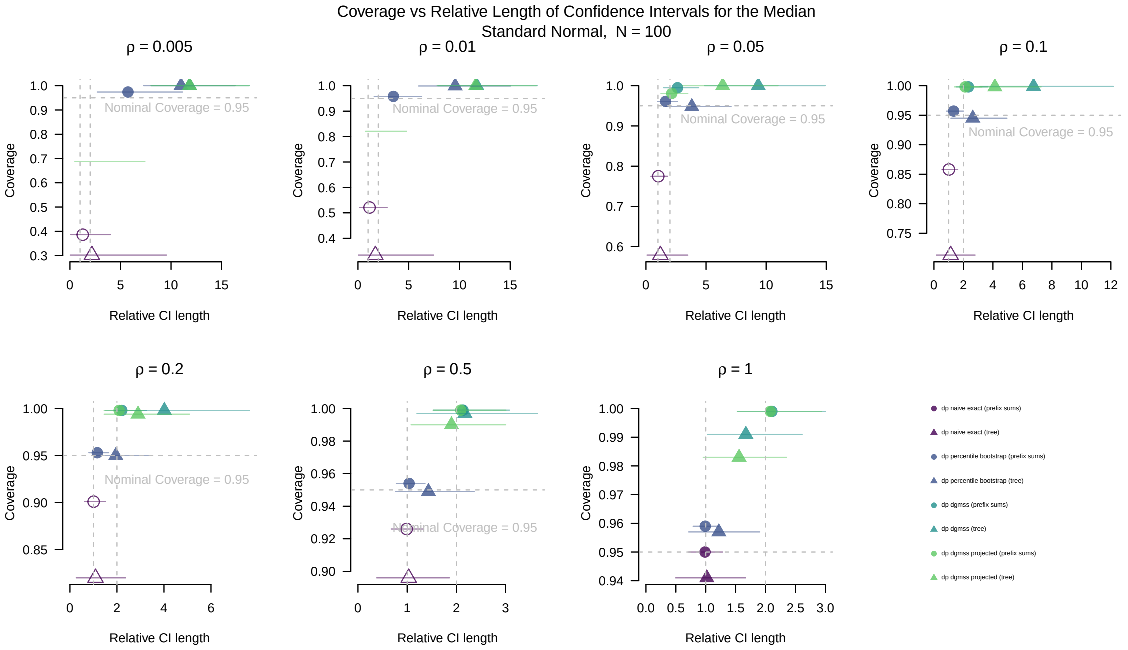

Figure 8: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different values of ρ drawn from a standard normal distribution with $n = 100$.

A.2 Log-normal DGP

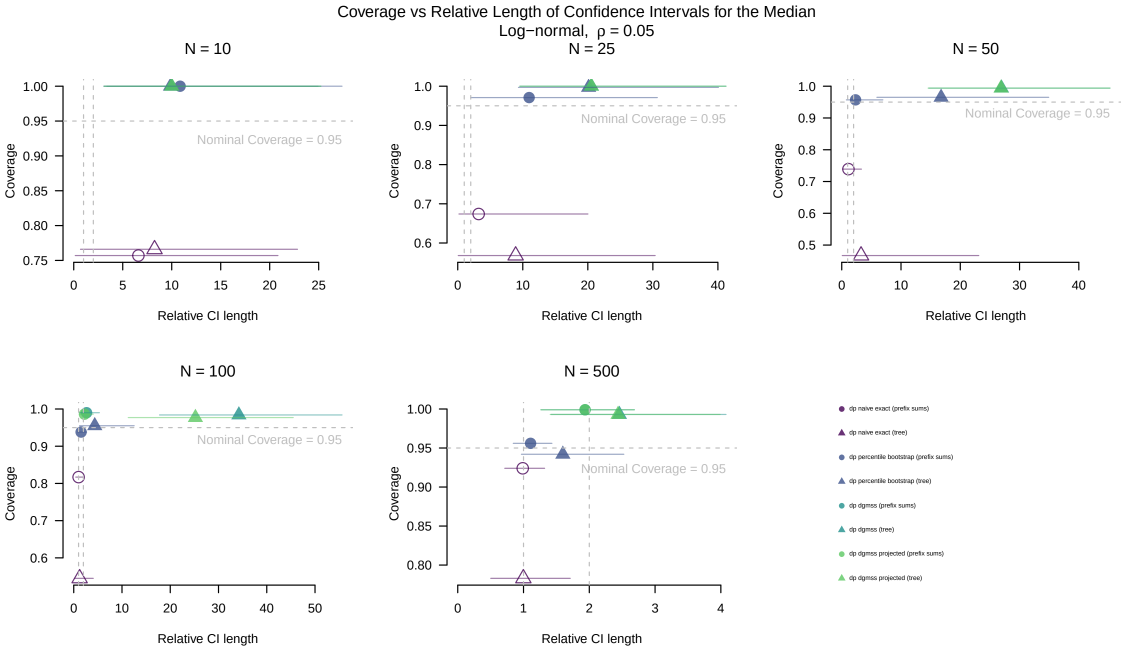

Figure 9: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different sample sizes drawn from a log-normal distribution with $\rho = 0.05$.

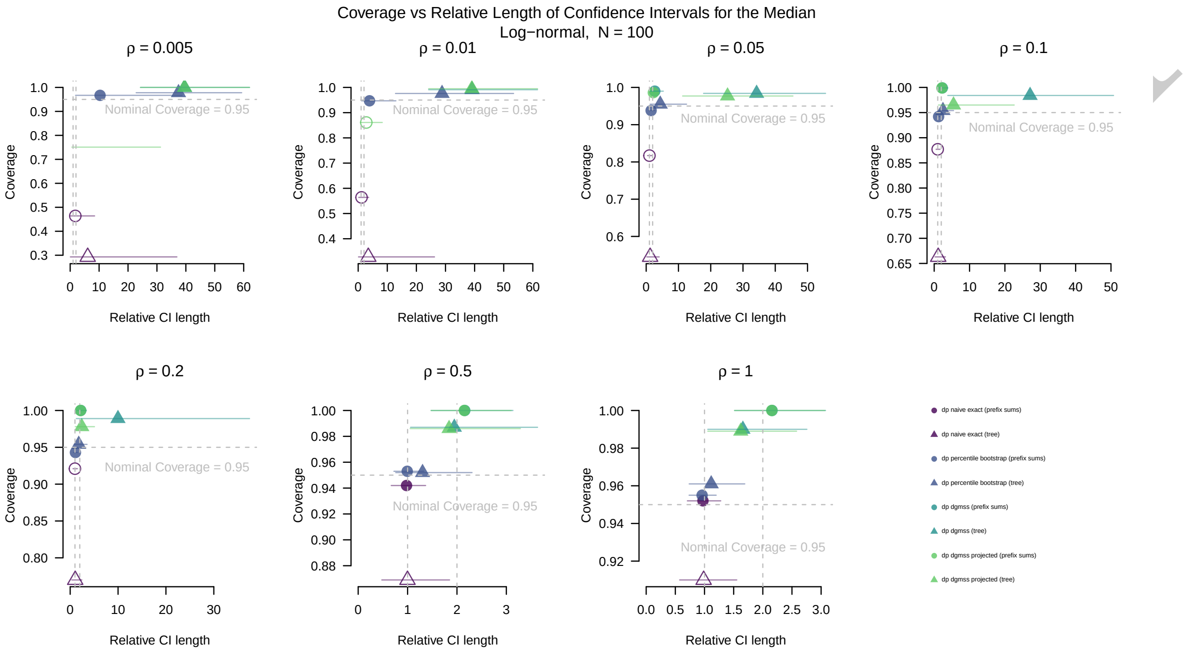

Figure 10: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different values of ρ drawn from a log-normal distribution with $n = 100$.

A.3 Bimodal DGP

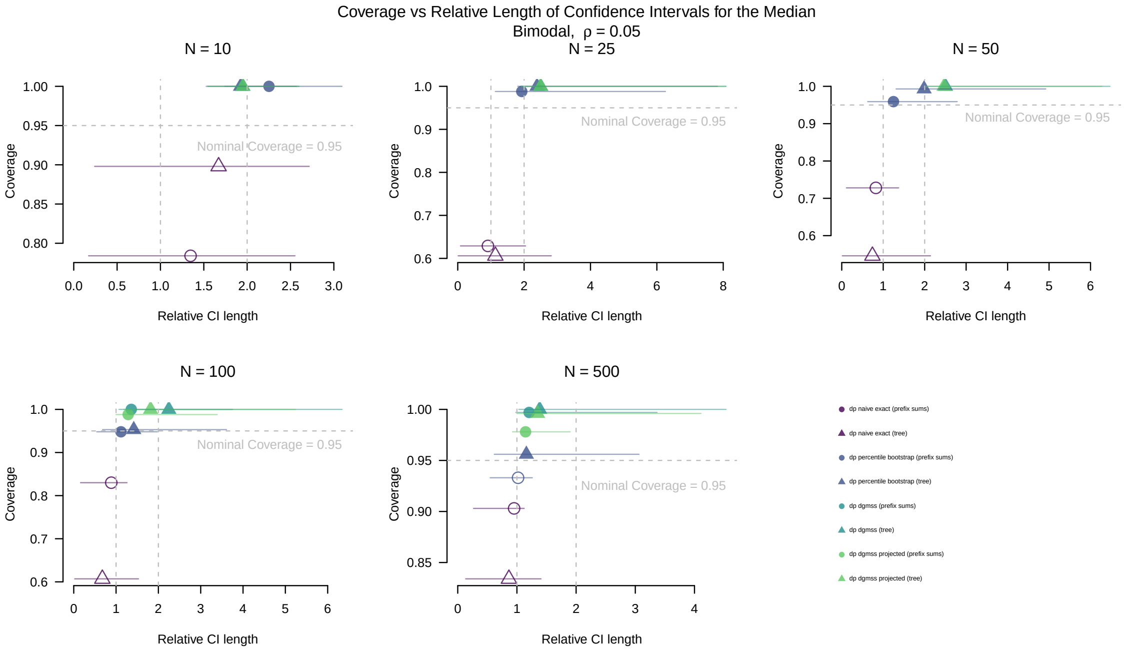

Figure 11: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different sample sizes drawn from a bimodal distribution with $\rho = 0.05$.

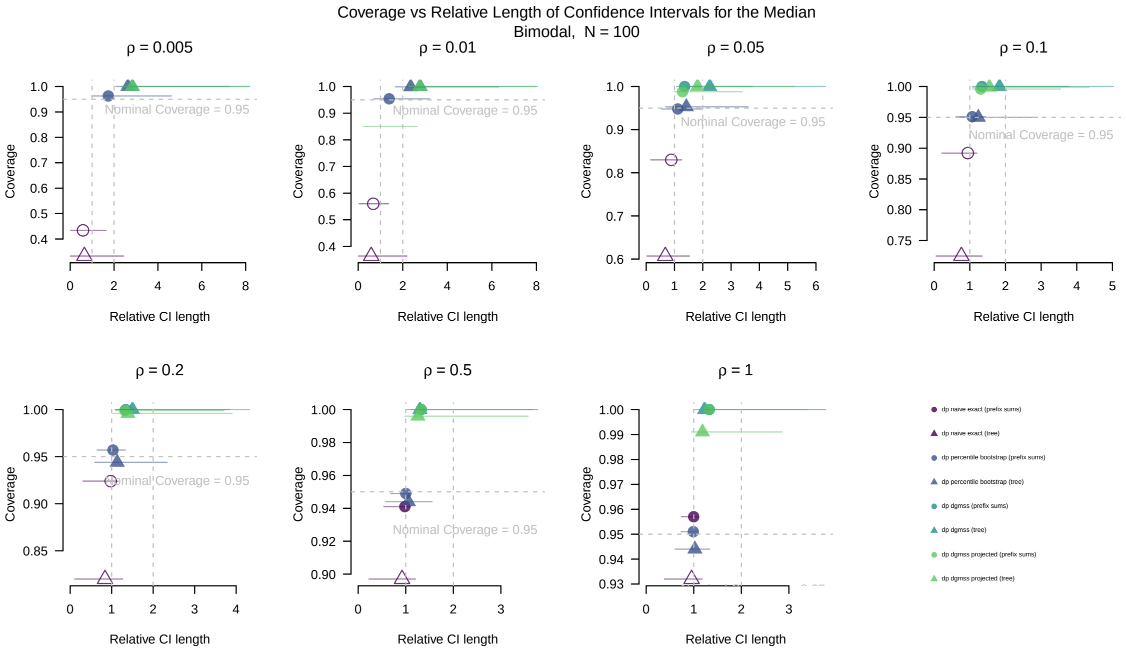

Figure 12: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different values of ρ drawn from a bimodal distribution with $n = 100$.

A.4 Adult Age DGP

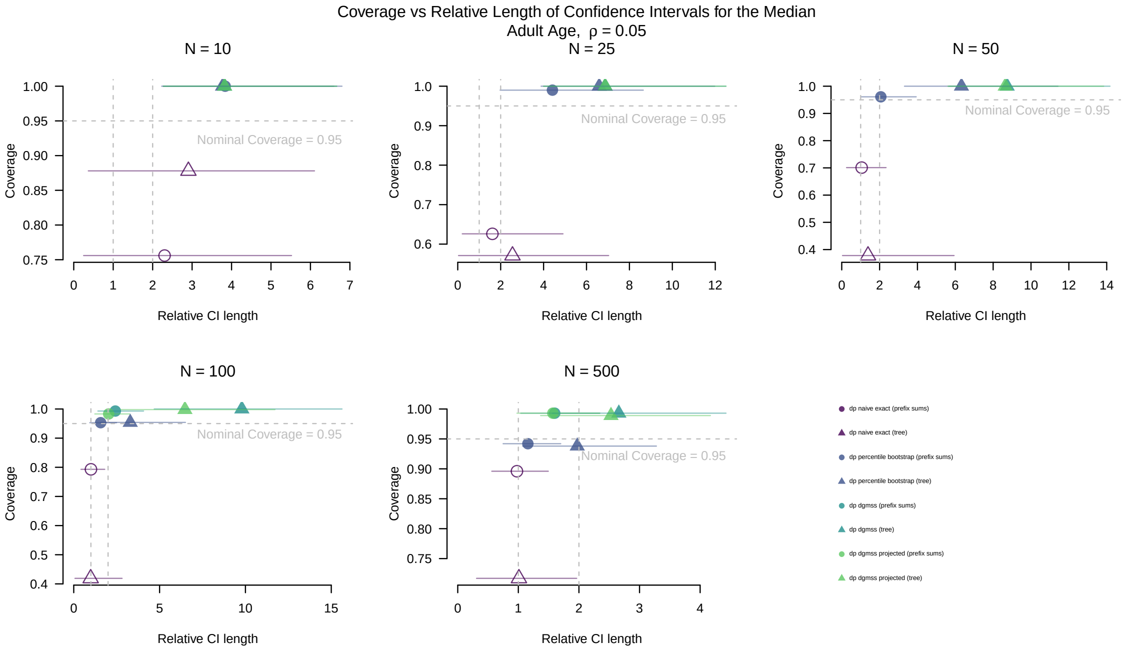

Figure 13: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different sample sizes drawn from the adult age population with $\rho = 0.05$.

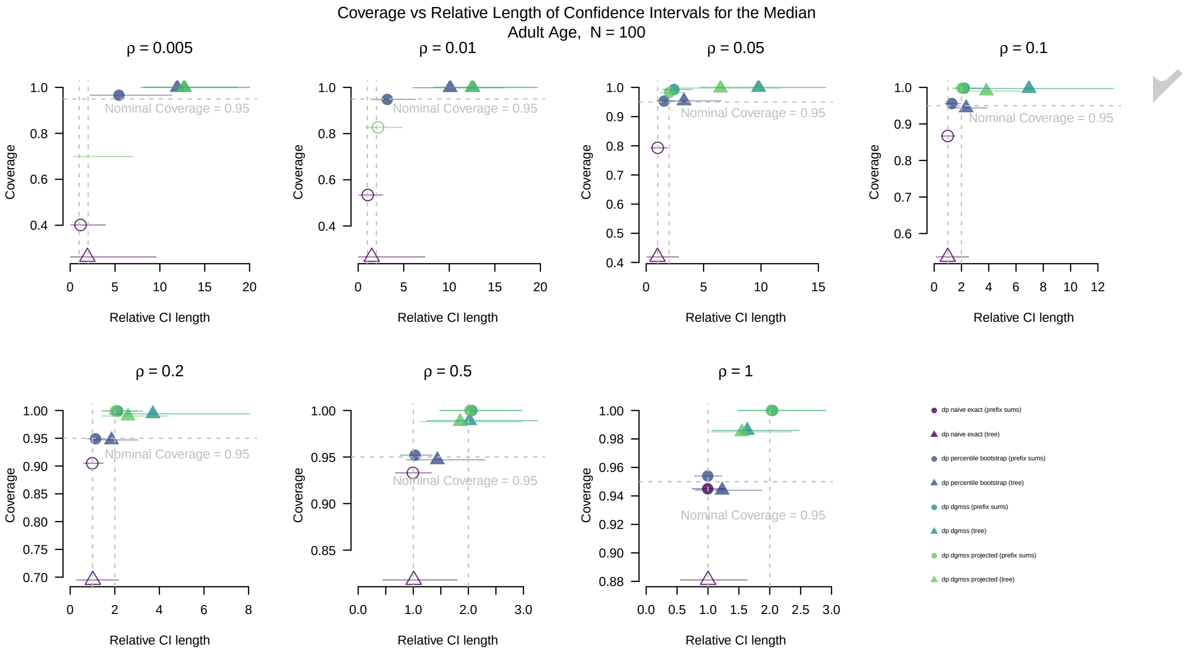

Figure 14: Relative confidence interval length plotted against the empirical coverage of confidence intervals for the median of samples with different values of ρ drawn from the adult age population with $n = 100$.