2022-06-28

Nonparametric differentially private confidence intervals for the median

*Journal of Survey Statistics and Methodology (2022) **10**, 804–829*

NONPARAMETRIC DIFFERENTIALLY PRIVATE CONFIDENCE INTERVALS FOR THE MEDIAN

JÖRG DRECHSLER
IRA GLOBUS-HARRIS
AUDRA MCMILLAN*
JAYSHREE SARATHY
ADAM SMITH

Differential privacy is a restriction on data processing algorithms that provides strong confidentiality guarantees for individual records in the data. However, research on proper statistical inference, that is, research on properly quantifying the uncertainty of the (noisy) sample estimate

Jörg Drechsler is a Distinguished Researcher at the Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg, Germany, and Associate Research Professor in the Joint Program in Survey Methodology at the University of Maryland, College Park, Maryland 20742, USA. Ira Globus-Harris is a PhD student at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. Audra McMillan is a Research Scientist at Apple, 1 Apple Park Way, Cupertino, California 95014, USA. Jayshree Sarathy is a PhD student at the Harvard John A. Paulson School of Engineering and Applied Sciences, 150 Western Ave, Boston, Massachusetts 02134, USA. Adam smith is Professor at the Department of Computer Science, Boston University, Boston, Massachusetts 02215, USA.

Part of this work was completed while the author Audra McMillan was at Boston University and Northeastern University.

The work of Drechsler, Globus-Harris, Sarathy, and Smith on this project was funded in part by United States Census Bureau cooperative agreements CB16ADR0160001 and CB20ADR0160001. The work of McMillan (while at Boston University) and Smith was also supported in part by National Science Foundation award CCF-1763786 as well as a Sloan Foundation research award. Part of this work was done while McMillan was supported by a Fellowship from the Cybersecurity & Privacy Institute at Northeastern University and National Science Foundation grant CCF-1750640. Globus-Harris' work at Boston University was supported by funding from the Hariri Institute for Computing and was supported by the Computer and Information Sciences PhD Graduate Fellowship at University of Pennsylvania. The opinions, findings, conclusions, and recommendations expressed herein are those of the authors and do not reflect the views of the United States Census Bureau or other funding sources.

Our work was prompted in part by discussions with sociologists John Logan and Brian Stults, in the context of their work on integrating data across time-varying tract boundaries (Logan, Zhang, Stults, and Gardner 2021). We are also grateful for helpful conversations with and comments from (in no particular order) Rolando Rodriguez, Ryan Cummings, Thomas Steinke, Shurong Lin, Eric Kolaczyk, and Salil Vadhan.

*Address correspondence to Audra McMillan, Apple, USA; E-mail: audra.mcmillan@apple.com.

https://doi.org/10.1093/jssam/smac021

regarding the true value in the population, is currently still limited. This article proposes and evaluates several strategies to compute valid differentially private confidence intervals for the median. Instead of computing a differentially private point estimate and deriving its uncertainty, we directly estimate the interval bounds and discuss why this approach is superior if ensuring privacy is important. We also illustrate that addressing both sources of uncertainty—the error from sampling and the error from protecting the output—simultaneously should be preferred over simpler approaches that incorporate the uncertainty in a sequential fashion. We evaluate the performance of the different algorithms under various parameter settings in extensive simulation studies and demonstrate how the findings could be applied in practical settings using data from the 1940 Decennial Census.

KEYWORDS: Robust; Statistical inference; Confidentiality; Disclosure limitation.

Statement of Significance

Differential privacy is a restriction on data processing algorithms that provides strong confidentiality guarantees. However, research on properly quantifying the uncertainty of the (noisy) sample estimate is still limited. This article proposes and evaluates several strategies to compute valid differentially private confidence intervals for the median. Instead of computing a differentially private point estimate and deriving its uncertainty, we directly estimate the interval bounds and discuss why this approach is superior in this context. We also illustrate that addressing the error from sampling and the error from protecting the output simultaneously should be preferred over simpler approaches.

1. INTRODUCTION

Statistical agencies constantly need to find the right balance between the two competing goals of disseminating useful information from their collected data and ensuring the confidentiality of the units included in the database. Many methods have been developed in the past decades to address this tradeoff. However, with the advent of modern computing and the massive amounts of data collected every day, many of the data protection strategies commonly used at statistical agencies are no longer adequate to protect the data (Abowd 2018; Garfinkel, Abowd, and Martindale 2019). The problem's difficulty is amplified by the continual appearance of new data sources that facilitate attacks.

One promising strategy to circumvent this dilemma is to rely on formal privacy guarantees such as those provided by differential privacy (DP) (Dwork, McSherry, Nissim, and Smith 2006b). These guarantees hold no matter what background knowledge a potential attacker might possess, or how much computational power they have. However, methodology for differential private statistical inference has mostly been studied from a theoretical perspective under asymptotic regimes. Although many algorithms have been proposed to ensure formal privacy guarantees for various estimation tasks, evaluations of their relative performance on real data with limited sample sizes and complex distributional properties are still limited, and only a small fraction of that literature has focused on inference and associated measures of uncertainty. Section 2.3 surveys related work.

In this article, we address these issues, focusing on one of the key measures of location: the median. We chose the median for two reasons. On one hand, it is a widely used summary statistic for skewed variables such as income (see, e.g., the U.S. Census Bureau's tables of median incomes for various subgroups of the population; U.S. Census Bureau 2020a). On the other hand, medians provide an interesting technical challenge for differentially private computation. The accuracy of differentially private median computations depends on the exact data distribution; as a result, providing sound and narrow confidence intervals appears to require releasing strictly more information about the data than is required for point estimation.

The discussion of confidence intervals is an important contribution of our study. None of the previously proposed algorithms for DP median estimation come equipped with a method for additionally releasing DP uncertainty estimates on the point estimator. In fact, as pointed out above the level of uncertainty in the point estimate is typically data dependent, and hence measuring it requires additional privacy budget. Thus, the optimal algorithm for differentially private point estimates can be different from the optimal algorithm for differentially private confidence intervals. Instead of deriving the variance of some differentially private point estimate, we suggest estimating DP confidence intervals directly. We show that our proposed methodology ensures proper confidence interval coverage in a frequentist sense and discuss why this strategy requires less privacy budget than starting from the protected point estimates.

When designing and analyzing differentially private algorithms it is tempting to separate the error due to sampling from the error due to privacy and bound the two separately. A main finding in our work is the limitation of this approach. We find that one can obtain considerably tighter confidence intervals by analyzing the relationship between the two sources of error. Unlike approaches which treat the analysis of the nonprivate algorithm as a black-box, this involves looking at the different ways that the sampling error can result in the confidence interval failing to capture the median, and considering how the error due to privacy affects each of these modalities.

We assume simple random sampling throughout the article. This assumption is often violated in survey practice. However, understanding the implications of complex sampling designs on the privacy guarantees is an open research problem (Drechsler 2021) and we are not aware of any DP applications that take complex sampling designs into account. We see our contribution as an important first step toward the goal of better serving the needs of statistical agencies, while acknowledging the limitations of the current findings.

We evaluate several algorithms for computing differentially private confidence intervals. We discuss algorithms that satisfy two versions of DP: the strictest version (Dwork et al. 2006b), now known as *pure differential privacy*, as well as a slight relaxation, *concentrated differential privacy* (CDP) (Bun and Steinke 2016; Dwork and Rothblum 2016). The focus of our study is on empirical evaluation, using a mix of simulated and real data. Nevertheless, we found that new methodology and theory was also needed to adapt existing algorithms for confidence interval computation. We include an application using data from the U.S. Census 1940 to illustrate how statistical agencies willing to adopt the methodology could decide which algorithm and parameter settings to pick for their data release.

The algorithms we developed are all *sound* in the nonparametric, frequentist sense: when run with nominal coverage $1 - \alpha$ the probability that the true population median is contained in the computed confidence interval is at least $1 - \alpha$, where the probability is taken over the entire process of sampling from the population and computing the private confidence intervals based on the drawn sample. Rubin (1996) terms this property *confidence validity* to distinguish it from *randomization validity*, which would require that the actual coverage rate matches the nominal coverage rate exactly (see Rubin 1996 for a discussion why the former should generally be preferred over the latter). Since all algorithms rely on nonparametric strategies for computing the confidence intervals, the proposed intervals are confidence valid for every i.i.d distribution on observations.[1]

Using nonparametric approaches is especially important under privacy constraints. This is because the typical safeguards around using parametric assumptions—such as visualizing the data and running goodness-of-fit tests—require using some of the privacy budget. Therefore, any potential gains in accuracy from using parametric models will likely be lost as more noise will have to be infused in the final output. However, without checking the parametric assumptions the analyst risks that the algorithms will produce misleading uncertainty measures. We illustrate this point in section 5, where we demonstrate that erroneously assuming a log-normal distribution for the income data used in our application will give severely biased results.

1. We will use the term *valid confidence interval* in the remainder of the article, whenever the interval is confidence valid in Rubin's sense.

We find that a specific algorithm, a variant of the exponential mechanism (McSherry and Talwar 2007), is the best choice across a range of settings. We also study in depth another DP algorithm, which provides slightly wider confidence intervals but has additional practical benefits.

The remainder of the article is organized as follows: In section 2, we present background on DP and nonprivate confidence intervals. In section 3, we discuss the design of private confidence intervals and give an overview of the two algorithms to be considered. In section 4, we present the results from extensive simulation studies that evaluate the performance of the algorithms under various parameter settings. Section 5 illustrates how the methodology could be applied in practice by replicating one of the income tables published by the U.S. Census Bureau using publicly available data from the 1940 U.S. Census. The notation defined in this section and elsewhere in the study is summarized in table S1 of the supplementary material online.

2. PRELIMINARIES

2.1 Confidence Intervals for the Median

Throughout this article we will refer to the median of the underlying population P as the *population median*, denoted $\texttt{med}(P)$, and the median of a given sample as the *sample median*. Let $\mathcal{P} \subset \Delta(\mathbb{R})$ be the set of possible population distributions over the data domain $\mathbb{R}$. Let $I_{\mathbb{R}}$ be the set of intervals in $\mathbb{R}$. Given a failure probability $\alpha \in [0, 1]$ and database size $n \in \mathbb{N}$, we say that a function $M : \mathcal{X}^n \to I_{\mathbb{R}}$ achieves the *nominal coverage* rate $1 - \alpha$ for the set $\mathcal{P}$ if for all $P \in \mathcal{P}$, $\Pr(\texttt{med}(P) \in M(y)) \geq 1 - \alpha$, where the randomness is taken over both the randomness in M and the randomness in the sample $y \sim P^n$. We will refer to $\Pr(\texttt{med}(P) \in M(y))$ as the *actual coverage* of M on database y.

Even if one is not concerned with privacy, researchers often prefer nonparametric approaches when computing confidence intervals for the median to avoid parametric assumptions that are seldom met in practice. A common nonparametric confidence interval for the median is computed using the order statistics of the sample (Lehmann and D'Abrera 1975, theorem 5, p. 182) If P has continuous CDF then the rank of the population median $\texttt{med}(P)$ in a simple random sample of size n drawn from P is distributed as the binomial $\text{Bin}(n, 1/2)$.[2] For a dataset $y \in \mathbb{R}^n$, let $y_{(k)}$ be the k-th order statistic (the k-th smallest value in y).

Lemma 2.1 (Nonprivate $(1 - \alpha)$-confidence interval). *Let F_{Bin} be the CDF of the binomial random variable $\text{Bin}(n, 1/2)$, $\mathtt{N}_L^\alpha = \max_{m \in \mathbb{N}}\{m | F_{\text{Bin}}(m-1) \leq \alpha/2\}$ and $\mathtt{N}_U^\alpha = \min_{m \in \mathbb{N}}\{m | F_{\text{Bin}}(m+1) \geq 1-\alpha/2\}$. For any dataset $y \in \mathbb{R}^n$,*

2. The assumption that P has continuous CDF can be made with essentially no loss of generality—see supplementary material online C.

let $\texttt{ci}_L^\alpha(y) = y_{(\mathrm{N}_L^\alpha)}$ *and* $\texttt{ci}_U^\alpha(y) = y_{(\mathrm{N}_U^\alpha)}$. *Then the function* $M : \chi^n \to I_{\mathbb{R}}$, *from the set of size* n *datasets to the set of intervals in* $\mathbb{R}$, *given by* $M(y) = [\texttt{ci}_L^\alpha(y), \texttt{ci}_U^\alpha(y)]$ *achieves nominal coverage* $1-\alpha$ *for* $\Delta(\mathbb{R})$ *(the set of all distributions on* $\mathbb{R}$*).*

This approach is fully nonparametric, which—as discussed in the introduction—is especially important given privacy constraints. Note that a confidence interval does not directly output a point estimate for the median itself. In the absence of privacy constraints, one can simply additionally release the sample median $\texttt{med}(y)$ as a point estimate for the population median. However, under privacy constraints, rather than allocating some of the privacy budget to providing a point estimate of the median, it is often preferable to allocate the entire budget to estimating the confidence interval, then use the midpoint of that interval as a point estimate of the median.

2.2 Differential Privacy

Since our algorithms often include hyperparameters, we state a definition of DP for algorithms that take as input not only the dataset, but also the desired privacy parameters and any required hyperparameters. Let $\mathcal{X}$ be a data universe (e.g., $\mathbb{R}$ for medians) and $\mathcal{X}^n$ be the space of datasets of size n. Two datasets $y, y' \in \mathcal{X}^n$ are neighboring, denoted $y \sim y'$, if they differ on a single record. Let $\mathcal{H}$ be the space of hyperparameters and $\mathcal{Y}$ be an output space. To build some intuition, let us first define pure and approximate DP.

Definition 2.1((ε, δ)-DP (Dwork, Kenthapadi, McSherry, Mironov, and Naor 2006a; Dwork et al. 2006b). *Given privacy parameters* $\varepsilon \geq 0$ *and* $\delta \in [0, 1]$, *a randomized mechanism* $M : \mathcal{X}^n \times \mathcal{H} \to \mathcal{Y}$ *is* (ε, δ)*-differentially private if for all datasets* $\boldsymbol{y} \sim \boldsymbol{y}' \in \mathcal{X}^n$, *hyperparams* $\in \mathcal{H}$, *and events* $E \subseteq \mathcal{Y}$,

$$\Pr[M(y, \mathit{hyperparams}) \in E] \leq e^\varepsilon \cdot \Pr[M(y', \mathit{hyperparams}) \in E] + \delta,$$

where the probabilities are taken over the randomness induced by M.

The key intuition for this definition is that the distribution of outputs on input dataset y is almost indistinguishable from the distribution of outputs on input dataset y'. Therefore, given the output of a differentially private mechanism, it is impossible to confidently determine whether the input dataset was y or y'. If $\delta = 0$, then we refer to this as ε*-pure differential privacy*. If $\delta > 0$, we refer to (ε, δ)*-approximate differential privacy*. For strong privacy guarantees, the privacy-loss parameter is typically taken to be a small constant <1 (note that $e^\varepsilon \approx 1 + \varepsilon$ as $\varepsilon \to 0$). However, in practice, larger values of ε are occasionally used to satisfy utility constraints while providing some level of nontrivial privacy guarantee.

The algorithms in this article actually satisfy a version of DP called *concentrated differential privacy* (CD). This notion of privacy lies between the more common notions of *pure differential privacy* and *approximate differential privacy*. While still satisfying a rigorous notion of privacy, this will allow our algorithms to be significantly more accurate than their corresponding purely differentially private counterparts. For most of our algorithms little accuracy is gained from transitioning to approximate DP. Additionally, CDP has the desirable property of being a one-parameter property, which allows for simpler privacy accounting. For a further discussion of concentrated DP, see supplementary material online B.

Our goal is to design differentially private algorithms that achieve the nominal coverage rate. That is, they output an interval that captures the true parameter with probability at least $1 - \alpha$, where the probability includes both the sampling uncertainty and any randomness in the algorithm.

2.3 Related Work

Computing confidence intervals for the median is one of the most fundamental statistical tasks. However, finding a differentially private estimator for this task that is accurate across a range of datasets and parameter regimes is surprisingly nuanced. There has been a significant amount of prior work on differentially private point estimators for the median (Nissim, Raskhodnikova, and Smith 2007; Bun and Steinke 2019; Asi and Duchi 2020; Alabi, McMillan, Sarathy, Smith, and Vadhan 2020; Tzamos, Vlatakis-Gkaragkounis, and Zadik 2020) and other quantiles (Gillenwater, Joseph, and Kulesza 2021). To the best of our knowledge, none of these works addressed DP confidence intervals for the median. However, there has been significant work on DP confidence intervals for other estimation tasks like (Gaussian or sub-Gaussian) mean estimation (Karwa and Vadhan 2018; Gaboardi, Rogers, and Sheffet 2019; Du, Foot, Moniot, Bray, and Groce 2020; Biswas, Dong, Kamath, and Ullman 2020), and linear regression (Barrientos, Reiter, Machanavajjhala, and Chen 2019; Evans and King 2022). There are also several works on designing more general DP confidence intervals using bootstrapping, or a technique called subsample-and-aggregate (Nissim et al. 2007), to account for the combined uncertainty from sampling and noise due to privacy (D'Orazio, Honaker, and King 2015; Brawner and Honaker 2018; Barrientos et al. 2019; Ferrando, Wang, and Sheldon 2020; Evans, King, Schwenzfeier, and Thakurta 2021). These algorithms typically require a parametric model on the data or a normality assumption on the quantity being estimated; neither hold in our setting.

The areas of differentially private Bayesian inference (Dimitrakakis, Nelson, Mitrokotsa, and Rubinstein 2014; Wang, Fienberg, and Smola 2015b; Foulds, Geumlek, Welling, and Chaudhuri 2016; Heikkilä, Lagerspetz, Kaski, Shimizu, Tarkoma, et al. 2017; Bernstein and Sheldon 2018, 2019; Gong

2019) and hypothesis testing (Vu and Slavkovic 2009; Wang, Lee, and Kifer 2015a; Gaboardi, Lim, Rogers, and Vadhan 2016; Degue and Ny 2018; Couch, Kazan, Shi, Bray, and Groce 2019) study related problems of quantifying uncertainty, but specific goals differ. Wang (2018), Du et al. (2020) and Biswas et al. (2020) perform experimental evaluations of DP confidence intervals, however they focus on different estimators (linear regression and mean estimation) and focus on large datasets of at least 1,000 observations.

To the best of our knowledge, our work is unique in focusing on nonparametric differentially private confidence intervals for the median. This approach allows us to define algorithms that provide accurate and private confidence intervals without requiring distributional assumptions on the underlying population.

3. DESIGNING DP CONFIDENCE INTERVALS

While there are several algorithms in the literature for privately estimating order statistics, there is no straightforward way to extend these algorithms to release a confidence interval. Providing such a measure of uncertainty for the sample median is especially important for differentially private statistics since randomness in the algorithm provides an additional source of uncertainty.

Naïvely, one might hope to obtain a private confidence interval by simply privately estimating the order statistics described in lemma 2.1 and adding a data independent, fixed-width interval around the order statistic point estimates. The issue is that differentially private algorithms for order statistics do not (and cannot) operate by adding data-independent noise to the statistics[3]; thus, there is no single-fixed width interval that will be generally valid for privately estimated order statistics. Privately releasing an estimate of the amount of uncertainty introduced by algorithms that add data-dependent noise is challenging since it may depend subtly on the entire data distribution. This will be evident in the algorithms described in section 3.2.

3.1 Accounting for All Sources of Randomness

Accurate and tight coverage analysis is a crucial component of designing good algorithms. Overly conservative coverage estimates can result in confidence intervals that are wider than necessary. Differentially private confidence intervals need to account for two sources of error; sampling error and error due to privacy. Sampling error, also present in the nonprivate context, captures how

3. Algorithms that enforce privacy constraints by simply adding noise to the nonprivate estimate must add noise whose standard deviation is roughly proportional to how much the statistic of interest can vary on *worst case* neighboring datasets. Since order statistics can be very sensitive to the input dataset, this means any data independent noise addition method will result in noise that will overwhelm the signal (Nissim et al.2007).

well the realized sample y represents the underlying population P. The error due to privacy takes into account the additional randomness in M as a result of the privacy guarantee. Our experimental results highlight that it is important to carefully exploit the dependence between the two sources of randomness.

As a primer, let us first consider the coverage analysis of the nonprivate algorithm described in lemma 2.1. This coverage analysis relies on the fact that if P is continuous then for all $m \in [n]$,

$$\Pr(\texttt{rank}_y(\texttt{med}(P)) = m) = \Pr(\text{Bin}(n, 1/2) = m).$$

There are two ways that the interval $[\texttt{ci}_L^{\alpha}(y), \texttt{ci}_U^{\alpha}(y)]$ can fail to capture $\texttt{med}(P)$; $\texttt{med}(P) < \texttt{ci}_L^{\alpha}(y)$ or $\texttt{med}(P) > \texttt{ci}_U^{\alpha}(y)$. Let us focus on the probability of the first type of failure, $\texttt{med}(P) < \texttt{ci}_L^{\alpha}(y)$. For every $P \in \Delta_{\mathcal{C}}(\mathbb{R})$,

$$\Pr(\texttt{med}(P) < \texttt{ci}_L^{\alpha}(y)) = \Pr(\texttt{med}(P) < y_{(\texttt{N}_L^{\alpha})}) \leq F_{\text{Bin}}(\texttt{N}_L^{\alpha} - 1) \leq \alpha/2,$$

where F_{Bin} is the CDF of the binomial random variable $\text{Bin}(n, 1/2)$. The probability of failure at the upper end of the confidence interval is analogous.

Now, let us turn to the coverage analysis of a ρ-CDP algorithm $M : \mathcal{X}^n \times \mathcal{H} \to I_{\mathbb{R}}$. Let $M(y) = [M(y)_L, M(y)_U]$. A naïve way to analyze the coverage error of M is to attempt to find β_1 and β_2 such that assuming β_1 denotes the failure probability for the nonprivate confidence interval, the $M(y)$ contains the nonprivate interval $[\texttt{ci}_L^{\beta_1}, \texttt{ci}_U^{\beta_1}]$ with probability $1 - \beta_2$. Then M has coverage at least $1 - (\beta_1 + \beta_2)$. Even if β_1 and β_2 are chosen carefully, this analysis can be overly conservative. In particular, it assumes that the only way that $M(y)$ can succeed in containing $\texttt{med}(P)$ is if both $\texttt{med}(P) \in [\texttt{ci}_L^{\beta_1}, \texttt{ci}_U^{\beta_1}]$ and $[\texttt{ci}_L^{\beta_1}, \texttt{ci}_U^{\beta_1}] \subset M(y)$. Neither of these events are inherently necessary.

A more careful analysis of the relationship between the sampling error and the error due to privacy results in a tighter coverage analysis. As in the nonprivate setting, there are two ways that $M(y)$ can fail to contain $\texttt{med}(P)$. We will focus on analyzing the probability that $\texttt{med}(P) < M(y)_L$.

$$\begin{aligned}\Pr(\texttt{med}(P) < M(y)_L) &= \sum_{m=0}^{n} \Pr(\texttt{rank}_y(\texttt{med}(P)) = m) \cdot \Pr(\texttt{med}(P) < M(y)_L \,|\, \texttt{rank}_y(\texttt{med}(P)) = m) \\ &= \sum_{m=0}^{n} \Pr(\text{Bin}(n, 1/2) = m) \cdot \Pr(\texttt{med}(P) < M(y)_L \,|\, \texttt{rank}_y(\texttt{med}(P)) = m) \end{aligned} \tag{1}$$

Now, we have reduced the problem to analyzing the failure probability conditioned on the empirical rank of $\texttt{med}(P)$ in the dataset y. This is a helpful reduction since, as we will see in the following section, most of our algorithms will come with accuracy guarantees on the rank of $M(y)_L$. Accuracy guarantees of this form can then be exploited, via equation (1), to obtain a coverage analysis of M. As we will illustrate in our experiments below, there is a stark difference between the performance of algorithms designed using the naive analysis, and those using the more careful analysis.

3.2 Private Confidence Intervals: `ExpMech` and `CDFPostProcess`

We will evaluate the performance of two algorithms for releasing CDP confidence intervals for the median in this article; `ExpMech` and `CDFPostProcess`.

The first algorithm, which we call `ExpMech`, is efficient and satisfies the stronger privacy guarantee of pure DP. It outputs the tightest, or close to the tightest confidence intervals in a majority of parameter regimes we studied. It is based on the exponential mechanism (McSherry and Talwar 2007), a standard tool from DP. The exponential mechanism has been used in prior work to give DP point estimates for the median (Dwork and Lei 2009; Thakurta and Smith 2013; Johnson and Shmatikov 2013; Alabi et al. 2020; Asi and Duchi 2020). Our extension to providing confidence intervals for the median, while using similar ideas to prior work, requires a careful coverage analysis that is new to this work.

The `CDFPostProcess` algorithm partially addresses a common frustration with differentially private data analysis; that exploratory data analysis to visualize the dataset and verify findings typically requires additional privacy budget. For many tasks, this means allocating privacy budget away from the primary task resulting in a noisier algorithm. A key feature of `CDFPostProcess` is that it releases a full CDP estimate to the empirical CDF without consuming additional privacy budget. It is then notable, and perhaps surprising, that in many settings this algorithm performs almost as well as `ExpMech`, which releases no side information. We will focus on a particular CDF estimator based on the tree-based mechanism introduced in Li, Hay, Rastogi, Miklau, and McGregor (2010); Dwork, Naor, Pitassi, and Rothblum (2010), and Chan, Shi, and D. Song (2011) and further refined in Honaker (2015).

A full description of both algorithms can be found in the supplementary material online accompanying this article. We also experimented with several other algorithms that are not discussed in this section since they are outperformed by `ExpMech` and `CDFPostProcess` in the majority of parameter regimes. Brief descriptions of these additional algorithms can be found in the supplementary material online. All our algorithms require hyperparameters (for a discussion, see supplementary material online H.

4. SIMULATION STUDIES

In this section we present extensive simulation studies to evaluate the different algorithms under various parameter settings.[4] We focus on log-normal data, as an example of a skewed distribution for which the median would generally be

4. Code for producing these simulations can be found at https://github.com/anonymous-conf-medians/dp-medians.

preferred over the mean. We expect many of our findings to extend to other types of skewed data. We evaluate the performance of the two algorithms described in section 3.2, as well as the nonprivate confidence interval described in lemma 2.1, in terms of width of confidence interval, coverage, and bias. Since we rely on a log-normal distribution for our simulations, a simple parametric approach can also be exploited: given that mean and median match for symmetric distributions, we can compute a confidence interval for the mean on the log-scale and exponentiate the bounds of this confidence interval to obtain the interval on the original scale. We add this parametric confidence interval in our evaluations. However, we emphasize that such a strategy will only give valid results if the parametric assumptions are met. In section 5.2, we illustrate the negative consequences if this is not the case.

4.1 Data Description

To visualize the distribution of the noisy confidence intervals, we run each private algorithm 5 times on 100 independently drawn datasets. Let $\mathbf{x}_1, \ldots, \mathbf{x}_{100}$, each contain $n = 1{,}000$ i.i.d. draws from the underlying log-normal distribution. The underlying normal random variable has mean $\mu = \ln(1.5)$ and standard deviation of either $\sigma = 1$ or $\sigma = 5$.

4.2 Utility Measures

We consider two main utility measures in our experiments. The first measure is the relative width of the CDP confidence interval $M(y)$ compared to the width of the nonprivate confidence interval, $[\mathtt{ci}_L^{\alpha}, \mathtt{ci}_U^{\alpha}]$. For a dataset $y \in \mathcal{D}^n$, $\alpha \in [0, 1]$, and interval $I = [I_L, I_U]$, the relative width is defined as $\text{rel-width}^{\alpha}(y, I) = \frac{I_U - I_L}{\mathtt{ci}_U^{\alpha}(y) - \mathtt{ci}_L^{\alpha}(y)}$. We are concerned with the distribution of $\text{rel-width}^{\alpha}(y, I)$ when $I = M(y)$. We expect the private confidence intervals to be wider than the nonprivate intervals, so if $I = M(y)$ then we expect $\text{rel-width}^{\alpha}(y) \geq 1$ with high probability. If $\text{rel-width}^{\alpha}(y) \leq 2$, then the additional uncertainty due to privacy is less than the uncertainty due to sampling. We are interested in the distribution of the relative width over multiple trials, so for each algorithm we show boxplots of this metric over 500 trials (100 datasets times 5 trials of the DP mechanism on each).

The second utility measure is an empirical estimate of the actual coverage of the DP confidence interval. For intervals $I_1, \ldots, I_T$ and distribution P, let $\text{cov}_T(P, I_1, \ldots, I_T) = \frac{1}{T}\sum_{t=1}^{T} 1_{\mathtt{med}(P) \in I_t}$. Given $n \in \mathbb{N}$ and a confidence interval M, for all $t \in [T]$, let $y_t \sim P^n$ and $I_t = M(y_t)$ then $\text{cov}_{T,n,P}(M) = \text{cov}_T(P, I_1, \ldots, I_T)$ gives an estimate of the actual coverage of M on the distribution P. We estimate the coverage over 5,000 trials (1,000 drawn samples of size $n = 1{,}000$ times 5 trials of the DP mechanisms on each). The actual coverage may, and in many settings will,

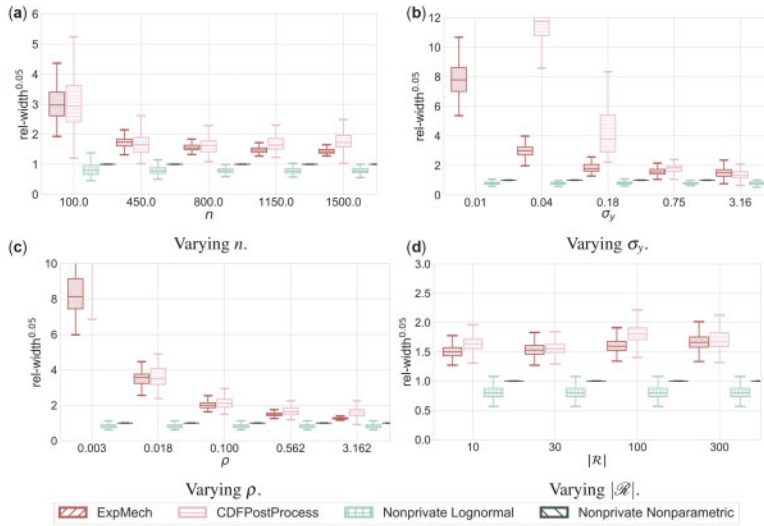

Figure 1. Relative Widths of DP Confidence Intervals as We Vary (a) Dataset Size n, (b) Dataset Standard Deviation σ_y, (c) Privacy Parameter ρ, and (d) Size of Range $|\mathcal{R}|$. The data are generated from a log-normal distribution with mean $\mu=\ln(1.5)$. Parameters are varied one at a time. All other parameters are fixed at their default values: standard deviation $\sigma_y=1.0$, number of datapoints $n = 1{,}000$, privacy parameter $\rho=0.5$, granularity $\theta=0.05$ (for all except `ExpMech`, where we use $\theta=0.01$), and nominal coverage rate $\alpha=0.05$. By definition, rel-width$^{\alpha}(\boldsymbol{y}, \boldsymbol{I})=1$ when $\boldsymbol{I}=[\texttt{ci}_L^{\alpha}, \texttt{ci}_U^{\alpha}]$.

exceed the nominal coverage (which is achieved by all the confidence intervals discussed).

4.3 Results and Discussion

4.3.1 Comparison among algorithms. Figure 1 demonstrates the performance of our two CDP confidence interval algorithms across a range of parameter regimes on log-normal data, in terms of the relative width metric. Notice that in a variety of regimes, including large n, large ρ and large σ_y both CDP algorithms provide confidence intervals that are at most twice the width of the nonprivate confidence interval with high probability. Our results indicate that `ExpMech` provides the tightest, or close to the tightest, confidence intervals in most parameter regimes we studied. This algorithm is the most targeted of the CDP algorithms we discuss and is carefully calibrated to not waste privacy budget on estimating additional information about P. It is a good general choice when one is solely interested in confidence intervals for the median. There are a few regimes in which `CDFPostProcess` outperforms `ExpMech` which we will discuss in this section.

The `CDFPostProcess` algorithm is appealing in practice since it allows a CDP estimate of the CDF to be released without consuming additional privacy budget. Surprisingly, in a variety of parameter regimes, `CDFPostProcess` provides confidence intervals that are almost as tight as those obtained by `ExpMech`. In fact, when σ_y is large, `CDFPostProcess` can result in tighter confidence intervals than `ExpMech` (figure 1b). We explore this further in supplementary material online K. Conversely, when σ_y is small, or $|\mathcal{R}|$ is large, `CDFPostProcess` is not a good choice. These are regimes where `CDFPostProcess` spends a lot of its privacy budget estimating the CDF in regions that are far from the median. These regimes are better served by the algorithms `BinSearch` + `CDF` and `NoisyBinSearch` discussed in the supplementary material online. Not surprisingly, the nonprivate parametric confidence interval is shorter than its nonparametric counterpart in most of the simulation runs.

4.3.2 Actual coverage analysis. A key component of the algorithmic design of each of the CDP confidence intervals was the coverage analysis. We discussed in section 3.1 how a careful coverage analysis that leverages the relationship between the two sources of the randomness potentially results in a much tighter coverage analysis than the naïve analysis that separates the sources of randomness. Our experimental results presented in figure 2 highlight two key findings regarding the coverage; that the careful analysis does result in substantially tighter intervals, and that the actual coverage of the CDP confidence intervals is still notably above the nominal coverage.

Figure 2a compares the relative width of the different confidence intervals. `ExpMechUnion` and `CDFPostProcessUnion` refer to the versions of `ExpMech` and `CDFPostProcess` resulting from the naïve coverage analysis. While the relative width is only slightly reduced for `ExpMech`, the relative width of `CDFPostProcess` is almost halved if the improved approach is used to produce the confidence intervals. These findings are also reflected in figure 2b, which compares the actual coverage of the naïve coverage analyses of `ExpMechUnion` and `CDFPostProcessUnion` and the more careful analyses described in section 3.1. While theoretically we can show that the careful analysis will result in actual coverage that is much closer to the target coverage, figure 2a shows that this improvement is practically relevant. While the improved analysis only leads to modest reduction in the overcoverage for `ExpMech`, the changes for `CDFPostProcess` are more substantial. The naïve approach results in coverage rates that are close to 1 irrespective of the target failure rate α. The improved approach leads to coverage rates that are much closer to the nominal coverage rates.

Despite the substantial improvement, figure 3 shows that all the CDP algorithms exhibit actual coverage higher than the nominal coverage for moderate values for ρ. As expected, figure 3a reveals that the coverage rates get closer to

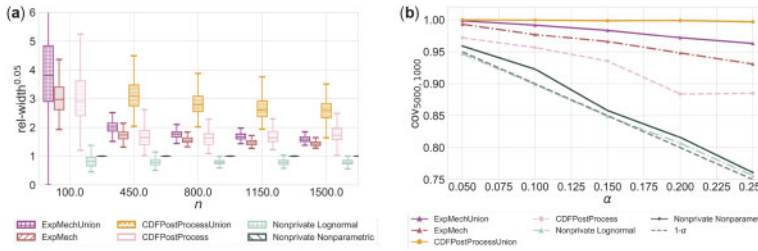

Figure 2. Comparing the Performance (in Terms of Relative Widths and Actual Coverage Rates) of Naive versus More Carefully Constructed DP Confidence Intervals on Log-Normal Data with Parameters $\mu = \ln(1.5), \sigma_y = 5.0$. (a) Relative widths of confidence intervals as we vary number of datapoints, n, from 100 to 1,500. Nominal coverage rate $\alpha = 0.05$ and privacy parameter $\rho = 0.5$. (b) Actual coverage rate of confidence intervals as we vary the nominal failure rate, α, from 0.05 to 0.25. Dataset contains $n = 1{,}000$ datapoints. Privacy parameter $\rho = 1.0$.

the nominal coverage rate for increasing values of ρ as each algorithm trends toward outputting the nonprivate confidence interval $[\texttt{ci}_L^\alpha, \texttt{ci}_U^\alpha]$ when $\rho = \infty$ (we use an unconventionally large value of $\alpha = 0.2$ to better visualize the difference in convergence rates for the different algorithms). Figure 3b shows the actual coverage rates as n increases. Note that the nonzero granularity in both private algorithms prevents the actual coverage from approaching the actual coverage of the nonprivate algorithms. The granularity is held constant in these simulations, but could be decreased with n to limit its effect. Over-coverage does not necessary correspond to substantially larger confidence intervals. We see in figure 1 that in a wide range of parameter regimes our CDP algorithms still result in confidence intervals that are at most twice as wide as their nonprivate counterparts. However, it does suggest an opportunity for improvement. An important question for future work is to what degree this over-coverage is necessary? In particular, is there an inherent tension between the privacy guarantee, and learning enough about the dataset to accurately quantify the uncertainty?

In many estimation tasks defining a nonparametric confidence interval that gives close to nominal coverage rates is difficult. Without information regarding the underlying distribution, the confidence intervals need to be wide enough to ensure valid coverage rates for any possible distribution. This can result in the nonparametric confidence intervals having higher than expected actual coverage when the data is drawn from a nice distribution. This effect is one possible explanation for the fact that the CDP confidence intervals have actual coverage higher than $1 - \alpha$ in figure 2a. In fact, we see evidence of this in the analysis of `ExpMech`. The error of the exponential mechanism is data dependent (and hence distribution dependent), but in our coverage analysis we

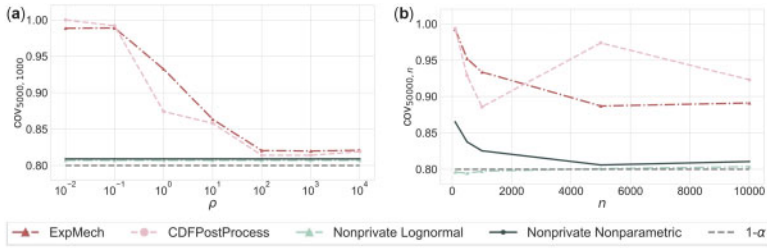

Figure 3. Actual Coverage Rate of DP Confidence Intervals as We Vary (a) Privacy Parameter ρ and (b) Number of Datapoints n on Log-Normal Data. (a) Actual coverage rate as we vary privacy parameter, ρ, from 0.01 to 10,000. Dataset contains $n = 1{,}000$ datapoints sampled i.i.d. from a log-normal distribution with parameters $\boldsymbol{\mu} = \ln(1.5), \boldsymbol{\sigma_y} = 5.0$. Nominal coverage rate $\boldsymbol{\alpha} = 0.2$. (b) Actual coverage rate as we vary number of datapoints, n, from 100 to 10,000. Dataset contains n datapoints sampled i.i.d. from a log-normal distribution with $\boldsymbol{\mu} = \ln(1.5), \boldsymbol{\sigma_y} = 5.0$. Nominal coverage rate $\boldsymbol{\alpha} = 0.2$. Privacy parameter $\boldsymbol{\rho} = 1.0$.

are forced to use the worst case error of the exponential mechanism over all datasets.

4.3.3 Bias. The goal of this work was to design algorithms that output valid confidence intervals for the median, not to estimate the median itself. An ad hoc estimate of the median can be obtained from a confidence interval by taking the estimate to be the mid-point of the interval. This approach is preferable in the DP context since it allocates its entire budget to the object of interest (the confidence interval) and we discussed in section 3 some of the reasons why direct estimators of the median are difficult to generalize to CDP confidence intervals. For all of our CDP algorithms, as well as the nonprivate confidence interval, this results in a biased estimator for the median, if the underlying distribution is skewed. In figure 4, we explore the bias of the inherited median estimators. As expected, the bias increases with the skew of the data. The bias of most of the DP algorithms is not substantially different from the bias for the nonprivate estimate. This implies that most of the bias can be attributed to the ad-hoc strategy of using the mid-point of the confidence interval as the point estimate for the median. As mentioned earlier, one benefit of `CDFPostProcess` is that it comes with additional information about the distribution which could potentially be used to release a less biased estimate of the median. We leave this for future work.

5. REAL DATA APPLICATION

In this section, we illustrate how the findings from the previous sections could inform the implementation of a differentially private median release strategy in

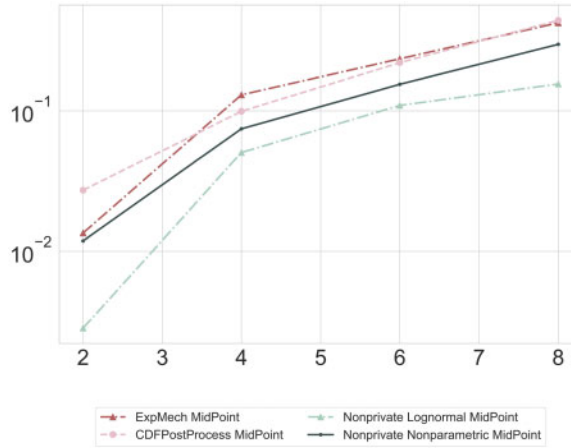

Figure 4. Bias of the Algorithms (Difference between the Center of the Confidence Intervals Obtained and the True Median, Averaged over 5 Trials on 100 Datasets) for Log-Normal Data with Parameters $\mu = 1.0$ and σ_y Varied from 2 to 8.

practice. We also demonstrate what level of accuracy one could reasonably expect for realistic applications. Our motivating example is the median income tables published by the U.S. Census Bureau for various subgroups of the population. Specifically, we aim to replicate a subset of statistics from table A1, *Income Summary Measures by Selected Characteristics: 2018 and 2019*, which appears in Semega, Kollar, Shrider, and Creamer (2020). This table reports median household income broken down by Type of household, Race and Hispanic Origin of Householder, Age of Householder, Nativity of Householder, Region, and Residence. For each of the 32 subgroups specified, the table provides the estimated median income and estimated margin of error (based on $\alpha = 0.1$) for 2018 and 2019. The estimates are computed using the Current Population Survey, 2019 and 2020 Annual Social and Economic Supplements (CPS ASEC).

Since we want to assess the accuracy of the CDP estimates, we use income data from the 1940 Decennial Census (Ruggles, Flood, Foster, Goeken, Pacas, et al. 2021), which enables us to compare the noisy estimates to the true values in the population. We restrict the population data to heads of households in the mountain division region (Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming) and focus on the variables type of household (two categories), metropolitan area (two categories) and age (two categories). To mimic the illustrative application described above, we repeatedly sample from this population and treat the

resulting data as the survey from which the (noisy) estimated medians will be computed. For simplicity, we draw 1 percent simple random samples without replacement. We acknowledge that the sampling design for the CPS ASEC is far more complex. However, understanding the subtle effects of complex sampling designs on the privacy guarantees is currently an area of active research and is beyond the scope of this article.

The variable we use for our evaluations is "INCWAGE," which "reports each respondent's total pre-tax wage and salary income for the previous year." The amounts are displayed in "contemporary dollars," which means they are not adjusted for inflation. In the 1940's dataset, the variable is topcoded at 5,001 dollars. We remove all N/A and missing values from the dataset, and only consider records corresponding to the head of each household. We note that we do not propose simply dropping all cases with missing values in practice as this will likely introduce bias. However, properly integrating any nonresponse adjustments into the DP algorithms is beyond the scope of the article. Thus, we treat the fully observed data on household heads as our population of interest. Finally, for the purpose of error analysis we treat the empirical distribution of the entire population dataset (from which we sample 1 percent) as the true underlying distribution P.[5]

5.1 Selecting the Algorithm and Hyperparameters

To generate the privatized confidence intervals, we use the algorithm identified as the winner in a wide range of regimes in the simulation studies: `ExpMech`. The hyperparameters are set to $\mathcal{R} = [0, 5001]$, and $\theta = 5$. The lower and upper bounds are chosen based on the assumption that the threshold used for top coding is public knowledge and that the median income will not be less than zero. The granularity parameter θ is chosen based on Census Bureau data visualizations that report median incomes from 1967 to present, which are rounded to the nearest \$100 (U.S. Census Bureau 2020b) indicating that a granularity of \$5 for 1940 median incomes is likely sufficient for data users. We split the overall privacy budget of $\rho = 0.5$ equally across three characteristics: type of household, metropolitan status, and age.

5.2 Parametric versus Nonparametric Approaches

As indicated above, it is beneficial under privacy constraints to use nonparametric approaches over parametric approaches. This is because the usual

5. Note that many respondents report incomes rounded to the nearest \$5, \$10, or \$50, which results in substantial overcoverage even for the nonprivate confidence intervals due to the spikes in the data. Therefore, we add a negligible amount of noise ($\mathcal{N}(0, 0.01)$) to each population income data point, so that the distribution being sampled is continuous. See supplementary material online C for further discussion.

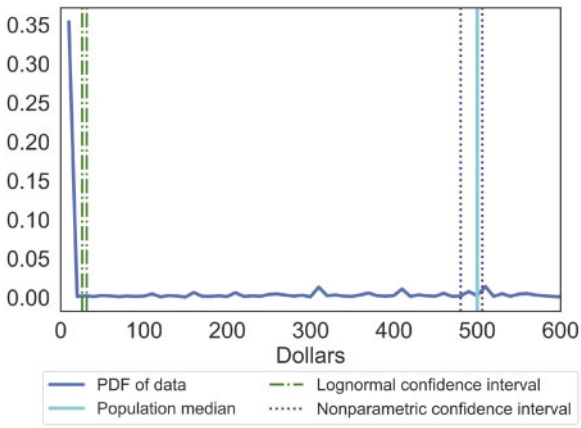

Figure 5. Comparing Nonprivate Parametric versus Nonparametric Confidence Intervals under Model Mismatch. The data contain incomes of family households ($n = 9{,}142$) based on a a 1 percent sample of mountain division households in the 1940 Decennial Census. The PDF of data shows that about 35 percent of the reported incomes are zero, which means that the data do not follow a log-normal distribution. Accordingly, the log-normal confidence interval is biased toward zero, while the nonparametric confidence interval captures the true population median.

guardrails around using parametric assumptions—such as visualizing the data and running goodness-of-fit tests—require privacy budget. Therefore, an analyst using private algorithms will find it more difficult to notice failures of parametric assumptions. Additionally, these failures will likely result in misleading uncertainty measures.

Figure 5 illustrates the dangers of using a parametric approach. The figure shows the empirical CDF of incomes from a 1 percent sample of family households in the Mountain region from the 1940 Decennial Census. Income data are typically assumed to be log-normally distributed; however, the data shown in figure 5 do not satisfy this assumption, as demonstrated by the large percentage of datapoints at 0. Due to this mismatch, a parametric confidence interval[6] computed under the assumption of log-normality (shown in green) is biased away from the true median (in cyan). Under privacy constraints, one would need to use privacy budget in order to visualize the distribution of the data and to notice this failure of the assumption. Therefore, it is preferable to instead compute a nonparametric confidence interval (as shown in purple) which maintains confidence validity.

6. The parametric confidence interval is generated by log-transforming the data, computing a normal 95% confidence interval for the mean of the log-transformed data, and back-transforming the limits of the confidence interval. More sophisticated parametric approaches exist in the literature; however, we focus on this one because it has the most natural private analog.

5.3 Results

Results based on the first simulation run are included in table 1. The CDP confidence intervals and median estimate are the result of a single run of `ExpMech` so this table is indicative of what we would expect in practice. The CDP point estimates for the median incomes are chosen as the midpoint of the corresponding CDP confidence intervals. Note that we could also leverage a prior assumption of the right-skewness of income data by choosing the CDP point estimator from the left half of the CDP confidence interval, rather than from its center, but we leave this type of parametric estimation to future work. We leverage the assumption that the incomes are nonnegative, so we set the lower endpoint of the CDP confidence intervals at the maximum of the output of the algorithm and zero. However, we compute the point estimates before the truncation step to avoid introducing bias. The table also provides nonprivate and private 90 percent confidence intervals (the margin of error reported in the Census tables could be computed as the half-width of these intervals).

The nonprivate median estimates are closer to the true values than the DP estimates for all sub-populations. However, for many statistics, the difference between the point estimates is small relative to the width of the confidence interval indicating that the bias introduced by the ad-hoc approach of using the center of the confidence interval to estimate the median is minor. Except for the householders aged 65 and above, the relative increase in uncertainty also seems to be acceptable. The relative increase in confidence interval length ranges between 20.7 and 81.7 percent, that is, the uncertainty from data protection is always less than the uncertainty from sampling. The large relative increase of the confidence intervals for householders aged 65 and above can be explained by noting that the width of the CDP intervals is lower bounded by the granularity parameter (which we chose to be $\theta = 5$), which leads to a large relative uncertainty if the nonprivate interval has width close to zero. However, the absolute increase in uncertainty is still acceptable.

In figures 6 and 7, we explore the performance of `ExpMech` and the nonprivate algorithm over 1,000 randomly sampled datasets. Figure 6, which contains boxplots showing the width of the private and nonprivate confidence intervals, confirms the findings based on one simulation run. While the private confidence intervals are typically wider than the nonprivate intervals, the increase in width is less than a multiplicative factor of two for all sub-populations except for head of households older than 65 and is less than $100 in all sub-populations. The figure also reports the actual coverage rates for the nonprivate and private confidence intervals. The actual coverage rates are computed over 1,000 simulation runs and 20 trails of the CDP algorithm within each simulation run. While the nonprivate actual coverage rates are close to the nominal 90 percent coverage, the CDP confidence intervals overcover substantially with actual coverage rates between 95.9 and 100 percent.

Table 1. Income Summary Measures by Selected Characteristics based on 1 Percent Simple Random Sample of Mountain Division Householder Records (i.e., in Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming) from the 1940 Decennial Census

Characteristics household(er)	Nb.Obs. in sample	CDP median Income	DP 90% CI	Nonprivate Median Income	Nonprivate 90% CI	Population Median Income
Type of household						
Family households	9,142	489.00	(469.99, 508.01)	499.97	(480.01, 500.93)	499.95
Nonfamily households	1,479	65.50	(0.0*, 136.01)	20.12	(0.17, 99.89)	0.20
Metropolitan status						
Not in metropolitan area	8,243	324.99	(290.03, 359.95)	329.85	(300.06, 359.85)	360.07
In metropolitan area	2,380	708.12	(640.00, 776.23)	699.99	(659.94, 749.85)	699.91
Age						
Age $<$65 years	9,259	564.89	(529.94, 599.84)	560.05	(540.03, 597.95)	540.10
Age $\geq$ 65 years	1,366	0.00	(0.0*, 5.0)	0.03	(0.02, 0.03)	0.03

NOTE.— Income is shown in 1940s dollars and is top-coded at \$5001. Differentially private estimates are obtained using `ExpMech` on a single sample with total privacy budget $\rho = 0.5$, range $\mathcal{R} = [0, 5001]$ and granularity $\theta = 5$. Note that the lower DP confidence interval values with a * are truncated to zero (0.0) based on the assumption that incomes are nonnegative. The DP point estimates are computed before the truncation to avoid introducing bias. All values are rounded to the nearest cent.

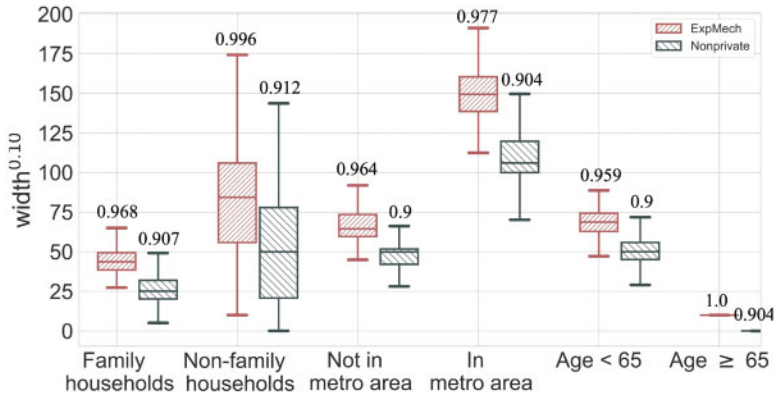

Figure 6. Comparing Widths of 90 Percent `ExpMech` and Nonprivate Confidence Intervals. Algorithms are run on 1,000 samples of income data by selected characteristics from the 1940 Decennial Census (the DP algorithm is run 20 times for each sample). Actual coverage rates are displayed for each algorithm.

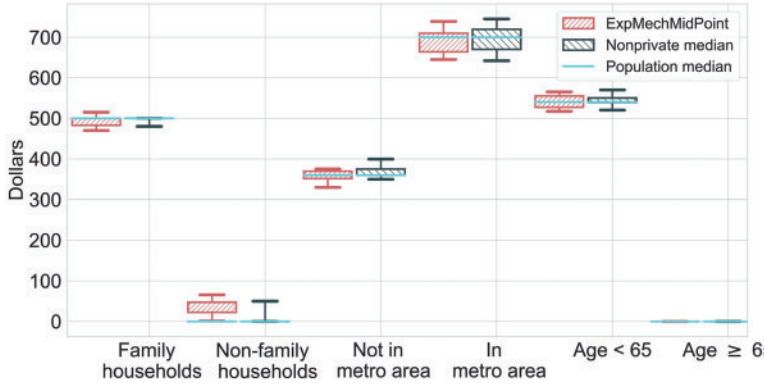

Figure 7. Comparing Distribution of Point Estimates. `ExpMechMidPoint` is the private median estimator obtained by taking the mid-point of the confidence interval produced by `ExpMech`. Algorithms are run on 1,000 samples of income data by selected characteristics from the 1940 Decennial Census. Population medians by characteristic are denoted by the cyan lines. For the Age $\geq$ 65 category, both the private and nonprivate estimators consistently return zero (0).

Figure 7 evaluates the bias introduced by taking the mid point of the confidence interval as a point estimate for the median. It contains boxplots showing the variability of the private and nonprivate point estimates (where `ExpMechMidPoint` is the private median estimator obtained by taking the midpoint of the confidence interval produced by `ExpMech`). The whiskers of the boxplots indicate the 5th and 95th quantile of the empirical distribution of

the point estimates computed over 1,000 trials. For most of the estimates, the range of the boxplots overlap to a large extent and similar to table 1 the bias is small relative to the variability in the estimates. The only estimate for which we find noticeable bias in the private midpoint median estimate is for nonfamily households. The bias arises because of the large fraction of zeros among this sub-population in the original data. Since 51 percent of the records in the original data report an income that is essentially zero (except for the small amount of noise that we introduce to make our data approximately continuous), the sample median will also be close to zero in many simulation runs. However, any point estimator based on the midpoint of a confidence interval (private or nonprivate) will almost always be strictly positive since the upper limit of this confidence interval will almost always be larger than the 51st quantile in the population, that is, almost always be larger than zero. So, while in some specific settings the bias of the midpoint estimator may be notable, typically the bias is small compared to the variance of both of the private and nonprivate median estimates.

6. CONCLUSION

In this article, we designed and evaluated several strategies to obtain differentially private confidence intervals for the median. We demonstrated that accounting for both sources of randomness, the sampling error as well as the error from the DP algorithm, simultaneously allowed us to give tighter confidence bounds than relying on naive approaches that account for the two components sequentially. Our simulation results showed that an algorithm `ExpMech` produced reliable and consistent confidence intervals which were less than twice the width of the nonprivate confidence intervals in a wide variety of parameter regimes. An algorithm `CDFPostProcess` provides confidence intervals that are almost as tight, or slightly tighter, than `ExpMech` in a variety of regimes. This algorithm is practically appealing since it releases a wealth of additional information about the distribution P without consuming additional privacy budget.

The private confidence intervals in the application based on the 1940 Decennial Census were not substantially wider than the intervals in the nonprivate setting, illustrating that the extra uncertainty due to data protection can be small in practice. We also found that the bias introduced by the ad-hoc strategy of using the midpoint of the confidence interval to estimate the median was limited for most estimates in our real data application. We also note that the `CDFPostProcess` algorithm allows us to release a direct estimate of the median without consuming additional privacy budget.

We saw in our experiments on both simulated and real data that the actual coverage rate of our private confidence intervals was often (sometimes substantially) higher than the nominal coverage rate. An interesting open question is whether this

is inherent for nonparametric CDP confidence intervals for the median. Furthermore, if this is unavoidable, then what distributional assumptions are required to narrow the gap between the actual coverage and nominal coverage rates?

Finally, perhaps the strongest limitation of our study is the reliance on the assumption that the sample is drawn using simple random sampling with replacement. Such a sampling design will never be used in the survey context in practice. Thus, the important next step will be to extend the methodology to allow for more complex designs.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam. The Supplementary material contains further details, psuedocode, and privacy analysis for the `ExpMech` and `CDF post-process` algorithms. We also provide comparison to several other potential algorithms for producing confidence intervals for the median that are outperformed by `ExpMech` and `CDF post-process` in most regimes.

REFERENCES

Abowd, J. M. (2018), "Staring-Down the Database Reconstruction Theorem," in American Statistical Association *Proceedings of Committee on Professional Ethics*. Available at https://ww2.amstat.org/meetings/jsm/2018/onlineprogram/AbstractDetails.cfm?abstractid=326575.

Alabi, D., A. McMillan, J. Sarathy, A. Smith, and S. Vadhan (2020), "Differentially Private Simple Linear Regression," arXiv.2007.05157.

Asi, H. and J. C. Duchi (2020), "Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms." In *Advances in Neural Information Processing Systems (Volume 33)*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, pp. 14106–14117, Vancouver, BC, Canada: Curran Associates, Inc.

Barrientos, A. F., J. Reiter, A. Machanavajjhala, and Y. Chen (2019), "Differentially Private Significance Tests for Regression Coefficients," *Journal of Computational and Graphical Statistics*, 28, 440–453.

Bernstein, G. and D. R. Sheldon (2018), "Differentially Private Bayesian Inference for Exponential Families," in *Advances in Neural Information Processing Systems (Volume 31)*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Vancouver, BC, Canada: Curran Associates, Inc. Available at https://papers.nips.cc/paper/2018/file/08040837089cdf46631a10aca5258e16-Paper.pdf.

———. (2019), "Differentially Private Bayesian Linear Regression," In *Advances in Neural Information Processing Systems (Volume 32)*, eds. Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Edward A. Fox, Roman Garnett, pp. 523–533, Vancouver, BC, Canada: Curran Associates, Inc.

Biswas, S., Y. Dong, G. Kamath, and J. Ullman (2020), "CoinPress: Practical Private Mean and Covariance Estimation," arXiv.2006.06618.

Brawner, T. W. and J. Honaker (2018), "Bootstrap Inference and Differential Privacy: Standard Errors for Free." Summer Meetings of the Society for Political Methodology, Provo, UT.

Bun, M. and T. Steinke (2016), "Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds," in *Theory of Cryptography Conference*, eds. M. Hirt and A. Smith, pp. 635–658, Berlin, Heidelberg: Springer.

———. (2019), "Average-Case Averages: Private Algorithms for Smooth Sensitivity and Mean Estimation," In *Advances in Neural Information Processing Systems Volume 32*, eds. Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Edward A. Fox, Roman Garnett, pp. 181–191, Vancouver, BC, Canada: Curran Associates, Inc.

Chan, T.-H. H., E. Shi, and D. Song (2011), "Private and Continual Release of Statistics," *ACM Transactions on Information and System Security*, 14, 1–24.

Couch, S., Z. Kazan, K. Shi, A. Bray, and A. Groce (2019), "Differentially Private Nonparametric Hypothesis Testing," Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS'19, pp. 737–751.

Degue, K. H. and J. L. Ny (2018), "On Differentially Private Gaussian Hypothesis Testing," in *Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing*, pp. 842–847.

Dimitrakakis, C., B. Nelson, A. Mitrokotsa, and B. I. P. Rubinstein (2014), "Robust and Private Bayesian Inference," in *Algorithmic Learning Theory*, eds. P. Auer, A. Clark, T. Zeugmann, and S. Zilles, pp. 291–305, Cham: Springer International Publishing.

D'Orazio, V., J. Honaker, and G. King (2015), "Differential Privacy for Social Science Inference," in *Alfred P. Sloan Foundation Economic Research Paper Series*, Sloan Foundation Economics Research Paper No. 2676160, Available at SSRN: http://dx.doi.org/10.2139/ssrn.2676160.

Drechsler, J. (2021), "Differential Privacy for Government Agencies—Are We There Yet?," *arXiv:2102.08847*.

Du, W., C. Foot, M. Moniot, A. Bray, and A. Groce (2020), "Differentially Private Confidence Intervals," *arXiv:2001.02285*.

Dwork, C. and J. Lei (2009), "Differential Privacy and Robust Statistics," *STOC*, 9, 371–380.

Dwork, C. and G. N. Rothblum (2016), "Concentrated Differential Privacy," *arXiv*:1603.01887.

Dwork, C., K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor (2006a), "Our Data, Ourselves: Privacy via Distributed Noise Generation," in *Advances in Cryptology—EUROCRYPT*, eds. S. Vaudenay, pp. 486–503, Berlin, Heidelberg: Springer.

Dwork, C., F. McSherry, K. Nissim, and A. D. Smith (2006b), "Calibrating Noise to Sensitivity in Private Data Analysis," Theory of Cryptography, in *Proceedings of the Third Theory of Cryptography Conference*, TCC 2006, New York, NY, USA, pp. 265–284.

Dwork, C., M. Naor, T. Pitassi, and G. N. Rothblum (2010), "Differential Privacy under Continual Observation," in *Proceedings of the Forty-Second ACM Symposium on Theory of Computing, STOC'10*, pp. 715–724. New York: Association for Computing Machinery.

Evans, G. and G. King (2022), "Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook Urls Dataset," *Political Analysis*, pp. 1–21.

Evans, G., G. King, M. Schwenzfeier, and A. Thakurta (2021), "Statistically Valid Inferences from Privacy Protected Data," Working Paper. Available at https://tinyurl.com/yd4xbnb8.

Ferrando, C., S.-F. Wang, and D. Sheldon (2020), "General-Purpose Differentially-Private Confidence Intervals," arXiv:2006.07749.

Foulds, J., J. Geumlek, M. Welling, and K. Chaudhuri (2016), "On the Theory and Practice of Privacy-Preserving Bayesian Data Analysis," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI'16*, pp. 192–201. Arlington, Virginia: AUAI Press.

Gaboardi, M., H. Lim, R. Rogers, and S. Vadhan (2016), "Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing," in *Proceedings of The 33rd International Conference on Machine Learning*, *Volume 48 of Proceedings of Machine Learning Research*, eds. M. F. Balcan and K. Q. Weinberger, pp. 2111–2120. New York: PMLR.

Gaboardi, M., R. Rogers, and O. Sheffet (2019), "Locally Private Mean Estimation: Z-Test and Tight Confidence Intervals," in *Proceedings of Machine Learning Research (Volume 89)*, eds. K. Chaudhuri and M. Sugiyama, pp. 2545–2554. New York: PMLR.

Garfinkel, S. L., J. M. Abowd, and C. Martindale (2019), "Understanding Database Reconstruction Attacks on Public Data," *Communications of the ACM*, 62, 46–53.

Gillenwater, J., M. Joseph, and A. Kulesza (2021), "Differentially Private Quantiles," arXiv:2102.08244.

Gong, R. (2019), "Exact Inference with Approximate Computation for Differentially Private Data via Perturbations," arXiv:1909.12237.

Heikkilä, M., E. Lagerspetz, S. Kaski, K. Shimizu, S. Tarkoma, and A. Honkela (2017), "Differentially Private Bayesian Learning on Distributed Data," in *Advances in Neural Information Processing Systems* (*Volume* 30), eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Vancouver, BC, Canada: Curran Associates, Inc.

Honaker, J. (2015), "Efficient Use of Differentially Private Binary Trees," Theory and Practice of Differential Privacy (TPDP 2015), London, UK.

Johnson, A. and V. Shmatikov (2013), "Privacy-Preserving Data Exploration in Genome-Wide Association Studies," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13*, pp. 1079–1087, New York: Association for Computing Machinery.

Karwa, V. and S. Vadhan (2018), "Finite Sample Differentially Private Confidence Intervals," 9th Innovations in Theoretical Computer Science Conference (ITCS 2018), volume 94 of Leibniz International Proceedings in Informatics (LIPIcs), pp. 44:1–44:9. Available at https://drops.dagstuhl.de/opus/volltexte/2018/8344/.

Lehmann, E. and H. D'Abrera (1975), Nonparametrics: Statistical Methods Based on Ranks. Holden-Day Series in Probability and Statistics. Holden-Day. ISBN 9780816249961. Available at https://books.google.com/books?id=BQ3YAAAAMAAJ.

Li, C., M. Hay, V. Rastogi, G. Miklau, and A. McGregor (2010), "Optimizing Linear Counting Queries under Differential Privacy," in *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 123–134.

Logan, J. R., C. Zhang, B. Stults, and T. Gardner (2021), "Improving Estimates of Neighborhood Change with Constant Tract Boundaries," *Applied Geography*, 132, 102476.

McSherry, F. and K. Talwar (2007), "Mechanism Design via Differential Privacy," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007)*, pp. 94–103.

Nissim, K., S. Raskhodnikova, and A. D. Smith (2007), "Smooth Sensitivity and Sampling in Private Data Analysis," *Proceedings of the 39th Annual ACM Symposium on Theory of Computing,* USA, pp. 75–84.

Rubin, D. B. (1996), "Multiple Imputation after 18+ Years," *Journal of the American Statistical Association*, 91, 473–489.

Ruggles, S., S. Flood, S. Foster, R. Goeken, J. Pacas, M. Schouweiler, and M. Sobek (2021), "IPUMS USA: Version 11.0 1940 decennial census." Available at https://doi.org/10.18128/D010.V11.0.

Semega, J., M. Kollar, E. Shrider, and J. F. Creamer (2020), "Current Population Reports, p60-270, Income and Poverty in the United States: 2019," Technical Report, U.S. Census Bureau, U.S. Government Publishing Office, Washington, DC.

Thakurta, A. G. and A. Smith (2013), "Differentially Private Feature Selection via Stability Arguments, and the Robustness of the Lasso," in *Proceedings of the 26th Annual Conference on Learning Theory, Volume 30 of Proceedings of Machine Learning Research*, eds. S. Shalev-Shwartz and I. Steinwart, pp. 819–850. Princeton, NJ: PMLR.

Tzamos, C., E.-V. Vlatakis-Gkaragkounis, and I. Zadik (2020), "Optimal Private Median Estimation under Minimal Distributional Assumptions," in *Advances in Neural Information Processing Systems* (*Volume* 33), eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, pp. 3301–3311, Vancouver, BC, Canada: Curran Associates, Inc.

U.S. Census Bureau (2020a), "Income Data Tables." Available at: https://www.census.gov/topics/income-poverty/income/data/tables.html. Accessed March 14, 2020.

U.S. Census Bureau (2020b), "Slides Available at Income, Poverty, and Health Insurance: 2019," Live Press Conference, September 15, 2020. Available at: https://www.census.gov/content/dam/Census/newsroom/press-kits/2020/iphi/20200915-iphi-slides-plot-points.pdf. Accessed March 16, 2020.

Vu, D. and A. Slavkovic (2009), "Differential Privacy for Clinical Trial Data: Preliminary Evaluations," in *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pp. 138–143.

Wang, Y. (2018), "Revisiting Differentially Private Linear Regression: Optimal and Adaptive Prediction & Estimation in Unbounded Domain," in *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018*, Monterey, California, USA, pp. 93–103.

Wang, Y., J. Lee, and D. Kifer (2015a), "Differentially Private Hypothesis Testing, Revisited," arXiv:1511.03376.

Wang, Y.-X., S.E. Fienberg, and A.J. Smola (2015b), "Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo," in *Proceedings of the International Conference on Machine Learning, PMLR*, pp. 2493–2502.