Electronic Journal of Statistics
Vol. 18 (2024) 1455–1494
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2234

Differentially private confidence intervals for proportions under stratified random sampling

Shurong Lin

Department of Mathematics and Statistics, Boston University, Boston, MA
e-mail: shrlin@bu.edu

Mark Bun and Marco Gaboardi

Department of Computer Science, Boston University, Boston, MA
e-mail: mbun@bu.edu; gaboardi@bu.edu

Eric D. Kolaczyk

Department of Mathematics and Statistics, McGill University, Canada
e-mail: eric.kolaczyk@mcgill.ca

Adam Smith

Department of Computer Science, Boston University, Boston, MA
e-mail: ads22@bu.edu

Abstract: Confidence intervals are a fundamental tool for quantifying the uncertainty of parameters of interest. With the increase of data privacy awareness, developing a private version of confidence intervals has gained growing attention from both statisticians and computer scientists. Differential privacy is a state-of-the-art framework for analyzing privacy loss when releasing statistics computed from sensitive data. Recent work has been done around differentially private confidence intervals, yet to the best of our knowledge, rigorous methodologies on differentially private confidence intervals in the context of survey sampling have not been studied. In this paper, we propose three differentially private algorithms for constructing confidence intervals for proportions under stratified random sampling. We articulate two variants of differential privacy that make sense for data from stratified sampling designs, analyzing each of our algorithms within one of these two variants. We establish analytical privacy guarantees and asymptotic properties of the estimators. In addition, we conduct simulation studies to evaluate the proposed private confidence intervals, and two applications to the 1940 Census data are provided.

MSC2020 subject classifications: Primary 68P27, 62G15; secondary 62Dxx.

Keywords and phrases: Differential privacy, confidence intervals, stratified sampling, population proportion.

Received January 2023.

Contents

1. Introduction

With the increase of privacy awareness in the modern information era, establishing privacy-preserving methodologies for statistics and machine learning has become an active research area. *Differential privacy*, a state-of-the-art privacy protection technique [14], is considered a gold standard for rigorous privacy guarantees. Not only has it drawn significant attention in academia [15, 16], but also it has been deployed by governments, firms, and other data agencies, such as the U.S. Census Bureau [1], Google [20], Microsoft [8], and Apple [38].

Recently, the U.S. Census Bureau released a new demonstration of its differentially private Disclosure Avoidance System (DAS) for the 2020 Census [5, 24]. At the intersection of differential privacy and statistics, both statisticians and computer scientists are working on developing private versions of statistical inference procedures. Early work discussing differential privacy in the context of statistics includes [17, 13, 42, 37]. More recent work has explored statistical inference and estimation under the constraint of differential privacy [6, 29, 31].

As one of the most fundamental tools for statistical inference, confidence intervals are ubiquitous in quantifying the uncertainty of parameters of interest. In this paper, we propose three differentially private algorithms for constructing confidence intervals for the population proportion under stratified random sampling. To the best of our knowledge, our work is the first to establish rigorous methodologies on differentially private confidence intervals in the context of survey sampling. Survey sampling is an important area in statistics that involves selecting a sample of individuals from a target population to conduct a survey. It provides timely and cost-efficient estimates of population characteristics of interest and is widely used in broad-scale data gatherings, such as the American Community Survey (ACS), the Survey of Income and Program Participation (SIPP), and the Current Population Survey (CPS).

This paper provides the first study of differentially private confidence intervals for data from stratified sampling designs. Specifically:

- We articulate two specific variants of differential privacy that are appropriate for data from stratified sampling designs. In addition to the standard notion of differential privacy, we also consider settings in which the sample stratum sample sizes are fixed and public. This latter setting allows for simpler algorithms and tighter confidence intervals.
- We give methods to propagate the uncertainty due to the application of differentially private mechanisms (adding random noise) into the construction of confidence intervals. A necessary bias correction is made to achieve (asymptotic) unbiased variance estimates. Central limit theorem (CLT)-type statements are provided to guarantee the confidence level asymptotically.
- We assess the performance of the proposed algorithms both in theory and through simulations. The theoretical analysis comparing the non-private and private methods gives practitioners a sense of how the algorithms would work prior to applying them to real data.
- To support the theoretical analysis of one of the algorithms, we study the behavior of a reciprocal normal variable in depth. A general form of the Taylor expansion (for conditional moments) is obtained to solve the problem of the non-existence of moments due to its heavy-tailed nature.

The paper is organized as follows. We briefly discuss the existing work on differentially private confidence intervals in Section 1.1. Section 2 provides preliminaries on confidence intervals of population proportions and differential privacy. In Section 3, we discuss the methodology of three differentially private algorithms. Section 4 provides theorems on both privacy and asymptotic cov-

erage guarantees. Numerical experiments, including simulation studies and two applications to the 1940 Census data, are conducted in Section 5. Section 6 discusses the implications of our methods and general research directions on differentially private confidence intervals.

1.1. Related work

Differentially private confidence intervals have recently been studied for other settings. Some studied differentially private confidence intervals for the population mean of normally distributed data [27, 11, 23]. Other tasks on confidence intervals have also been explored. Drechsler et al. designed and evaluated several strategies to obtain differentially private confidence intervals for the median [10]. Wang et al. provided confidence intervals for differentially private models trained with objective or output perturbation algorithms [41].

Besides, bootstrapping is a popular technique for constructing more general differentially private confidence intervals. Ferrando et al. proposed a general-purpose approach to construct confidence intervals for a population parameter [21]. A numerical confidence interval for the difference of mean was provided [9]. The nonparametric bootstrap was considered in [3]. Covington et al. described a method to induce distributions of mean and covariance estimates via the bag of little bootstraps (BLB), which can further produce private confidence intervals [7].

Our work is the first to study design-based approaches to sampling. In a design-based setting, the values of interest are viewed as fixed but unknown constants. Randomness only comes from the sampling design. The selection probabilities introduced with the design will be used for estimation. On the contrary, in a model-based setting, a parametric model is postulated. In many cases, especially with natural populations where no accurate prior information about the population distribution is available, design-based sampling methods can be more reassuring. More discussion of design-based versus model-based approaches in sampling can be found in [39].

2. Preliminaries

In this section, we provide some preliminaries on population proportion estimation and differential privacy. We first review the classic Wald confidence interval for the population proportion under stratified random sampling. Then we define a notion of differential privacy specifically for stratified data. Some properties of differential privacy are revisited in preparation for the theoretical analysis in Section 4.

2.1. Confidence intervals for the population proportion

In stratified random sampling, a population of N individuals is partitioned into H strata, where stratum h has N_h individuals, and simple random sampling

of n_h individuals is conducted within each stratum. When the objective is to estimate the proportion of individuals having some attribute in the population, one can estimate it by the sample proportion. Let y_{hi} be the corresponding indicator variable: $y_{hi} = 1$ when the individual i in stratum h has the attribute and $y_{hi} = 0$ otherwise. One can estimate the population proportion

$$p = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi}$$

by the sample proportion

$$\hat{p} = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = \sum_{h=1}^{H} w_h \hat{p}_h$$

where $w_h \overset{def}{=} \frac{N_h}{N}$ and $\hat{p}_h \overset{def}{=} \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$. Its variance $\mathrm{Var}(\hat{p}) = \sum_{h=1}^{H} w_h^2 \mathrm{Var}(\hat{p}_h)$, where

$$\mathrm{Var}(\hat{p}_h) = \left(\frac{N_h - n_h}{N_h - 1} \right) \frac{p_h(1 - p_h)}{n_h}.$$

An unbiased estimator for $\mathrm{Var}(\hat{p}_h)$ is given by the sample variance in the stratum

$$\widehat{\mathrm{Var}}(\hat{p}_h) = \left(\frac{N_h - n_h}{N_h} \right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}. \tag{1}$$

Then an unbiased estimator for $\mathrm{Var}(\hat{p})$ is given by $\widehat{\mathrm{Var}}(\hat{p}) = \sum_{h=1}^{H} w_h^2 \widehat{\mathrm{Var}}(\hat{p}_h)$. An approximate $100\%(1-\alpha)$ confidence interval for p based on a normal distribution can be constructed:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\mathrm{Var}}(\hat{p})}, \tag{2}$$

where $z_{1-\frac{\alpha}{2}}$ denotes the $1 - \frac{\alpha}{2}$ quantile of standard normal distribution. The normal approximation is useful when all the sample sizes are moderate to large. Otherwise, the t distribution with appropriate degrees of freedom is typically used to replace the standard normal distribution. For small sample sizes, various specialized confidence intervals have been developed [22].

2.2. Differential privacy

Differential privacy ensures that the output of data analysis or a query does not differ much when the data set is changed by one record, such that one can not infer the presence or absence of any individual. If two data sets $\boldsymbol{x}, \boldsymbol{x}'$ differ by one record, we say that $\boldsymbol{x}, \boldsymbol{x}'$ are *adjacent* or *neighboring*, written as $\boldsymbol{x} \sim \boldsymbol{x}'$. The definition of differential privacy depends on how we define adjacency. For the partitioned data under stratified sampling, there are two ways to change a

record: (1) one way is to substitute one record within a stratum, with all the stratum sample sizes fixed. We refer to this adjacency relation as "*substitute-one relation within a stratum*" and denote it by $\sim_{ss}$. This relation corresponds to the case where the sample sizes are public and fixed; (2) another way to obtain an adjacent data set is to remove or add one record from one stratum; we refer to the corresponding relation as, which we call "*remove/add-one relation*", denoted by $\sim_r$. In this case, one of the stratum sample sizes will change by one, as will the overall sample size. This relation corresponds to the case where the sample sizes are private.

Under either adjacency relation, we can define *zero-concentrated differentially private* (ρ-zCDP) as in [4]:

Definition 1 (ρ-zCDP)**.** *Let $\mathcal{X}^*$ denote the space of the input data with an arbitrary finite dimension. Under the adjacency relation $\sim$, a randomized algorithm $M : \mathcal{X}^* \to \mathcal{Y}$ is ρ-zero-concentrated-differentially private (ρ-zCDP) if, for every pair of adjacent data sets $\boldsymbol{x} \sim \boldsymbol{x}' \in \mathcal{X}^*$, and all $\alpha \in (1, \infty)$,*

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

where $\mathrm{D}_\alpha(M(\boldsymbol{x}) \| M(\boldsymbol{x}'))$ is the α-Rényi divergence [40] between the distribution of $M(\boldsymbol{x})$ and the distribution of $M(\boldsymbol{x}')$.

The parameter ρ indicates the *privacy level*. A smaller ρ means a more restrictive distance control between $M(\boldsymbol{x})$ and $M(\boldsymbol{x}')$. As a result, the outputs on two adjacent data sets are harder to tell apart and the algorithm achieves higher privacy. We call ρ the *privacy budget* when we deliberately design an algorithm to satisfy ρ-zCDP.

Depending on the adjacency notion, there are two types of differential privacy: *bounded* and *unbounded differential privacy* [28]. Definition 1 under the "remove/add-one relation" corresponds to the standard unbounded differential privacy. The sample size of the data set changes when one record is added or removed to obtain an adjacent data set. With "substitute-one within a stratum" relation $\sim_{ss}$, the resulting notion corresponds to the bounded version of differential privacy where the sizes of two adjacent data sets are the same. But it is somewhat different from the standard notion of bounded differential privacy in that for the latter, substitutions can happen across strata. That is, we can change both the record and the stratum it is part of.

In the literature on differential privacy, (ϵ, δ)-DP ([15] Definition 2.4) is considered the classic notion. We consider ρ-zCDP because (1) ρ-zCDP implies (ϵ, δ)-DP ([4] Proposition 1.3), (2) the application of the Gaussian mechanism to achieve zCDP facilitates the theoretical analyses, and (3) the composition of ρ-zCDP is straightforward. The Gaussian mechanism is a prototypical example of a mechanism satisfying zCDP, which perturbs the true values by adding Gaussian noise. We provide the Gaussian mechanism and the composition and post-processing properties of ρ-zCDP in the following propositions. All propositions can be found in [4] and will be used in the analyses of privacy guarantees in Section 4.

Definition 2 (Sensitivity)**.** *A function q: $\mathcal{X}^* \to \mathbb{R}$ has sensitivity Δ if for all pairs of adjacent data sets $x \sim x' \in \mathcal{X}^*$, we have $|q(x) - q(x')| \leq \Delta$.*

Proposition 1 (Gaussian Mechanism of ρ-zCDP)**.** *Let $q : \mathcal{X}^* \to \mathbb{R}$ be a sensitivity-Δ query. Consider the mechanism $M : \mathcal{X}^* \to \mathbb{R}$ that on input x, releases a sample from $N(q(x), \Delta^2/(2\rho))$. Then, M satisfies ρ-zCDP.*

A smaller budget leads to larger noise added to the query on average. Consequently, the output is more private.

Proposition 2 (Composition)**.** *Let $M : \mathcal{X}^* \to \mathcal{Y}$ and $M' : \mathcal{X}^* \to \mathcal{Z}$ be two randomized algorithms. Suppose M satisfies ρ-zCDP and M' satisfies ρ'-zCDP, then algorithm $M'' = (M, M') : \mathcal{X}^* \to \mathcal{Y} \times \mathcal{Z}$ is $(\rho + \rho')$-zCDP.*

Proposition 3 (Post-processing)**.** *Let $M : \mathcal{X}^* \to \mathcal{Y}$ and $f : \mathcal{Y} \to \mathcal{Z}$ be randomized algorithms. If M is ρ-zCDP, then so is the composed algorithm $M' = f \circ M : \mathcal{X}^* \to \mathcal{Z}$.*

3. Methodology

Our goal is to release a ρ-zCDP confidence interval for the population proportion p under stratified random sampling. To construct a confidence interval as in (2), we need to estimate both p and the variance of the estimator privately. Recall that the non-private estimator of population proportion is given by the sample proportion

$$\hat{p} = \sum_{h=1}^{H} w_h \hat{p}_h.$$

We assume the stratum sizes N_h are all public and fixed, thus so are w_h. To get a private estimator for p, denoted by $\tilde{p}$, we can add noise at the level of either the (non-private) estimator $\hat{p}$ or the estimator $\hat{p}_h$. With $\tilde{p}$, we further devise a private estimator for $\mathrm{Var}(\tilde{p})$. Based on this idea, two algorithms for the case of public sample sizes are designed by adding noise at the stratum or population level in section 3.1. In section 3.2, we additionally propose adding noise at the stratum level when sample sizes are private. Throughout the paper, the accents $\hat{\cdot}$ and $\tilde{\cdot}$ are used to represent non-private and private estimators, respectively.

3.1. Estimating with public sample sizes

When sample sizes n_h are fixed, there are two natural strategies for perturbing $\hat{p}$: add Gaussian noise to (1) the stratum-level statistics $\hat{p}_h$'s, or (2) the overall statistic $\hat{p}$. Adding noise to the $\hat{p}_h$'s has the advantage of producing private estimators for stratum-level proportions simultaneously.

3.1.1. Adding noise at the stratum level

We apply the Gaussian mechanism to each stratum to derive a private estimator $\tilde{p}_h \overset{def}{=} \hat{p}_h + e_h$ where e_h is the Gaussian noise. Then the private estimator for the population proportion is

$$\tilde{p} \overset{def}{=} \sum_{h=1}^{H} w_h \tilde{p}_h.$$

As a result, the variance of $\tilde{p}$ consists of both the intrinsic variances of estimating p_h's by $\hat{p}_h$'s and the additional variability from added noise:

$$\mathrm{Var}(\tilde{p}) = \sum_{h=1}^{H} w_h^2 \left(\mathrm{Var}(\hat{p}_h) + w_h^2 \,\mathrm{Var}(e_h)\right) \tag{3}$$

where $\mathrm{Var}(e_h), h = 1, ..., H$ are public since they do not depend on the data.

To obtain a private confidence interval for $\hat{p}$, we will need to privately estimate $\mathrm{Var}(\hat{p}_h)$. Note that the added noise biases the term $\hat{p}_h(1-\hat{p}_h)$ in the non-private estimate of $\mathrm{Var}(\hat{p}_h)$ in (1). More specifically, $\mathbb{E}_e[\tilde{p}_h(1-\tilde{p}_h)] = \hat{p}_h(1-\hat{p}_h) - \mathrm{Var}(e_h)$ where $\mathbb{E}_e$ denotes the expectation taken on the randomness of the added noise. Then a private unbiased estimator of $\mathrm{Var}(\hat{p}_h)$ in the right-hand side in (3) is given by

$$\widetilde{\mathrm{Var}}(\hat{p}_h) \overset{def}{=} \left(\frac{N_h - n_h}{N_h}\right) \frac{\tilde{p}_h(1-\tilde{p}_h) + \mathrm{Var}(e_h)}{n_h - 1}. \tag{4}$$

To estimate $\mathrm{Var}(\tilde{p})$, we set

$$\widetilde{\mathrm{Var}}(\tilde{p}) \overset{def}{=} \sum_{h=1}^{H} w_h^2 \left(\widetilde{\mathrm{Var}}(\hat{p}_h) + \mathrm{Var}(e_h)\right)$$

This approach is formulated in Algorithm 1 which we call **StrNz-PubSz** (adding noise at the stratum level with public sample sizes). The theoretical results regarding privacy level and asymptotic coverage are provided in Theorems 4.1 and 4.2.

3.1.2. Adding noise at the population level

An alternative strategy is to directly add noise to the non-private estimator of p, i.e., $\hat{\mathrm{p}}$. The sensitivity of $\hat{p}$ is

$$\Delta_p = \max_h \frac{w_h}{n_h}.$$

Since w_h and n_h are public, Δ_p can be made public. We set $\tilde{p} = \hat{p} + e$ where e is the Gaussian noise with standard deviation proportional to Δ_p. Then, the variance of $\tilde{p}$ becomes

$$\mathrm{Var}(\tilde{p}) = \mathrm{Var}(\hat{p}) + \mathrm{Var}(e). \tag{5}$$

Algorithm 1 Adding noise at the stratum level with public sample sizes, StrNz-PubSz

Input: $\hat{p}_h$, n_h, N_h, w_h, ρ, α.
Output: ρ-zCDP $(1-\alpha)$ CI for the population proportion.

1: **for** $h = 1$ to H **do**
2: Generate Gaussian noise $e_h \sim \mathcal{N}(0, \frac{1}{2\rho n_h^2})$, and let

$$\tilde{p}_h \leftarrow \hat{p}_h + e_h.$$

3: Estimate $\mathrm{Var}(\tilde{p}_h)$ by

$$\widetilde{V}_h \leftarrow \left(\frac{N_h - n_h}{N_h}\right) \frac{\tilde{p}_h(1-\tilde{p}_h) + \frac{1}{2\rho n_h^2}}{n_h - 1} + \frac{1}{2\rho n_h^2}.$$

4: **end for**
5: Estimate p by $\tilde{p} \leftarrow \sum_{h=1}^{H} w_h \tilde{p}_h$ and $\mathrm{Var}(\tilde{p})$ by $\widetilde{V} \leftarrow \sum_{h=1}^{H} w_h^2 \widetilde{V}_h$.
6: Return

$$\tilde{p} \pm z_{1-\alpha/2}\sqrt{\widetilde{V}},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Recall that

$$\widehat{\mathrm{Var}}(\hat{p}) = \sum_{h=1}^{H} w_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$$

is an unbiased estimator for $\mathrm{Var}(\hat{p})$. To get a private estimator for $\mathrm{Var}(\tilde{p})$, we again apply the Gaussian mechanism to $\widehat{\mathrm{Var}}(\hat{p})$ based on the sensitivity of $\mathrm{Var}(\hat{p})$:

$$\Delta_V = \max_h \left(\frac{C_h}{n_h}\left(1 - \frac{1}{n_h}\right)\right),$$

where $C_h = w_h^2 \frac{N_h - n_h}{N_h} \frac{1}{n_h - 1}$.

Since we apply the Gaussian mechanism twice, the total privacy budget should be divided into two parts: $\rho = \rho_1 + \rho_2$ to spend on adding noise to $\hat{p}$ and $\mathrm{Var}(\hat{p})$, respectively. The composition property (Proposition 2) ensures the total privacy level is ρ. The resulting algorithm, **PopNz-PubSz**, is presented in Algorithm 2.

Remark 1. *When there are multiple strata with similar sampling rates, Algorithm 1 yields a wider confidence interval for p than Algorithm 2 does, given the same privacy budget. However, Algorithm 1 additionally produces private confidence intervals for $\hat{p}_h$ which may be of interest for release. In Section 4.2.1, we compare the two algorithms quantitatively.*

Remark 2. *Proportions are always between 0 and 1. One can post-process proportion estimates ($\tilde{p}_h$ in Algorithm 1 and $\tilde{p}$ in Algorithm 2) by clipping them onto interval [0,1] without undermining privacy. When the privacy budget is very small, the noisy proportion estimates are likely to lie outside [0,1]. Thus, clipping moves the confidence interval toward the truth and a higher coverage*

Algorithm 2 Adding noise at the population level with public sample sizes, PopNz-PubSz

Input: $\hat{p}$, $\hat{p}_h$, n_h, N_h, w_h, ρ, α.
Output: A ρ-zCDP $(1-\alpha)$ CI for Population Proportion.

1: Split the budget $\rho = \rho_1 + \rho_2$.
2: Generate noise $e \sim \mathcal{N}(0, \frac{\Delta_p^2}{2\rho_1})$ where $\Delta_p = \max_h \frac{w_h}{n_h}$ and let
$$\tilde{p} \leftarrow \hat{p} + e.$$
3: Generate noise $e_V \sim \mathcal{N}(0, \frac{\Delta_V^2}{2\rho_2})$ where $\Delta_V = \max_h \left(\frac{C_h}{n_h}\left(1 - \frac{1}{n_h}\right)\right)$ and $C_h = w_h^2 \frac{N_h - n_h}{N_h} \frac{1}{n_h - 1}$. Let
$$\widetilde{V} \leftarrow \sum_{h=1}^{H} w_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1} + \frac{\Delta_p^2}{2\rho_1} + e_V.$$
4: Return
$$\tilde{p} \pm z_{1-\alpha/2}\sqrt{\widetilde{V}},$$
where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

rate will be observed. With a moderate or large budget, clipping does not make a noticeable difference.

Lastly, one can always clip the output confidence intervals onto [0,1] without privacy loss.

3.2. Estimating with private sample sizes

When sample sizes are public information, keeping the proportions private is essentially protecting only the numerator, i.e., the counts of individuals with $y = 1$. In some cases where subpopulation proportions also need to be estimated, Algorithms 1 and 2 with public sample sizes can lead to privacy leakage since the counts become the denominator. For example, one may ask the following queries: (1) what is the proportion of females in the US; and (2) what is the proportion of unemployed among females in the US. The number of females is the numerator in query (1) but becomes the denominator in query (2). Employing Algorithms 1 or 2 protects the number of females in query (1) but reveals it in query (2). Therefore, a method of constructing confidence intervals for proportions to keep both the counts and sample sizes private is necessary for subpopulation analysis. We protect the sample sizes by adding noise to them. As a result, sample sizes are not fixed and therefore we need the unbounded notion of differential privacy with the adjacency relation $\sim_r$.

In the following, we extend Algorithm 1 to serve the needs of privacy protection of sample sizes by adding noise at the stratum level. (It is not obvious how to extend Algorithm 2, which adds noise at the population level. It requires more sophisticated mechanisms; we briefly discuss in Section 6.)

To begin, we first consider the setting of simple random sampling. The idea is to add independent Gaussian noise to both the numerator and denominator

for each stratum. For ease of notation, we first consider a single stratum with count $c = \sum_{i=1}^{n} x_i$. We know

$$c \sim \text{Hypergeometric}(N, K, n),$$

where K is the total number of individuals with the attribute of interest. The count c has mean $n\frac{K}{N} = np$ and variance $n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1} = n^2 \operatorname{Var}(\hat{p})$. By applying the Gaussian mechanism to c and n with privacy budgets ρ_1 and ρ_2, respectively, we have private count $\tilde{c}$ and sample size $\tilde{n}$:

$$\tilde{c} \mid c \sim \mathcal{N}(c, \frac{1}{2\rho_1})$$

and

$$\tilde{n} \sim \mathcal{N}(n, \frac{1}{2\rho_2}).$$

The unconditional mean and variance for c are

$$\mathbb{E}(\tilde{c}) = \mathbb{E}[\mathbb{E}(\tilde{c} \mid c)] = \mathbb{E}(c) = np$$

and

$$\operatorname{Var}(\tilde{c}) = \mathbb{E}\operatorname{Var}(\tilde{c} \mid c) + \operatorname{Var}\mathbb{E}(\tilde{c} \mid c) = \frac{1}{2\rho_1} + n^2 \operatorname{Var}(\hat{p}). \tag{6}$$

By the composition property of zCDP, we get a private estimator for proportion p, denoted by $\tilde{p}$, with privacy level $\rho = \rho_1 + \rho_2$. Since $\tilde{c}$ and $\tilde{n}$ are independent variables, in principle,

$$\mathbb{E}(\tilde{p}) = \mathbb{E}\left(\frac{\tilde{c}}{\tilde{n}}\right) = \mathbb{E}(\tilde{c})\mathbb{E}\left(\frac{1}{\tilde{n}}\right), \tag{7}$$

and

$$\operatorname{Var}(\tilde{p}) = \mathbb{E}\left(\frac{\tilde{c}}{\tilde{n}}\right)^2 - \left(\mathbb{E}\left(\frac{\tilde{c}}{\tilde{n}}\right)\right)^2 = \mathbb{E}\tilde{c}^2\mathbb{E}\left(\frac{1}{\tilde{n}^2}\right) - (\mathbb{E}\tilde{c})^2\left(\mathbb{E}\frac{1}{\tilde{n}}\right)^2. \tag{8}$$

However, the moments of $\frac{1}{\tilde{n}}$ do not exist, thus neither do those of $\tilde{p}$. Generally speaking, the ratio of two independent normal random variables has a heavy-tailed distribution with no moments [34, 18]. The shape of the distribution could be unimodal, bimodal, symmetric, or asymmetric. It is primarily determined by the coefficient of variation of the denominator variable, CV. When CV is sufficiently small, a normal distribution approximation can be effective. It has been shown theoretically that a normal distribution can be arbitrarily close to the ratio variable in an interval centered at the ratio of means of two normal random variables [18]. Experiments have provided guidelines for when the normal approximation can be used. For example, a simple rule is that the approximation is reasonable whenever CV is less than 0.1 [30]. Other practical rules are mentioned in [26, 34].

We take advantage of the normal approximation to construct a ρ-zCDP confidence interval for the proportion. We present the following estimation strategy

Algorithm 3 Adding noise at the stratum level with private sample sizes, StrNz-PrivSz

Input: N_h, w_h, n_h, c_h, ρ, α.
Output: A ρ-zCDP $(1-\alpha)$ CI for the population proportion.

1: Split the budget $\rho = \rho_1 + \rho_2$.
2: **for** $h = 1$ to H **do**
3: Generate $e_h^{(1)} \sim \mathcal{N}(0, \frac{1}{2\rho_1})$ and $e_h^{(2)} \sim \mathcal{N}(0, \frac{1}{2\rho_2})$, and let

$$\begin{cases} \tilde{c}_h \leftarrow c_h + e_h^{(1)} \\ \tilde{n}_h \leftarrow \max(n_h + e_h^{(2)}, 2) \end{cases} \tag{9}$$

4: Let

$$\tilde{p}_h \leftarrow \frac{\tilde{c}_h}{\tilde{n}_h} \tag{10}$$

5: Let

$$\widetilde{V}_h \leftarrow \left(\frac{N_h - \tilde{n}_h}{N_h - 1}\right) \frac{\tilde{p}_h(1-\tilde{p}_h)}{\tilde{n}_h} + \frac{1}{2\rho_1 \tilde{n}_h^2} + \frac{\tilde{p}_h^2}{2\rho_2 \tilde{n}_h^2}. \tag{11}$$

6: **end for**
7: Estimate p by $\tilde{p} \leftarrow \sum_{h=1}^{H} w_h \tilde{p}_h$ and let $\widetilde{V} \leftarrow \sum_{h=1}^{H} w_h^2 \widetilde{V}_h$.
8: Return

$$\tilde{p} \pm z_{1-\alpha/2} \sqrt{\widetilde{V}},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

in Algorithm 3, **StrNz-PrivSz**. In the algorithm, we clip $\tilde{n}_h$ in (9) to ensure the denominator is not too small. Otherwise, the ratio can be arbitrarily large. Such a post-processing step preserves the same privacy guarantee. For the theoretical analysis, we do not clip $\tilde{n}_h$, but instead, we consider the ratio variable $\tilde{c}_h/\tilde{n}_h$ given the event $S_h = \{1 \le \tilde{n}_h \le 2n_h - 1\}$ (a symmetric area around the mean of $\tilde{n}_h$). It is more convenient for the analysis. The asymptotic behaviors of $\tilde{p}_h$ in the algorithm and $\tilde{c}_h/\tilde{n}_h \mid S_h$ are essentially the same since $\Pr(\tilde{n}_h \ge 2) \to 1$ and $\Pr(S_h) \to 1$ as $n \to \infty$. We will see the private estimator of the variance of $\tilde{p}_h$ we derive from the analysis of $\tilde{c}_h/\tilde{n}_h \mid S_h$ works well and the algorithm does achieve the desired coverage level.

We consider the ratio of two independent normal variables. By independence, what remains unclear is the behavior of the reciprocal of a normal distribution. (We should mention that the Inverse Gaussian distribution is a different distribution than the reciprocal distribution we discuss here.) In Theorem 3.1, we provide a general form of the Taylor series of conditional mean and variance of a reciprocal normal distribution. To our best knowledge, this is the first complete result of the Taylor series, with the remainder term specified. We prove the theorem in the Proofs section. We use $k = 2$ to derive an estimator for the variance of $\tilde{p}$ Algorithm 3, which leads to (11).

Theorem 3.1 (Conditional mean and variance of a reciprocal normal distribution)**.** *For random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu > 1$ and $\sigma^2 > 0$, given the event $S = \{1 \le X \le 2\mu - 1\}$, for any integer $k > 0$, the first two moments of*

$\frac{1}{X} \mid S$ *have the following expansions:*

$$\mathbb{E}\left(\frac{1}{X} \mid S\right) = \frac{1}{\mu}\sum_{j=0}^{k}\frac{(2j-1)!!\sigma^{2j}}{\mu^{2j}} + O\left(\frac{\sigma^{2k+2}}{\mu^{2k+2}}\right) \tag{12}$$

and

$$\mathbb{E}\left(\frac{1}{X^2} \mid S\right) = \frac{1}{\mu^2}\sum_{j=0}^{k}\frac{(2j+1)!!\sigma^{2j}}{\mu^{2j}} + O\left(\frac{\sigma^{2k+2}}{\mu^{2k+2}}\right). \tag{13}$$

4. Theoretical results

In this section, we present the theoretical results of both privacy and asymptotic coverage guarantees. In addition, comparisons of the three algorithms in terms of variance and width ratios are discussed.

4.1. Privacy and coverage guarantees

Our theoretical results are two-fold. First, the proposed algorithms satisfy the desired privacy level under the corresponding adjacency relation, which is presented in Theorem 4.1.

Theorem 4.1 (Privacy Guarantee)**.** *Algorithms 1 and 2 satisfy ρ-zCDP under the adjacency relation $\sim_{ss}$; Algorithm 3 satisfies ρ-zCDP under the adjacency relation $\sim_r$.*

Proofs are presented in the Proofs section.

On the other hand, for the confidence intervals to be useful, we provide theorems that guarantee the asymptotic coverage for each algorithm. The central limit theorem (CLT) asserts (essentially) that the sample mean is asymptotically normally distributed regardless of the original distribution. Therefore, the sample mean can be used to construct a confidence interval for the population mean. In the finite-population sampling designs we are considering, variants of CLTs can be found among [19, 25, 32] and others. We restate a general form of the finite-population CLT for simple random sampling in Theorem A.2 and provide asymptotic coverage guarantees in the following theorems.

Theorem 4.2 (Algorithm 1)**.** *For a population of size N, let p be the proportion in the population with the attribute of interest. Under stratified random sampling with sample sizes n_h within the stratum of size N_h, $h = 1, .., H$, let* $\widetilde{V} = \sum_{h=1}^{H} w_h^2 \widetilde{V}_h$ *where*

$$\widetilde{V}_h = \left(\frac{N_h - n_h}{N_h}\right)\frac{\tilde{p}_h(1-\tilde{p}_h) + \frac{1}{2\rho n_h^2}}{n_h - 1} + \frac{1}{2\rho n_h^2}. \tag{14}$$

for $\rho > 0$ as described in Algorithm 1. If $\rho = \omega(1/n_h)$ for all h, then as $N_h - n_h$ and n_h both tend to infinity for every stratum,

(i) $\widetilde{V} \xrightarrow{p} \mathrm{Var}(\tilde{p})$, *more specifically, for all* h,

$$\widetilde{V}_h - \mathrm{Var}(\tilde{p}_h) = \widehat{\mathrm{Var}}(\hat{p}_h) - \mathrm{Var}(\hat{p}_h) + O_P\left(\frac{1}{\sqrt{\rho} n_h^2}\right) = O_P\left(\frac{1}{n_h^{3/2}}\right),$$

where $\widehat{\mathrm{Var}}(\hat{p}_h)$ *is the non-private estimator for* $\mathrm{Var}(\hat{p}_h)$*;*

(ii) for $0 < \alpha < 1$,

$$\Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \to 1 - \alpha. \tag{15}$$

Theorem 4.3 (Algorithm 2)**.** *For a population of size* N*, let* p *be the proportion in the population with the attribute of interest. Under stratified random sampling with sample sizes* n_h *within the stratum of size* N_h*,* $h = 1, .., H$*, let*

$$\widetilde{V} = \sum_{h=1}^{H} w_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1} + \frac{\Delta_p^2}{2\rho_1} + e_V \tag{16}$$

where $e_V \sim \mathcal{N}(0, \frac{\Delta_V^2}{2\rho_2})$ *for* $\rho_1, \rho_2 > 0$ *as described in Algorithm 2. If* $\rho_1 = \omega(1/n_h)$ *and* $\rho_2 = \omega(1/n_h)$ *for all* h*, then as* $N_h - n_h$ *and* n_h *both tend to infinity for every stratum,*

(i) $\widetilde{V} \xrightarrow{p} \mathrm{Var}(\tilde{p})$, *more specifically,*

$$\widetilde{V} - \mathrm{Var}(\tilde{p}) = \widehat{\mathrm{Var}}(\hat{p}) - \mathrm{Var}(\hat{p}) + O_P\left(\frac{1}{\max_h \sqrt{\rho_2} n_h^2}\right) = O_P\left(\frac{1}{\max_h n_h^{3/2}}\right);$$

where $\widehat{\mathrm{Var}}(\hat{p})$ *is the non-private estimator for* $\mathrm{Var}(\hat{p})$*;*

(ii) for $0 < \alpha < 1$,

$$\Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \to 1 - \alpha. \tag{17}$$

Proofs of the above theorems use the finite-population CLT and are provided in the Proofs section.

For Algorithm 3, the asymptotic behavior of $\tilde{p}$ is grounded on the normal approximation to a ratio variable in addition to the CLT. We revisit the result of normal approximation by [18] in Theorem A.4. Based on the approximation, we have shown the consistency of $\tilde{p}$ in the case of simple random sampling.

Theorem 4.4. *Under simple random sampling, let* c *be the count of individuals having the attribute of interest and* n *be the sample size. The true population proportion is denoted by* p*. Let* $\tilde{p} = \tilde{c}/\tilde{n}$ *where* $\tilde{c} \sim \mathcal{N}(c, \frac{1}{2\rho_1})$ *and* $\tilde{n} \sim \mathcal{N}(n, \frac{1}{2\rho_2})$ *for* $\rho_1, \rho_2 > 0$*. Under the conditions that* $\rho_2 = \omega(1/n)$*,* $\rho_1 = \omega(1/n)$*,* $\tilde{p}$ *is a consistent estimator for* p*.*

With the foundation of the above consistency, we establish the asymptotic properties:

Theorem 4.5 (Algorithm 3)**.** *For a population of size N, let p be the proportion in the population with the attribute of interest. Under stratified random sampling with sample sizes n_h within the stratum of size N_h, $h = 1, .., H$, let $\widetilde{V} = \sum_{h=1}^{H} w_h^2 \widetilde{V}_h$ where*

$$\widetilde{V}_h = \left(\frac{N_h - \tilde{n}_h}{N_h - 1}\right) \frac{\tilde{p}_h(1-\tilde{p}_h)}{\tilde{n}_h} + \frac{1}{2\rho_1 \tilde{n}_h^2} + \frac{\tilde{p}_h^2}{2\rho_2 \tilde{n}_h^2} \tag{18}$$

for $\rho_1, \rho_2 > 0$ as described in Algorithm 3. If $\rho_1 = \omega(1/n_h)$ and $\rho_2 = \omega(1/n_h)$ for all h, then as $N_h - n_h$ and n_h both tend to infinity for every stratum,

(i) $\widetilde{V} \xrightarrow{p} \mathrm{Var}(\tilde{p} \mid S)$ where S is an event with $\Pr(S) \to 1$, more specifically, for all h,

$$\tilde{V}_h - \mathrm{Var}(\tilde{p}_h \mid S) = \widehat{\mathrm{Var}}(\hat{p}_h) - \mathrm{Var}(\hat{p}_h) + O_p\left(\frac{1}{\rho_1 n_h^2} + \frac{1}{\rho_2 n_h^2}\right) = o_p\left(\frac{1}{n_h}\right),$$

where S_h is an event with $\Pr(S_h) \to 1$;

(ii) for $0 < \alpha < 1$,

$$\Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \to 1 - \alpha. \tag{19}$$

The event S_h is discussed in Section 3.2 and S can be set to $\cap_h S_h$. As the variance of both $\tilde{p}$ and $\tilde{p}_h$ does not exist, we resort to the conditional variance under high probability events. To prove Theorem 4.5, we start with a single stratum. We use a normal distribution (denoted by p_h^*) to approximate that of the proportion estimator $\tilde{p}_h$, with the distance between the two distribution vanishing to zero in an interval. Then for multiple strata, we show that the linear combination of the normal variables (denoted by p^*) is an accurate approximation to $\tilde{p}$. Last but not least, due to the consistency stated in Theorem 4.4, the noisy estimator $\widetilde{V}$ is a consistent estimator for the variance of p^*. Then, a Wald confidence interval can be constructed using $\tilde{p}$ and $\widetilde{V}$. Details are presented in the Proofs section.

Note that, in addition to consistency for our estimates of the variance, the results above provide convergence rates. Compared to the estimation of variance in non-private settings, the additional biases are merely nuances given the conditions on ρ, ρ_1 and ρ_2. In fact, we impose these conditions to ensure that the introduced noise does not dominate when estimating the variance. In principle, these rates may be used in practice to adjust the length of confidence intervals accordingly, although we do not explore that direction here.

4.2. Comparisons of variances

The theorems presented in Section 4.1 ensure that, under proper conditions, the desired coverage is achieved asymptotically. Therefore, to compare the performance of the different proposed confidence intervals, we compare their widths, which are determined by their variance estimates. In this section, we will analyze our variance estimates and compare the resulting widths to that of the non-private confidence interval.

4.2.1. Extrinsic variances

To investigate how much additional uncertainty is caused by adding noise, we decompose the variances of the private estimators into two parts: (1) the inherent component coming from the estimation from the sampling data, i.e, $\mathrm{Var}(\hat{p})$, and (2) the extrinsic component introduced by the added noise, written as

$$V_{\mathrm{ex}} \stackrel{def}{=} \mathrm{Var}(\tilde{p}) - \mathrm{Var}(\hat{p}).$$

Table 1 provides the (approximate) variances of $\tilde{p}$ for three algorithms, where $w_h = \frac{N_h}{N}$ are the stratum weights. The variances are derived in the proofs of Theorems 4.2, 4.3, and 4.5. The additional variance terms, V_{ex}, can be rewritten in terms of $u_h \stackrel{def}{=} \frac{N_h}{n_h}$ instead of w_h, as shown in the second row of the table. In fact, u_h are called *sampling weights* in the literature on survey sampling. A sample weight is defined as the number of individuals that each respondent in the sample is representing in the population. It is the reciprocal of the sampling rate $\frac{n_h}{N_h}$ and plays an important role in statistical inference for survey data [35, 12]. Understanding the relation between sampling weights and the variance of the noisy estimators is helpful for practitioners to make survey designs and the choice of algorithms.

TABLE 1
(Approximate) variances of $\tilde{p}$.

Algorithm	StrNz-PubSz	PopNz-PubSz	StrNz-PrivSz (approximate)
$\mathrm{Var}(\tilde{p})$	$\mathrm{Var}(\hat{p}) + \frac{1}{2\rho}\sum_{h=1}^{H} \frac{w_h^2}{n_h^2}$	$\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1}\max_h \frac{w_h^2}{n_h^2}$	$\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1}\sum_{h=1}^{H} \frac{w_h^2}{n_h^2} + \frac{1}{2\rho_2}\sum_{h=1}^{H} \frac{w_h^2 p_h^2}{n_h^2}$
V_{ex}	$\frac{1}{2N^2}\sum_{h=1}^{H} \frac{u_h^2}{\rho}$	$\frac{1}{2N^2}\max_h \frac{u_h^2}{\rho_1}$	$\frac{1}{2N^2}\sum_{h=1}^{H} u_h^2(\frac{1}{\rho_1} + \frac{p_h^2}{\rho_2})$

With a fixed population size N and a chosen privacy level ρ, the extra variances V_{ex} induced by the added noise are primarily dictated by u_h. In PopNz-PubSz where we add noise at the population level, V_{ex} is solely determined by the largest sample weight among all strata. If noise is injected into each stratum, then sampling weights in all strata collectively affect V_{ex}. In particular, for StrNz-PrivSz, V_{ex} is impacted by p_h additionally. For all three algorithms, smaller sampling weights lead to lower extrinsic variance.

For comparison, we look at the ratio of V_{ex} with the budgeting $\rho_1 = \rho_2 = \rho/2$ for PopNz-PubSz and StrNz-PrivSz. The ratio of V_{ex} for StrNz-PubSz to PopNz-PubSz is

$$\frac{\sum_{h=1}^{H} u_h^2}{2 \max_h u_h^2}. \tag{20}$$

Roughly speaking, when there are many strata, adding noise at the population level gives a smaller variance. To compare StrNz-PrivSz and StrNz-PubSz, the ratio of V_{ex} is

$$\frac{2 \sum_{h=1}^{H} u_h^2 (1 + p_h^2)}{\sum_{h=1}^{H} u_h^2}, \tag{21}$$

which will always be greater than 2 (due to the cost it takes to protect sample sizes in StrNz-PrivSz) and at most 4.

4.2.2. Comparing with Non-Private CI: One Stratum Case

To assess the width in theory, we also look at the confidence interval width ratios by comparing them to the non-private one. Since the parameters N_h, n_h, p_h, ρ_h come into play together in the stratification setting, it is more practical to analyze the special case with one stratum.

Let the theoretical width ratio (TWR) be

$$\text{TWR} = \sqrt{\frac{\text{Var}(\tilde{p})}{\text{Var}(\hat{p})}}.$$

In the implementation, the real width ratio (WR), defined as $\sqrt{\widetilde{V} / \text{Var}(\hat{p})}$, will be very close to TWR in that $\widetilde{V}$ is a consistent estimator for $\text{Var}(\tilde{p})$. Table 2 displays some relevant quantities. Note that $\frac{N-1}{N-n}$ is always less than 1 but tends to 1 when the population size is far larger than the sample size.

TABLE 2
Theoretical width ratios and lower bounds. The budgeting $\rho_1 = \rho_2 = \rho/2$ are used for PopNz-PubSz and StrNz-PrivSz.

Algorithm	StrNz-PubSz	PopNz-PubSz	StrNz-PrivSz
$\tilde{p}$	$\hat{p} + \mathcal{N}(0, \frac{1}{2\rho n^2})$	$\hat{p} + \mathcal{N}(0, \frac{1}{\rho n^2})$	$(c + \mathcal{N}(0, \frac{1}{\rho}))/(n + \mathcal{N}(0, \frac{1}{\rho}))$
$\text{Var}(\tilde{p})$	$\text{Var}(\hat{p}) + \frac{1}{2\rho n^2}$	$\text{Var}(\hat{p}) + \frac{1}{\rho n^2}$	$\text{Var}(\hat{p}) + \frac{1+p^2}{\rho n^2}$
TWR	$\sqrt{1 + \frac{N-1}{N-n} \frac{1}{2p(1-p)n\rho}}$	$\sqrt{1 + \frac{N-1}{N-n} \frac{1}{p(1-p)n\rho}}$	$\sqrt{1 + \frac{N-1}{N-n} \frac{1+p^2}{p(1-p)n\rho}}$
Lower bound of TWR	$\sqrt{1 + \frac{2}{n\rho}}$	$\sqrt{1 + \frac{4}{n\rho}}$	$\sqrt{1 + \frac{2(1+\sqrt{2})}{n\rho}}$

We can obtain a lower bound for TWR by dropping the factor $\frac{N-1}{N-n}$ and minimizing over p. We can see that the width ratio mainly depends on p and the relative magnitude between n and ρ. If p is extreme (tends to 0 or 1), TWR is drastically large; when p is around 0.5, TWR is close to the lower bound. Also, the added noise induces a term involving ρ. For example, under the regime

$\rho = 1/n$, the three algorithms result in an interval of length at least $\sqrt{3} \approx 1.73$, $\sqrt{5} \approx 2.24$, and $\sqrt{3+2\sqrt{2}} \approx 2.41$ as wide, respectively. It is trivial that with one stratum, StrNz-PubSz produces a tighter confidence interval than PopNz-PubSz does in that the ratio of V_{ex} in (20) is 1/2. However, PopNz-PubSz will outperform StrNz-PubSz once there are enough strata such that (20) is greater than 1.

5. Numerical results

In this section, we conduct both simulation studies and applications to assess and illustrate the numerical performance of the proposed algorithms. The budgeting $\rho_1 = \rho_2 = \rho/2$ are used for PopNz-PubSz and StrNz-PrivSz. We clip the proportions $\tilde{p}_h$ onto $[0, 1]$ as mentioned in Remark 2.

5.1. Simulations

We set up a set of experiments to (1) check the normality of noisy estimators, and (2) evaluate the performance of the proposed confidence intervals by varying the number of strata H, the true population proportion p, and the privacy level ρ. To generate the data, we need to specify the strata sizes N_h and the sampling rates r_h. The setup of these parameters is presented in Table 3. We generate a proportion for each stratum to create heterogeneity across strata. The true population proportion is then calculated and reported in each experiment.

TABLE 3
Parameter setup. The resulting sample sizes are between 60 and 160.

Fixed parameter	Value / Distribution	Varying parameter	Value / Distribution
α	0.1	H	1 or 20
N_h	Uniform(1500, 2000)	p_h	0.5, Uniform(0.4, 0.6) or Uniform(0.05, 0.15)
r_h	Uniform(0.04, 0.08)	ρ	$1/\max(n_h)$ or specified in the axis of the plot

5.1.1. Normality Check

We first check whether the distributions of $\tilde{p}$ in the three algorithms are reasonably close to the theoretical normal distributions with the corresponding means and variances. Figure 1 displays the Q-Q plots of the theoretical distribution of $\tilde{p}$ versus its sample distribution:

- Non-private: $\mathcal{N}(p, \text{Var}(\hat{p}))$;
- StrNz-PubSz: $\mathcal{N}(p, \text{Var}(\tilde{p}))$ as $\text{Var}(\tilde{p})$ in (3);
- PopNz-PubSz: $\mathcal{N}(p, \text{Var}(\tilde{p}))$ as $\text{Var}(\tilde{p})$ in (5);
- StrNz-PrivSz: $\mathcal{N}\left(p + \sum_{h=1}^{H} \frac{w_h p_h}{2\rho_2 n_h^2}, \sum_{h=1}^{H} w_h^2 V_h\right)$ with V_h specified in (51).

Note that, $\tilde{p}$ in Algorithms StrNz-PubSz and PopNz-PubSz are unbiased for p while $\tilde{p}$ in StrNz-PrivSz is not. Nevertheless, under the condition that $\rho_2 = \omega(1/n_h)$ in Theorem 3, the bias term $\sum_{h=1}^{H} \frac{w_h p_h}{2\rho_2 n_h^2}$ is negligible and thus we

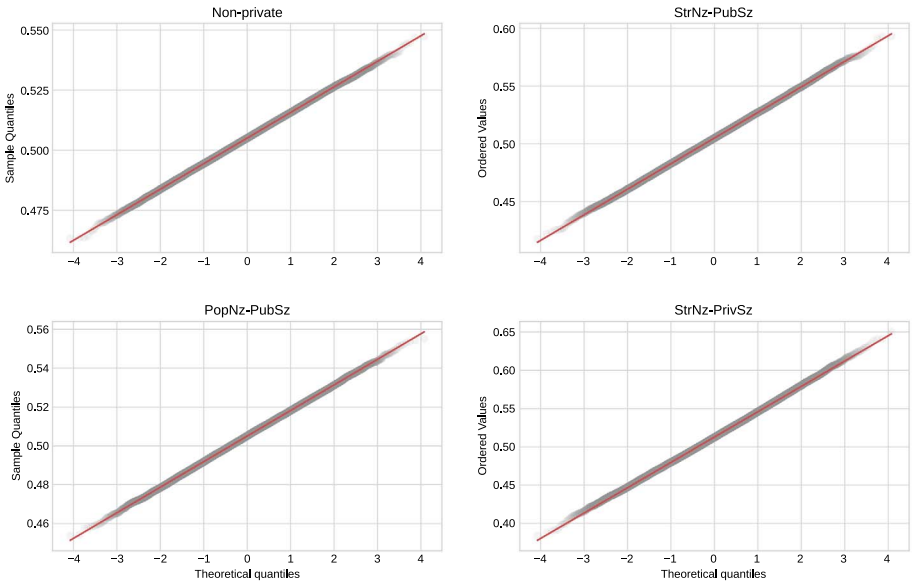


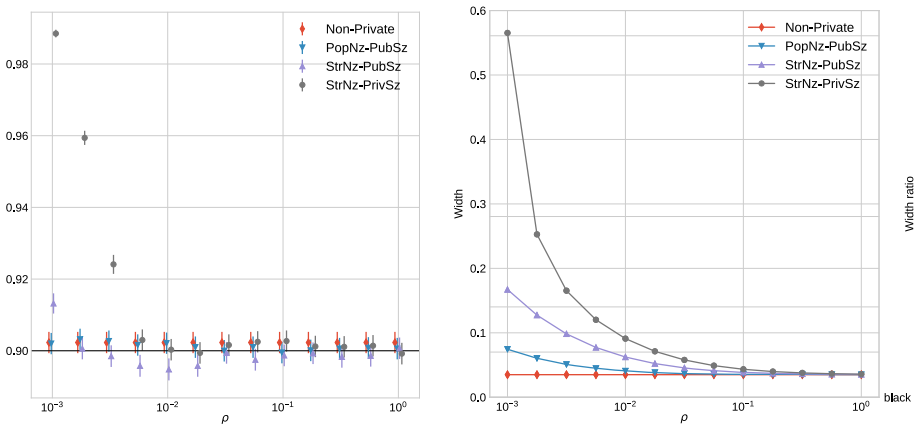
FIG 1: *Q-Q plots: Theoretical versus sample distributions of $\tilde{p}$ with 20 strata and $p = 0.505$ (resulting from $p_h \sim \text{Uniform}(0.4, 0.6)$), based on 10,000 repetitions each.*

do not make a bias correction in Algorithm 3. We observe great alignments between the theoretical and experimental distributions, indicating that the private estimators in all three algorithms are indeed highly close to being normally distributed.

5.1.2. Varying key parameters

Assured by the results of the normality check, we experiment with a wide range of the privacy budget, different numbers of strata, and true population proportions.

We examine the impact of ρ on the performance of the three private estimators. The simulation is run on 10,000 repetitions and therefore the empirical coverage falling into $90\% \pm 0.006$ (departure of two standard deviations) is considered appropriate. In Figure 2a, the empirical coverage is reasonable except that StrNz-PrivSz gives unnecessarily higher coverage when ρ is smaller than around 0.005. This is because the budget is so small for the method that, with clipping, it covers the truth more often than needed. In this case, the confidence intervals are too wide to be as useful, as shown in Figure 2b. For all three methods, the width grows as ρ becomes smaller. However, the rates of width growth differ: in the multiple strata case we simulate, the width of PopNz-PubSz grows the slowest, StrNz-PrivSz grows the fastest, and StrNz-PubSz is in the middle. Thus, the optimal privacy level should be chosen by taking into account the method, width, and coverage. For instance, if we want a 90% of confidence level



(A) *Empirical coverage* (B) *Width and ratio*

FIG 2: *Setup: 20 strata and* $p = 0.505$ *(*$p_h \sim$ Uniform*(0.4, 0.6)) with 10,000 repetitions. Figure (a) is the empirical coverage with the black solid line indicating the nominal confidence level of 90%. Error bars of one standard deviation are shown for coverage. The average width and width ratio are displayed in (b) with the non-private as the benchmark. Error bars of width are not visible in the plots and therefore not shown.*

and width under 0.1, one can choose the value for ρ as small as (1) 0.001 for PopNz-PubSz, (2) 0.003 for StrNz-PubSz, and (3) 0.01 for StrNz-PrivSz.

In addition, Table 4 shows the numerical results of three experiments with different combinations of the numbers of strata and the true population proportions. The simulation in the middle panel shares the same setting as the experiment shown in Figure 2 but has a fixed privacy level: $1/\max(n_h)$. This is an analogous regime to $\rho = 1/n$ (for simple random sampling) for multiple strata. In the literature on differential privacy, the regime $\rho = 1/n$ for a simple random sample is often considered to understand how small ρ can be as the sample size increases. Recall that a smaller ρ means a higher privacy level.

As argued above, clipping $\tilde{p}_h$ (or $\tilde{p}$) onto [0,1] will yield better results in some cases. The conclusions coincide with the analyses in Section 4. The empirical coverage of the three private ones in all simulations achieves the nominal level of 90%, as guaranteed by Theorems 4.2, 4.3, and 4.5. The case where StrNz-PrivSz gives a 91.9% confidence level in the bottom panel is due to clipping. (When the stratum proportions are close to the extreme, clipping is more noticeable.)

The average width and width ratio (WR) varies. With one single stratum, WRs are near the lower bounds of theoretical width ratios (TWR) given in Section 4.2.2, which suggests that the lower bounds are almost tight. StrNz-PubSz gives a narrower CI than PopNz-PubSz with one stratum. But with more strata, PopNz-PubSz outperforms StrNz-PubSz in terms of WR. Having more strata means splitting the total privacy budget into smaller portions, which leads to adding more noise on the whole. The CI needs to be wider to achieve the same confidence level. As for StrNz-PrivSz, however, it always yields the widest

TABLE 4

Simulation results under $\rho = 1/n$ (or $\rho = 1/\max(n_h)$) regime based on 10,000 repetitions. The strata sizes and sampling rates are drawn as described in Table 3. For the multiple strata case, the resulting sample sizes in n_h range from 72 to 152, and ρ is set to be $1/152 \approx 6.58 \times 10^{-3}$. For the one-stratum case, we set the sample size to 152 so that we have the same level of privacy.

	Non-Private	StrNz-PubSz	PopNz-PubSz	StrNz-PrivSz
	1 stratum, $p = 0.5$			
coverage	0.893	0.901	0.894	0.901
coverage SD	3.09×10^{-3}	2.99×10^{-3}	3.08×10^{-3}	2.99×10^{-3}
width	0.127	0.228	0.295	0.327
width SD	5.47×10^{-4}	9.85×10^{-4}	8.89×10^{-3}	3.15×10^{-2}
CI	(0.436, 0.564)	(0.386, 0.614)	(0.352, 0.648)	(0.34, 0.667)
WR	1	1.786	2.318	2.567
	20 strata, $p = 0.505$ ($p_h \sim \text{Uniform}(0.4, 0.6)$)			
coverage	0.902	0.895	0.902	0.902
coverage SD	2.97×10^{-3}	3.07×10^{-3}	2.97×10^{-3}	2.97×10^{-3}
width	0.035	0.073	0.043	0.111
width SD	1.08×10^{-4}	1.58×10^{-4}	5.87×10^{-4}	5.22×10^{-3}
CI	(0.488, 0.523)	(0.469, 0.542)	(0.483, 0.527)	(0.457, 0.568)
WR	1	2.074	1.239	3.168
	20 strata, $p = 0.103$ ($p_h \sim \text{Uniform}(0.05, 0.15)$)			
coverage	0.902	0.919	0.904	0.899
coverage SD	2.97×10^{-3}	2.73×10^{-3}	2.95×10^{-3}	3.01×10^{-3}
width	0.021	0.067	0.033	0.096
width SD	6.17×10^{-4}	5.21×10^{-4}	8.71×10^{-4}	3.94×10^{-3}
CI	(0.092, 0.113)	(0.073, 0.143)	(0.086, 0.119)	(0.072, 0.168)
WR	1	3.189	1.571	4.563

CI due to the additional price it pays to protect sample sizes simultaneously. On the other hand, with the same number of strata (20 here), we see that more extreme p_h leads to a larger WR than p_h in the middle range. This is because the factor $p_h(1 - p_h)$ comes into play as $p(1 - p)$ does in TWR in Table 2 for the one stratum case.

We also provide the sample standard deviation of the widths (width SD). In general, the non-private method results in a smaller standard deviation than the private ones. In some cases, clipping helps reduce the width SD for the private algorithms. With the same privacy level, there is more fluctuation in width for PopNz-PubSz compared to StrNz-PubSz. This is because we use one-half of the privacy budget and directly add noise to the variance estimate. As expected, StrNz-PrivSz has the largest width SD since the magnitude of width is the largest and the ratio variable is heavy-tailed by design. Nevertheless, compared to the width, the width SD for all methods is so small that it does not compromise the effectiveness of the confidence interval.

5.2. Applications

In this section, we apply the proposed methods to the 1940 Census full enumeration from IPUMS USA [36] and evaluate the performance of three differentially private confidence intervals. To conduct stratified random sampling on the data set, the state-level geographical variable "STATEICP" (49 categories, constituting the then-48 states and Washington, D.C.) is used for stratification. Under stratified random sampling with $H = 49$ strata, we estimate the national unemployment rate for the first application. In the second application, we are interested in studying the discrepancy in income levels between black and white men.

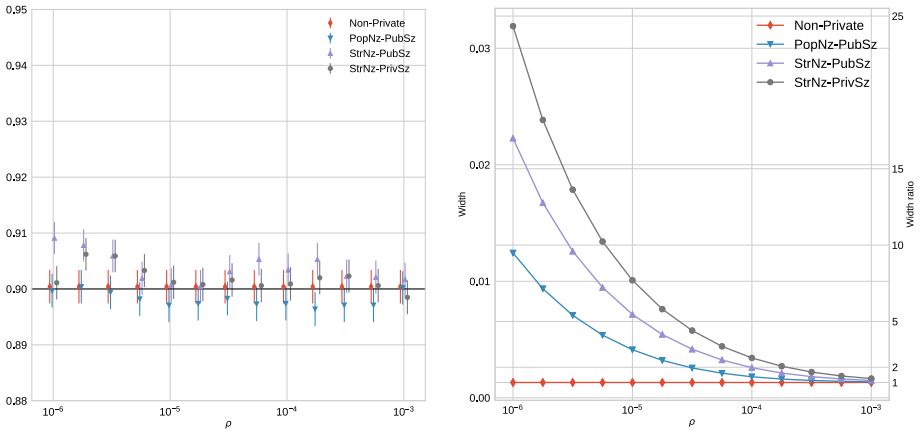
5.2.1. Confidence intervals for the unemployment rate

As an important indicator of the status of the national economy, the unemployment rate is the percentage of unemployed workers in the total labor force consisting of both the employed and unemployed. Thus, we consider all the individuals who are either employed or unemployed as the whole population. In the 1940 Census data set, the binary characteristic "EMPSTAT" represents employment status. The full enumeration is considered the truth and the true population proportion is $p = 9.346\%$. To carry out stratified random sampling, sample sizes or sampling rates are selected for all 49 strata. For modern relevance, we simulate in a manner intended to mimic the canonical design implemented in the current American Community Survey (ACS), by choosing a typical range of sampling rates used in ACS which is $[0.5\%, 15\%]$. See Table 5 for detail.

TABLE 5
Sampling rates.

Stratum size	Sampling rate
$n_h \leq 5 \times 10^4$	15%
$5 \times 10^4 < n_h \leq 10^5$	10%
$10^5 < n_h \leq 5 \times 10^5$	5%
$5 \times 10^5 < n_h \leq 10^6$	2%
$10^6 < n_h \leq 5 \times 10^6$	1%
$n_h > 5 \times 10^6$	0.5%

To apply and assess the proposed algorithms, we experiment with a wide range of small privacy budgets: $\rho \in [10^{-6}, 10^{-3}]$. Each method is repeated 10,000 times and the empirical coverage, the average CI width, and the average CI width ratio (WR) are computed. As shown in Figure 3a, the empirical coverage is always around the nominal level which is chosen at the level of 90% for the whole range of privacy levels. In Figure 3b, the CI width and CI width ratio with the non-private CI as the benchmark, share the same shape. Even when the CI given by StrNz-PrivSz is 8 times the non-private CI width, the CI width is only 0.01 due to the large sample size. Both CI width and width ratio should be taken into account when choosing an optimal privacy level.



(A) *Empirical coverage* (B) *Width and ratio*

FIG 3: *The empirical coverage with error bars, average width and width ratio of DP-CIs of the unemployment rate.*

5.2.2. Confidence intervals for the difference in income level

In the second application, we want to investigate whether there was a discrepancy between the income levels of white males (population 1) and that of black males (population 2). Note that only those who had valid income numbers in the 1940 Census are considered. Since the poverty thresholds were not developed until the 1960s and thus are not available for the 1940 data, the national income average is used as a threshold instead. We are interested in examining the difference in subpopulation proportions of those whose income levels passed this threshold.

The geographic feature "STATEICP" is used for stratification, yielding 49 strata, with stratum size ranges of $(41838, 4621442)$ for the population of white males and $(50, 309214)$ for the population of black males. Sampling rates are adaptively chosen based on stratum sizes. For the population of white males, the range of sampling rates is also $[0.5\%, 15\%]$, whereas the range of sampling rates is $[0.5\%, 30\%]$ for the population of black males given its small stratum sizes. Additionally, to allow solid approximations based on the asymptotic results, we impose that the sample sizes are adjusted to be 50 if the sampling rates give smaller sizes than 50. See Table 6 for detail.

Let p_1 and p_2 denote the proportions of eligible individuals who earned more than the national average income level \$442.12. The true values of proportions are $p_1 = 49.0223\%$ and $p_2 = 29.5152\%$. Let $p_{\text{diff}} = p_1 - p_2$, then the true difference in these two proportions is $p_{\text{diff}} = 19.5071\%$. By the additivity of two independent normal distributions, naturally, we use the following differentially private CI:

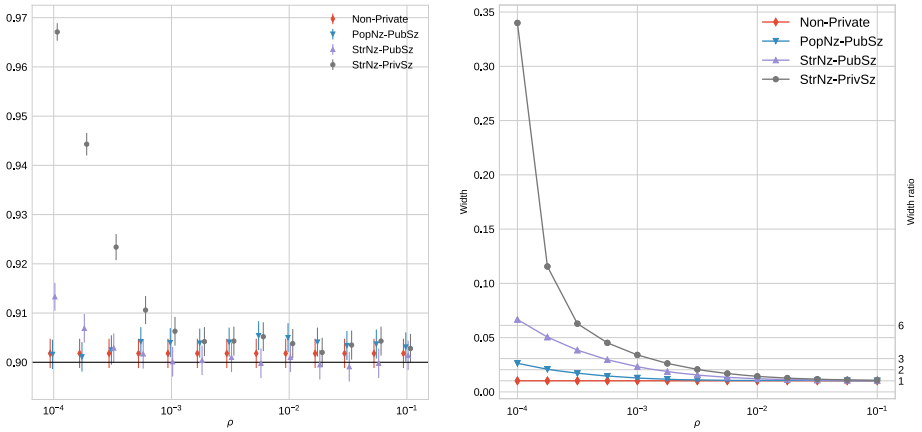
$$\tilde{p}_{\text{diff}} + z_{1-\alpha/2}\sqrt{\widetilde{\mathrm{V}}(\tilde{p}_{\text{diff}})}, \tag{22}$$

TABLE 6

*Sampling rates for two populations. Stratum sizes $n_h \in (4.1 \times 10^4, 4.7 \times 10^6)$ for the population of white males and stratum sizes $n_h \in (50, 3.1 \times 10^5)$ for the population of black males. *The sample size will be adjusted to be 50 if the above sampling rate results in a size smaller than 50.*

Stratum size N_h of white males	Sampling rate
$N_h \leq 5 \times 10^4$	15%
$5 \times 10^4 < N_h \leq 10^5$	10%
$10^5 < N_h \leq 5 \times 10^5$	5%
$5 \times 10^5 < N_h \leq 10^6$	2%
$10^6 < N_h \leq 4 \times 10^6$	1%
$N_h > 4 \times 10^6$	0.5%

Stratum size N_h of black males	Sampling rate*
$N_h \leq 500$	30%
$500 < N_h \leq 5 \times 10^3$	15%
$5 \times 10^3 < N_h \leq 10^4$	5%
$10^4 < N_h \leq 2 \times 10^4$	2%
$2 \times 10^4 < N_h \leq 3 \times 10^4$	1%
$n_h > 3 \times 10^4$	0.5%



(A) Empirical coverage *(B) Width and ratio*

FIG 4: *The empirical coverage with error bars, average width and width ratio of DP-CIs of the difference of the above-national-income-level proportions between black and white males with valid income values.*

where $\widetilde{V}(\cdot)$ denotes a private estimator of variance, $\tilde{p}_{\text{diff}}$ is defined as $\tilde{p}_1 - \tilde{p}_2$ and

$$\widetilde{\mathrm{V}}(p_{\text{diff}}) = \widetilde{\mathrm{V}}(p_1) + \widetilde{\mathrm{V}}(p_2).$$

In Figure 4, similar patterns are observed in this application as in the first. All CIs have empirical coverage around/above the nominal confidence level as in the simulation study in Section 5.1.2. The phenomenon of higher coverage is due to small ρ and effective clipping. When the range of stratum sizes is large (it is (50, 309214) in this application), that is, when the stratum sizes are very different, a large privacy budget ρ should be chosen. The choice of a small ρ harms the estimates of small-sized strata. We advise that the smallest ρ be chosen given the tolerance of uncertainty in terms of width and/or width ratio. For example, if the accuracy requirement is that the width should be under 0.05 or WR under 5, then the best choices of ρ among the experiments in Figure 4b are (1) 0.0001 for PopNz-PubSz, (2) 0.0018 for StrNz-PubSz, and (3) 0.0056 for StrNz-PrivSz.

6. Discussion

We have designed three algorithms to construct confidence intervals for the population proportion under stratified random sampling with zero concentrated differential privacy guarantees. We consider both the case where the sample sizes are public and the case where they are private information. Theoretical results including privacy guarantees and asymptotic properties are established. With proper conditions on the relation between the privacy budget and sample sizes, as stated in the theorems, the resulting confidence intervals will achieve the desired coverage asymptotically, and the width tends to be that of a non-private confidence interval when the sample sizes go to infinity.

In the simulation studies and two applications, we have experimented with a wide range of privacy budgets under a variety of parameter setups. The three algorithms always perform well in terms of empirical coverage. The width and width ratio are in a reasonable range even under the strict regime where $\rho = 1/\max_h n_h$. Typically in practice, a constant between 0.001 to 10 is chosen to be the privacy budget. According to our experiments, with the choice of the smallest budget in this range, 0.001, the three algorithms still have fairly good results even when the smallest stratum has only a size 50 (as demonstrated in the second application).

The comparative analysis of the three algorithms in Section 4.2 gives actionable guidance to practitioners. When releasing the population proportion is the only goal and there are enough strata (such that Eq.(20) regarding sample weights is greater than 1), PopNz-PubSz is the better option. However, if stratum proportions should also be released or there are just a few strata, StrNz-PubSz is preferable. On the other hand, when the population proportion and sample sizes must be protected simultaneously, StrNz-PrivSz is the only algorithm presented in this paper. StrNz-PrivSz, compared to the case with public sample sizes, needs a larger budget to meet the same width requirement on account of the additional cost of protecting sample sizes.

There are a few open questions worth considering for future research. In this paper, we discuss the classic case where the number of strata is fixed, and the sample sizes tend to infinity. In principle, asymptotic normality is also valid in other settings with finite sample sizes. For example, it has been shown in the non-private setting that as the total sample size $N \to \infty$ with many small samples or a few large samples, or some combination thereof, central limit theorems hold under certain (complex) conditions [2]. Under the constraints of differential privacy, we have shown that the trade-off between the privacy and accuracy (a.k.a., utility) of DP-CIs depends on the smallest sample, i.e, recall the condition, $\rho = \omega(1/n_h)$ for all h, in Section 4.1. The overall privacy loss is determined by the largest privacy loss among all strata. However, when the strata sizes N_h remain finite, the weights ($w_h = N_h/N$) of these strata tend to 0 as $N \to \infty$. Therefore, the noise injected into these small strata should not harm the overall accuracy of the intervals if N is sufficiently large.

More interestingly, we do not provide 'PopNz-PrivSz' – an analogous algorithm to PopNz-PubSz for the private sample sizes case. To protect both the

population proportion and the sample sizes, the direct addition of noise to the non-private aggregated estimator is not plausible. One should consider more sophisticated mechanisms other than directly adding noise to the statistics. If 'PopNz-PrivSz' were proposed, we shall expect it to yield a narrower confidence interval since we only need to publish the private population proportion without being able to provide private confidence intervals for stratum proportions at the same time.

Another direction for future research would be optimal budget allocation. We do not discuss how to best divide the total budget for PopNz-PubSz or StrNz-PrivSz. Budgeting for the composed application of the algorithms may also be of interest, like in Section 5.2.2 where we apply the algorithms twice for two independent populations.

Lastly, one broad direction is to develop the differentially private versions for other alternatives to the basic Wald interval, such as the Wilson Interval, Jeffreys interval, etc.(see [22] for a comparative summary of seven such types of confidence intervals for proportions). Many of these latter are specifically designed for the case of small sample sizes, which we do not consider here and for which we expect fundamentally different approaches to differential privacy likely to be necessary.

Appendix A: Proofs

A.1. Proof of Theorem 3.1

Lemma A.1. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $S = \{\mu - a \le X \le \mu + a\}$. For any $a > 0$ and an integer $k \ge 1$, the conditional even moments*

$$\mathbb{E}[(X-\mu)^{2k} \mid S] = \sigma^{2k}(2k-1)!! - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k-1}\right), \tag{23}$$

where the big-O hides a constant depending on σ and k.

Proof. Without loss of generality, we assume $\mu = 0$. We prove the lemma by induction. Set $k = 1$, integrate by parts,

$$\begin{aligned}\mathbb{E}[X^2 I_S] &= \int_{-a}^{a} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &= \frac{\sigma}{\sqrt{2\pi}} \left(-x e^{-\frac{x^2}{2\sigma^2}} \Big|_{-a}^{a} + \int_{-a}^{a} e^{-\frac{x^2}{2\sigma^2}}\, dx \right).\end{aligned}$$

Integrate by substitution, the integral in the second term becomes

$$\int_{-a}^{a} e^{-\frac{x^2}{2\sigma^2}}\, dx = \sigma\sqrt{2\pi}\, \mathrm{erf}\left(\frac{a}{\sigma\sqrt{2}}\right)$$

where

$$\mathrm{erf}(z) = \sigma \int_0^z e^{-t^2}\, dt$$

is the *error function*. Then,

$$\mathbb{E}[X^2 I_S] = \sigma^2 \operatorname{erf}\left(\frac{a}{\sigma\sqrt{2}}\right) - O\left(e^{-\frac{a^2}{2\sigma^2}} a\right).$$

Assuming

$$\mathbb{E}[X^{2k} I_S] = \sigma^{2k}(2k-1)!! \operatorname{erf}\left(\frac{a}{\sigma\sqrt{2}}\right) - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k-1}\right), \tag{24}$$

then integrate by parts for the $k+1$ case,

$$\begin{aligned}
\mathbb{E}[X^{2(k+1)} I_S] &= \int_{-a}^{a} x^{2k+2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-a}^{a} x^{2k+1} \cdot \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}\, dx \\
&= \frac{\sigma}{\sqrt{2\pi}} \left(-x^{2k+1} e^{-\frac{x^2}{2\sigma^2}} \Big|_{-a}^{a} + (2k+1) \int_{-a}^{a} x^{2k} e^{-\frac{x^2}{2\sigma^2}}\, dx\right) \\
&= \sigma^2(2k+1)\mathbb{E}[X^{2k} I_S] - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k+1}\right).
\end{aligned}$$

Plug in (24), we obtain

$$\mathbb{E}[X^{2(k+1)} I_S] = \sigma^{2k+2}(2k+1)!! \operatorname{erf}\left(\frac{a}{\sigma\sqrt{2}}\right) - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k+1}\right).$$

So far we have proved (24). Note that

$$\Pr(S) = \int_{-a}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\, dx = \operatorname{erf}\left(\frac{a}{\sigma\sqrt{2}}\right),$$

and that the image of $\operatorname{erf}(z)$ is between $(-1, 1)$. Therefore,

$$\mathbb{E}[X^{2k} \mid S] = \mathbb{E}[X^{2k} I_S] / \Pr(S) = \sigma^{2k}(2k-1)!! - O\left(e^{-\frac{a^2}{2\sigma^2}} a^{2k-1}\right).$$

□

Proof of (12) in Theorem 3.1. Consider the Taylor series of $\frac{1}{x}$ at $x = \mu$:

$$\frac{1}{x} = \sum_{j=0}^{\infty} \frac{(-(x-\mu))^j}{\mu^{j+1}} = \frac{1}{\mu} - \frac{x-\mu}{\mu^2} + \frac{(x-\mu)^2}{\mu^3} - \frac{(x-\mu)^3}{\mu^4} + \cdots$$

Let y_m be the partial sum of the above series, i.e., $y_m(x) = \sum_{k=0}^{m} \frac{(-(x-\mu))^k}{\mu^{k+1}}$. Then $y_m(x)$ converges to $\frac{1}{x}$ in $(0, 2\mu)$ which contains $[1, 2\mu - 1]$. Let

$$g(x) = \sum_{k=0}^{\infty} \frac{|x-\mu|^k}{\mu^{k+1}} = \begin{cases} \frac{1}{x}, & \text{if } 1 \le x \le \mu \\ \frac{1}{2\mu - x} & \text{if } \mu < x \le 2\mu - 1 \end{cases}$$

Then g is integrable as

$$\int_1^{2\mu-1} |g(x)|d\nu = \int_1^{\mu} \frac{1}{x}d\nu + \int_{\mu}^{2\mu-1} \frac{1}{2\mu - x}d\nu = 2\int_1^{\mu} \frac{1}{x}d\nu < \infty,$$

where $d\nu = f(x)dx$ is induced by $\mathcal{N}(\mu, \sigma^2)$ conditional on event S. Note also that $|y_m(x)| \leq g(x)$ for any naturals m and $x \in [1, 2\mu - 1]$. By the dominated convergence theorem, the operations of limit and integral are exchangeable for $y_m(x)$.

$$\begin{aligned}
\int_1^{2\mu-1} \frac{1}{x}d\nu &= \int_1^{2\mu-1} \lim_{m\to\infty} y_m(x)d\nu \\
&= \lim_{m\to\infty} \int_1^{2\mu-1} y_m(x)d\nu \\
&= \lim_{m\to\infty} \int_1^{2\mu-1} \left(\sum_{j=0}^{m} \frac{(-(x-\mu))^j}{\mu^{j+1}} \right) d\nu \\
&= \lim_{m\to\infty} \left(\sum_{j=0}^{m} \int_1^{2\mu-1} \frac{(-(x-\mu))^j}{\mu^{j+1}}d\nu \right)
\end{aligned} \tag{25}$$

Then,

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{X} \mid S\right) &= \sum_{j=0}^{\infty} \frac{1}{\mu^{j+1}} \mathbb{E}\left[(-(X-\mu))^j \mid S\right] \\
&= \sum_{j=0}^{\infty} \frac{1}{\mu^{2j+1}} \mathbb{E}\left[(X-\mu)^{2j} \mid S\right] \\
&= \sum_{j=0}^{k} \frac{1}{\mu^{2j+1}} \mathbb{E}\left[(X-\mu)^{2j} \mid S\right] + \frac{1}{\mu} \sum_{j=k+1}^{\infty} \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2j} \mid S\right].
\end{aligned} \tag{26}$$

The second equality is because the odd moments are zero due to symmetry.

Note that given event S, $|\frac{X-\mu}{\mu}| \leq \frac{\mu-1}{\mu} < 1$, then

$$\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right] \leq \left(\frac{\mu-1}{\mu}\right)^2 \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k} \mid S\right]. \tag{27}$$

It follows that

$$\begin{aligned}
\sum_{j=k+1}^{\infty} \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2j} \mid S\right] &\leq \sum_{j=0}^{\infty} \left(\frac{\mu-1}{\mu}\right)^{2j} \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right] \\
&= \frac{\mu^2}{2\mu-1} \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right] \\
&= O\left(\frac{1}{\mu^{2k+1}}\right) \cdot \mathbb{E}[(X-\mu)^{2k+2} \mid S].
\end{aligned}$$

Applying Lemma A.1, by the choice of $a = \mu - 1$, (26) becomes

$$\mathbb{E}\left(\frac{1}{X} \mid S\right) = \frac{1}{\mu}\sum_{j=0}^{k}\frac{(2j-1)!!\sigma^{2j}}{\mu^{2j}} + O\left(\frac{\sigma^{2k+2}}{\mu^{2k+2}}\right).$$

□

Proof of (13) in Theorem 3.1. We conduct a similar procedure for the second moment of $X \mid S$. Based on the Taylor expansion

$$\frac{1}{x^2} = \sum_{j=0}^{\infty}\frac{(j+1)(-(x-\mu))^j}{\mu^{j+2}} = \frac{1}{\mu^2} - \frac{2(x-\mu)}{\mu^3} + \frac{3(x-\mu)^2}{\mu^4} - \frac{4(x-\mu)^3}{\mu^5} + \cdots,$$

we have

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{X^2} \mid S\right) &= \sum_{j=0}^{\infty}\frac{j+1}{\mu^{j+2}}\mathbb{E}\left[(-(X-\mu))^j \mid S\right] \\
&= \sum_{j=0}^{\infty}\frac{2j+1}{\mu^{2j+2}}\mathbb{E}\left[(X-\mu)^{2j} \mid S\right] \\
&= \sum_{j=0}^{k}\frac{2j+1}{\mu^{2j+2}}\mathbb{E}\left[(X-\mu)^{2j} \mid S\right] \\
&\quad + \frac{1}{\mu^2}\sum_{j=k+1}^{\infty}(2j+1)\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2j} \mid S\right].
\end{aligned} \tag{28}$$

Due to (27), it follows that

$$\begin{aligned}
&\sum_{j=k+1}^{\infty}(2j+1)\mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2j} \mid S\right] \\
&\le \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right]\cdot\sum_{j=0}^{\infty}(2k+3+2j)\left(\frac{\mu-1}{\mu}\right)^{2j} \\
&= \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right]\cdot\left[(2k+3)\sum_{j=0}^{\infty}\left(\frac{\mu-1}{\mu}\right)^{2j} + 2\sum_{j=1}^{\infty}j\left(\frac{\mu-1}{\mu}\right)^{2j}\right] \\
&= \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right]\cdot\left[\frac{(2k+3)\mu^2}{2\mu-1} + 2\frac{\mu^2(\mu-1)^2}{(2\mu-1)^2}\right] \\
&= \mathbb{E}\left[\left(\frac{X-\mu}{\mu}\right)^{2k+2} \mid S\right]\cdot O(\mu^2) \\
&= O\left(\frac{1}{\mu^{2k}}\right)\cdot\mathbb{E}[(X-\mu)^{2k+2} \mid S],
\end{aligned} \tag{29}$$

where the term $\sum_{j=1}^{\infty} j \left(\frac{\mu-1}{\mu}\right)^{2j}$ is a sum of an arithmetic–geometric sequence. By Lemma A.1, (28) becomes

$$\mathbb{E}\left(\frac{1}{X^2} \mid S\right) = \frac{1}{\mu^2} \sum_{j=0}^{k} \frac{(2j+1)!!\sigma^{2j}}{\mu^{2j}} + O\left(\frac{\sigma^{2k+2}}{\mu^{2k+2}}\right). \tag{30}$$

□

A.2. Proof of Theorem 4.1

Proof for Algorithm 1. Under neighboring relation $\sim_{ss}$, only one record changes within one stratum and sample sizes remain the same. Applying the Gaussian mechanism to each stratum at the level of ρ gives ρ-zCDP guarantee. By post-processing, the confidence interval is also ρ-zCDP. □

Proof for Algorithm 2. The sensitivities of $\hat{p}$ and $\widehat{\mathrm{Var}}(\hat{p})$ are Δ_p and Δ_V, respectively. Applying the Gaussian mechanism, it follows that $\tilde{p}$ is ρ_1-zCDP and $\widetilde{V}$ is ρ_2-zCDP. By basic composition, the confidence interval $\tilde{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\widetilde{V}}$ is $(\rho_1+\rho_2)$-zCDP. □

Proof for Algorithm 3. By the Gaussian mechanism and the basic composition property of zCDP, we know that $\tilde{p}_h$ is ρ-zCDP. Under neighboring relation $\sim_r$, only one record changes within one stratum. Then, by post-processing, the confidence interval is ρ-zCDP. □

A.3. Proof of Theorem 4.2

Before proving the theorem, we revisit the finite-population CLT first:

Theorem A.2 (Theorem 1, [33]). *Consider a finite population* $\Pi = \{X_1, ..., X_N\}$ *of size* N. *Let* μ *be the population mean and* $\bar{X}_n$ *be the mean of a simple random sample of size* n *from* Π, *and* $\mathrm{Var}(\bar{X}_n)$ *is the variance of* $\bar{X}_n$. *The finite population variance of* Π *is denoted by*

$$v = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2.$$

As $N \to \infty$, *if*

$$\frac{1}{\min(n, N-n)} \cdot \frac{\max_{1\le i\le N}(X_i-\mu)^2}{v} \to 0, \tag{31}$$

we have

$$\frac{\bar{X}_n - \mu}{\sqrt{\mathrm{Var}(\bar{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1). \tag{32}$$

The variance of $\bar{X}_n$ is determined by the population variance v which is unknown. Nevertheless, the sample variance $\widehat{\mathrm{Var}}(\bar{X}_n)$ can be used to estimate v. To make sure the CLT still holds when substituting $\mathrm{Var}(\bar{X}_n)$ by $\widehat{\mathrm{Var}}(\bar{X}_n)$, the consistency of $\widehat{\mathrm{Var}}(\bar{X}_n)$ is crucial, as stated in the following lemma.

Lemma A.3. *Let $\widehat{\mathrm{Var}}(\bar{X}_n)$ be the sample variance. $\widehat{\mathrm{Var}}(\bar{X}_n)$ is an unbiased estimator for $\mathrm{Var}(\bar{X}_n)$. Moreover, under the condition in Theorem A.2, as $N \to \infty$,*

$$\widehat{\mathrm{Var}}(\bar{X}_n)/\mathrm{Var}(\bar{X}_n) \xrightarrow{p} 1.$$

Now we prove Theorem 4.2:

Proof of Theorem 4.2. It suffices to show $\frac{\tilde{p}_h - p_h}{\sqrt{\widetilde{V}_h}} \xrightarrow{d} \mathcal{N}(0,1)$ for all h. By the finite-population CLT in Theorem A.2, we know

$$\frac{\hat{p}_h - p_h}{\sqrt{\mathrm{Var}(\hat{p}_h)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Since $\tilde{p}_h = \hat{p}_h + e_h$ where $e_h \sim \mathcal{N}(0, \frac{1}{2\rho n_h^2})$, we have

$$\frac{\tilde{p}_h - p_h}{\sqrt{\mathrm{Var}(\tilde{p}_h)}} \xrightarrow{d} \mathcal{N}(0,1) \tag{33}$$

where

$$\mathrm{Var}(\tilde{p}_h) = \mathrm{Var}(\hat{p}_h) + \frac{1}{2\rho n_h^2}.$$

Let

$$\begin{aligned} \widetilde{V}_h &= \left(\frac{N_h - n_h}{N_h}\right) \frac{\tilde{p}_h(1-\tilde{p}_h) + \frac{1}{2\rho n_h^2}}{n_h - 1} + \frac{1}{2\rho n_h^2} \\ &= \widehat{\mathrm{Var}}(\hat{p}_h) + \left(\frac{N_h - n_n}{N_h}\right) \frac{e_h - 2\tilde{p}_h e_h - e_h^2 + \frac{1}{2\rho n_h^2}}{n_h - 1} + \frac{1}{2\rho n_h^2}. \end{aligned} \tag{34}$$

Since $e_h \sim \mathcal{N}(0, \frac{1}{2\rho n_h^2})$, we have $e_h = O_P(\frac{1}{\sqrt{\rho} n_h})$. Then, the second term of (34) is $O_P(\frac{1}{\sqrt{\rho} n_h^2})$, and thus, $\widetilde{V}_h - \mathrm{Var}(\tilde{p}_h) = \widehat{\mathrm{Var}}(\hat{p}_h) - \mathrm{Var}(\hat{p}_h) + O_P\left(\frac{1}{\sqrt{\rho} n_h^2}\right)$. Note that $\widehat{\mathrm{Var}}(\hat{p}_h)$ is of order $\frac{1}{n_h}$, and that by Lemma A.3, $\widehat{\mathrm{Var}}(\hat{p}_h) \xrightarrow{p} \mathrm{Var}(\hat{p}_h)$. Therefore, $\widetilde{V}_h \xrightarrow{p} \mathrm{Var}(\tilde{p}_h)$, and thus, $\widetilde{V} \xrightarrow{p} \mathrm{Var}(\tilde{p})$.

Combining the consistency of $\widetilde{V}$ with (33), we have

$$\frac{\tilde{p}_h - p_h}{\sqrt{\widetilde{V}_h}} \xrightarrow{d} \mathcal{N}(0,1) \tag{35}$$

by Slutsky's Theorem. Then, $\frac{\tilde{p}-p}{\sqrt{\widetilde{V}}} \xrightarrow{d} \mathcal{N}(0,1)$. Therefore, the confidence interval given by $p \pm z_{1-\alpha/2}\sqrt{\widetilde{V}}$ has asymptotic coverage level $1-\alpha$. □

A.4. Proof of Theorem 4.3

Proof. Since $\frac{\hat{p}-p}{\sqrt{\mathrm{Var}(\hat{p})}} \xrightarrow{d} \mathcal{N}(0,1)$ and $\tilde{p} = \hat{p} + \mathcal{N}(0, \Delta_p^2/2\rho_1)$ with $\Delta_p = \max_h \frac{w_h}{n_h}$, it follows that

$$\frac{\tilde{p}-p}{\sqrt{\mathrm{Var}(\tilde{p})}} \xrightarrow{d} \mathcal{N}(0,1),$$

and

$$\mathrm{Var}(\tilde{p}) = \mathrm{Var}(\hat{p}) + \frac{\Delta_p^2}{2\rho_1}.$$

In Algorithm 2, we set

$$\widetilde{V} = \widehat{\mathrm{Var}}(\hat{p}) + \frac{\Delta_p^2}{2\rho_1} + e_V, \tag{36}$$

where $e_V \sim \mathcal{N}(0, \frac{\Delta_V^2}{2\rho_2})$ with $\Delta_V = \max_h \left(\frac{C_h}{n_h}\left(1 - \frac{1}{n_h}\right)\right)$ and $C_h = w_h^2 \frac{N_h - n_h}{N_h} \frac{1}{n_h - 1}$. Since $\Delta_V = O(\frac{1}{\max_h n_h^2})$, we have $e_V = O_P(\frac{1}{\max_h \sqrt{\rho_2} n_h^2})$. Thus, $\widetilde{V} - \mathrm{Var}(\tilde{p}) = \widehat{\mathrm{Var}}(\hat{p}) - \mathrm{Var}(\hat{p}) + O_P\left(\frac{1}{\max_h \sqrt{\rho} n_h^2}\right)$. Since $\widehat{\mathrm{Var}}(\hat{p}) \xrightarrow{p} \mathrm{Var}(\hat{p})$ by finite-population CLT, we have $\widetilde{V} \xrightarrow{p} \mathrm{Var}(\tilde{p})$.

Therefore, by Slutsky's Theorem,

$$\frac{\tilde{p}-p}{\sqrt{\widetilde{V}}} \xrightarrow{d} \mathcal{N}(0,1). \tag{37}$$

Then, the confidence interval given by $p \pm z_{1-\alpha/2}\sqrt{\widetilde{V}}$ has the asymptotic coverage level $1-\alpha$. □

A.5. Proof of Theorem 4.4

Proof. For $\tilde{n} \sim \mathcal{N}(n, \frac{1}{2\rho_2})$, by Proposition 3.1, we derive the kth-order Taylor series of the conditional expectation of $\tilde{p}$ given $S = \{1 \le \tilde{n} \le 2n-1\}$:

$$\mathbb{E}(\tilde{p} \mid S) = p \sum_{j=0}^{k} \frac{(2j-1)!!}{n^{2j}(2\rho_2)^j} + O\left(\frac{1}{n^{2k+1}\rho_2^{k+1}}\right). \tag{38}$$

For example, when $k = 2$,

$$\mathbb{E}(\tilde{p} \mid S) = p\left(1 + \frac{1}{2n^2\rho_2} + \frac{3}{4n^4\rho_2^2}\right) + O\left(\frac{1}{n^5\rho_2^3}\right). \tag{39}$$

To obtain a Taylor expansion for the conditional variance, we plug

$$\mathbb{E}\left(\frac{1}{\tilde{n}} \mid S\right) = \frac{1}{n} \sum_{j=0}^{k} \frac{(2j-1)!!}{n^{2j}(2\rho_2)^j} + O\left(\frac{1}{n^{2k+2}\rho_2^{k+1}}\right)$$

and

$$\mathbb{E}\left(\frac{1}{\tilde{n}^2} \mid S\right) = \frac{1}{n^2}\sum_{j=0}^{k}\frac{(2j+1)!!}{n^{2j}(2\rho_2)^j} + O\left(\frac{1}{n^{2k+2}\rho_2^{k+1}}\right)$$

into

$$\mathrm{Var}(\tilde{p} \mid S) = \mathbb{E}(\tilde{p}^2 \mid S) - (\mathbb{E}(\tilde{p} \mid S))^2 = \mathbb{E}\tilde{c}^2\mathbb{E}\left(\frac{1}{\tilde{n}^2} \mid S\right) - (\mathbb{E}(\tilde{p} \mid S))^2,$$

by which we derive a general expansion for the conditional variance:

$$\begin{aligned}\mathrm{Var}(\tilde{p} \mid S) = \mathrm{Var}(\hat{p})\sum_{j=0}^{k}\frac{(2j+1)!!}{n^{2j}(2\rho_2)^j} + p^2\left(\sum_{j=0}^{k}\frac{(2j+1)!!}{n^{2j}(2\rho_2)^j} - \left(\sum_{j=0}^{k}\frac{(2j-1)!!}{n^{2j}(2\rho_2)^j}\right)^2\right)\\ + \frac{1}{2\rho_1}\sum_{j=0}^{k}\frac{(2j+1)!!}{n^{2j+2}(2\rho_2)^j} + O\left(\frac{1}{n^{2k}\rho_2^{k+1}}\right) + O\left(\frac{1}{n^{2k+2}\rho_1\rho_2^{k+1}}\right).\end{aligned} \tag{40}$$

When $k = 2$,

$$\begin{aligned}\mathrm{Var}(\tilde{p} \mid S) = \mathrm{Var}(\hat{p})\left(1 + \frac{3}{2n^2\rho_2} + \frac{15}{4n^4\rho_2^2}\right)\\ + p^2\left(\frac{1}{2n^2\rho_2} + \frac{2}{n^4\rho_2^2} - \frac{6}{8n^6\rho_2^3} - \frac{9}{16n^8\rho_2^4}\right)\\ + \frac{1}{2\rho_1}\left(\frac{1}{n^2} + \frac{3}{2n^4\rho_2} + \frac{15}{4n^6\rho_2^2}\right) + O\left(\frac{1}{n^4\rho_2^3}\right) + O\left(\frac{1}{n^6\rho_1\rho_2^3}\right).\end{aligned} \tag{41}$$

Based on Taylor expansion with $k = 2$ for both conditional mean and variance given in (39) and (41), under the condition $\frac{1}{\rho_1 n} = o(1)$ and $\frac{1}{\rho_2 n} = o(1)$, we have

$$\mathbb{E}(\tilde{p} \mid S) = p + o\left(\frac{1}{n}\right)$$

and

$$\mathrm{Var}(\tilde{p} \mid S) = \mathrm{Var}(\hat{p}) + o\left(\frac{1}{n}\right).$$

Then, $\tilde{p} \mid S$ is asymptotically unbiased. Note that $\mathrm{Var}(\hat{p})$ is of order $\frac{1}{n}$ and thus $\tilde{p} \mid S$ has a vanishing variance. Therefore, $\tilde{p} \mid S$ converges to p in probability. Note also that $\Pr(S) \to 1$ as $n \to \infty$, then for any $\epsilon > 0$,

$$\Pr(|\tilde{p} - p| > \epsilon) = \Pr(|\tilde{p} - p| > \epsilon \mid S) + \Pr(|\tilde{p} - p| > \epsilon \mid S^{\mathsf{c}}) \to 0.$$

That is, $\tilde{p}$ is a consistent estimator for p. □

A.6. Proof of Theorem 4.5

To prove Theorem 4.5, we need the following theorem and lemmas.

Theorem A.4 (Theorem 1, [18]). *Let X be a normal random variable with positive mean μ_x, variance σ_x^2 and coefficient of variation $\delta_x = \sigma_x/\mu_x$ such that $0 < \delta_x < \lambda \leq 1$, where λ is a known constant. For every $\epsilon > 0$, there exists $\gamma(\epsilon) \in (0, \sqrt{\lambda^2 - \delta_x^2})$ and also a normal random variable Y independent of X, with positive mean μ_y, variance σ_y^2 and coefficient of variation $\delta_y = \sigma_y/\mu_y$ that satisfy the conditions,*

$$0 < \delta_y \leq \gamma(\epsilon) \leq \sqrt{\lambda^2 - \delta_x^2} < \lambda \tag{42}$$

for which the following result holds. Any z that belongs to the interval

$$I = \left[\beta - \frac{\sigma_z}{\lambda}, \beta + \frac{\sigma_z}{\lambda}\right],$$

where $\beta = \mu_x/\mu_y$, $\sigma_z = \beta\sqrt{\delta_x^2 + \delta_y^2}$, satisfies that

$$|G(z) - F_Z(z)| < \epsilon,$$

where $G(z)$ is the cumulative distribution function of $\mathcal{N}(\beta, \sigma_z^2)$, and F_Z is that of $Z = X/Y$. Note that once a given Y fulfills the closeness between the corresponding G to F_Z, any other random variables with a smaller coefficient of variation will satisfy this result too.

Lemma A.5. *For a population of size N, let p be the true proportion in the population with the attribute of interest. Consider simple random sampling with sample size n. Let $Z^* \sim \mathcal{N}(p, V)$ where $V = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n} + \frac{1}{2\rho_1 n^2} + \frac{p^2}{2\rho_2 n^2}$. If $\rho_1 = \omega(1/n^2)$ and $\rho_2 = \omega(1/n)$, as $N - n$ and n both tend to infinity, then for any $z \in (0, 2p)$,*

$$|F_{\tilde{p}}(z) - F_{Z^*}(z)| \to 0. \tag{43}$$

Proof. By the CLT in Theorem A.2, we know that $\hat{p} \sim \mathcal{AN}(p, \mathrm{Var}(\hat{p}))$. Recall that $\tilde{c} = n\hat{p} + \mathcal{N}(n, \frac{1}{2\rho_1})$, then $\tilde{c} \sim \mathcal{AN}(np, n^2\,\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1})$.

Let $\tilde{X} \sim \mathcal{AN}(np, n^2\,\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1})$ and $X \sim \mathcal{N}(np, n^2\,\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1})$. Therefore, for any $\epsilon > 0$, there exists some $n_0 = n_0(\epsilon)$ such that for any x and $n > n_0$,

$$|F_{\tilde{X}}(x) - F_X(x)| < \epsilon, \tag{44}$$

where F denotes the cumulative density function. Let $Y \sim \mathcal{N}(n, \frac{1}{2\rho_2})$, $\tilde{Z} = \tilde{X}/Y$ and $Z = X/Y$, then

$$F_{\tilde{Z}}(z) = \Pr\left(\frac{\tilde{X}}{Y} < z\right) = \Pr(\tilde{X} < Yz) = \int_{-\infty}^{\infty} F_{\tilde{X}}(yz) f_y(y) dy,$$

where $f_y(y)$ is the density function of Y. From (44), $|F_{\tilde{X}}(yx) - F_X(yx)| < \epsilon$. It follows that,

$$\int_{-\infty}^{\infty}(F_X(yx)-\epsilon)f_y(y)dy < \int_{-\infty}^{\infty}F_{\tilde{X}}(yx)f_y(y)dy < \int_{-\infty}^{\infty}(F_X(yx)+\epsilon)f_y(y)dy,$$

which is equivalent to

$$\left|\int_{-\infty}^{\infty}F_{\tilde{X}}(yx)f_y(y)dy - \int_{-\infty}^{\infty}F_X(yx)f_y(y)dy\right| < \epsilon,$$

i.e.,

$$|F_{\tilde{Z}}(z) - F_Z(z)| < \epsilon. \tag{45}$$

Let δ_x and δ_y be the coefficients of variation of X and Y, respectively, then $\delta_x^2 = (\mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1 n^2})/p^2$ and $\delta_y^2 = \frac{1}{2\rho_2 n^2}$. Under the condition $\frac{1}{\rho_1 n} = o(1)$, we have $\delta_x^2 = O(\frac{1}{n})$ since $\mathrm{Var}(\hat{p}) = O(\frac{1}{n})$. Under the condition $\frac{1}{\rho_2 n} = o(1)$, we know $\delta_y^2 = o(\frac{1}{n})$ and then $\delta_y = o(\delta_x)$. When n is sufficiently large, δ_y is sufficiently small. Let $\lambda = \sqrt{\delta_x^2 + 2\delta_y^2}$ and $F_{Z^*}(z)$ be the distribution function of $Z^* \sim \mathcal{N}(p, \mathrm{Var}(\hat{p}) + \frac{1}{2\rho_1 n^2} + \frac{p^2}{2\rho_2 n^2})$. By Lemma A.4, for a normal random variable Y independent of X, with small enough δ_y, the condition (42) is satisfied and we have

$$|F_Z(z) - F_{Z^*}(z)| < \epsilon, \tag{46}$$

for any $z \in I = \left[p - \frac{\sigma_{z^*}}{\lambda}, p + \frac{\sigma_{z^*}}{\lambda}\right]$ where $\sigma_{z^*} = p\sqrt{\delta_x^2 + \delta_y^2}$. Hence, for $z \in I$,

$$|F_{\tilde{Z}}(z) - F_{Z^*}(z)| < |F_{\tilde{Z}}(z) - F_Z(z)| + |F_Z(z) - F_{Z^*}(z)| < 2\epsilon. \tag{47}$$

Note also that as $n \to \infty$, $\frac{\sigma_{z^*}}{\lambda} \to p$, and the limit of I is $(0, 2p)$.

So far, we have shown that as n goes to infinity, under the conditions $\frac{1}{\rho_1 n} = o(1)$ and $\frac{1}{\rho_2 n} = o(1)$, for $z \in I_h$,

$$|F_{\tilde{Z}}(z) - F_{Z^*}(z)| \to 0. \tag{48}$$

□

Lemma A.6. *Let $Z_1, ..., Z_H$ and $Z_1^*, ..., Z_H^*$ be independent continuous random variables which depend on n. Let F denote the distribution function. As $n \to \infty$, if*

$$|F_{Z_h}(z) - F_{Z_h^*}(z)| \to 0$$

holds for any $h = 1, ..., H$ and z in an interval (a_h, b_h) and $\Pr(Z_h \in (a_h, b_h)) \to 1$, $\Pr(Z_h^* \in (a_h, b_h)) \to 1$. Then,

$$\left|F_{\sum_{h=1}^H c_h Z_h}(z) - F_{\sum_{h=1}^H c_h Z_h^*}(z)\right| \to 0$$

for any $z \in \left(\sum_{h=1}^H c_h a_h, \sum_{h=1}^H c_h b_h\right)$, where c_h's are constants.

Proof. It suffices to show that, for any $z \in (a_1c_1 + a_2c_2, b_1c_1 + b_2c_2)$,

$$\left|F_{c_1Z_1+c_2Z_2}(z) - F_{c_1Z_1^*+c_2Z_2^*}(z)\right| \to 0$$

as $n \to \infty$. We have

$$\begin{aligned}
&F_{c_1Z_1+c_2Z_2}(z)\\
&= \Pr(c_1Z_1 + c_2Z_2 < z)\\
&= \Pr\left(Z_1 < \frac{z - c_2Z_2}{c_1}\right)\\
&= \int_{\mathbb{R}} F_{Z_1}\left(\frac{z - c_2x}{c_1}\right) f_{Z_2}(x)dx\\
&= \int_{\mathbb{R}} \left[F_{Z_1}\left(\frac{z - c_2x}{c_1}\right) - F_{Z_1^*}\left(\frac{z - c_2x}{c_1}\right)\right] f_{Z_2}(x)dx\\
&\quad + \int_{\mathbb{R}} F_{Z_1^*}\left(\frac{z - c_2x}{c_1}\right) f_{Z_2}(x)dx\\
&= \int_{\mathbb{R}} \left[F_{Z_1}\left(\frac{z - c_2x}{c_1}\right) - F_{Z_1^*}\left(\frac{z - c_2x}{c_1}\right)\right] f_{Z_2}(x)dx + F_{c_1Z_1^*+c_2Z_2}(z).
\end{aligned} \tag{49}$$

When $a_1 < (z - c_2x)/c_1 < b_1$, we know

$$\left|F_{Z_1}\left(\frac{z - c_2x}{c_1}\right) - F_{Z_1^*}\left(\frac{z - c_2x}{c_1}\right)\right| \to 0. \tag{50}$$

Since $F_{Z_1}(b_1) - F_{Z_1}(a_1) \to 1$ and $F_{Z_1^*}(b_1) - F_{Z_1^*}(a_1) \to 1$, for any $a < a_1$, it holds that $F_{Z_1}(a) \to 0$ and $F_{Z_1^*}(a) \to 0$, and for any $b > b_1$, $F_{Z_1}(b) \to 1$ and $F_{Z_1^*}(b) \to 1$. Thus, (50) also holds when $(z - c_2x)/c_1$ is outside (a_1, b_1). Therefore, the first term of the right-hand side of (49) converges to 0. Then

$$\left|F_{c_1Z_1+c_2Z_2}(z) - F_{c_1Z_1^*+c_2Z_2}(z)\right| \to 0.$$

Similarly, we have

$$\left|F_{c_1Z_1^*+c_2Z_2}(z) - F_{c_1Z_1^*+c_2Z_2^*}(z)\right| \to 0.$$

By the triangle inequality,

$$\left|F_{c_1Z_1+c_2Z_2}(z) - F_{c_1Z_1^*+c_2Z_2^*}(z)\right| \to 0.$$

□

Proof of Theorem 4.5. By Lemma A.5, for each stratum, under the conditions $\rho_1 = \omega(1/n_h)$ and $\rho_2 = \omega(1/n_h)$, the distribution function of $\tilde{p}_h$ converges to that of $\mathcal{N}(p_h, V_h)$ in the interval $(0, 2p_h)$ where

$$V_h = \left(\frac{N_h - n_h}{N_h - 1}\right)\frac{p_h(1 - p_h)}{n_h} + \frac{1}{2\rho_1 n_h^2} + \frac{p_h^2}{2\rho_2 n_h^2}. \tag{51}$$

Let $p^* \sim \mathcal{N}(p, V)$ where $V = \sum_{h=1}^{H} w_h^2 V_h$. By Lemma A.6, in the interval $(0, 2p)$, we have

$$|F_{\tilde{p}}(z) - F_{p^*}(z)| \to 0. \tag{52}$$

where $F_{\tilde{p}}$ denotes the distribution function of $\tilde{p}$ designed in Algorithm 3 and F_{p^*} is the distribution function of p^*.

Let $L = p - z_{1-\alpha/2}\sqrt{V}$, $U = p + z_{1-\alpha/2}\sqrt{V}$, $\tilde{L} = p - z_{1-\alpha/2}\sqrt{\widetilde{V}}$ and $\tilde{U} = p + z_{1-\alpha/2}\sqrt{\widetilde{V}}$. Note that L and U are constants whereas $\tilde{L}$ and $\tilde{U}$ are random variables. Provided that n_h's are sufficiently large, U and L lie in the interval where the following hold due to (52),

$$|F_{\tilde{p}}(U) - F_{p^*}(U)| \to 0 \tag{53}$$

and

$$|F_{\tilde{p}}(L) - F_{p^*}(L)| \to 0. \tag{54}$$

On the other hand, by Theorems 3.1 and 4.4, we know that $\tilde{p}_h \xrightarrow{p} p_h$ and $\frac{1}{\tilde{n}_h} \xrightarrow{p} \frac{1}{n_h}$ under the conditions $\rho_1 = \omega(1/n_h)$ and $\rho_2 = \omega(1/n_h)$. By the continuous mapping theorem, $\widetilde{V}_h \xrightarrow{p} V_h$ as $n_h \to \infty$, and, hence, $\widetilde{V} \xrightarrow{p} V$. Therefore, $\tilde{U} \xrightarrow{p} U$ and $\tilde{L} \xrightarrow{p} L$. Since $F_{\tilde{p}}$ is continuous, we have

$$|F_{\tilde{p}}(\tilde{U}) - F_{\tilde{p}}(U)| \xrightarrow{p} 0 \tag{55}$$

and

$$|F_{\tilde{p}}(\tilde{L}) - F_{\tilde{p}}(L)| \xrightarrow{p} 0. \tag{56}$$

Therefore,

$$\begin{aligned}
&\Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \\
&= \Pr\left(p - z_{1-\alpha/2}\sqrt{\widetilde{V}} < \tilde{p} < p + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right) \\
&= \left(F_{\tilde{p}}(\tilde{U}) - F_{\tilde{p}}(U)\right) + \left(F_{\tilde{p}}(U) - F_{p^*}(U)\right) \\
&\quad - \left(F_{\tilde{p}}(\tilde{L}) - F_{\tilde{p}}(L)\right) - \left(F_{\tilde{p}}(L) - F_{p^*}(L)\right) + \left(F_{p^*}(U) - F_{p^*}(L)\right).
\end{aligned}$$

Putting together (53) through (56) and $F_{p^*}(U) - F_{p^*}(L) = 1 - \alpha$, we have

$$\lim_{n\to\infty} \Pr\left(p \in \left(\tilde{p} - z_{1-\alpha/2}\sqrt{\widetilde{V}}, \tilde{p} + z_{1-\alpha/2}\sqrt{\widetilde{V}}\right)\right) \to 1 - \alpha.$$

Since $\frac{1}{\tilde{n}_h} \xrightarrow{p} \frac{1}{n_h}$, under the conditions $\rho_1 = \omega(1/n_h)$ and $\rho_2 = \omega(1/n_h)$, it holds that $\frac{1}{2\rho_1 \tilde{n}_h^2} = o_P(\frac{1}{n_h})$ and $\frac{\tilde{p}_h^2}{2\rho_2 \tilde{n}_h^2} = o_P(\frac{1}{n_h})$. Therefore, the additional error in estimating the conditional variance of $\tilde{p}$ caused by the injected noise is $O_p\left(\frac{1}{\rho_1 n_h^2} + \frac{1}{\rho_2 n_h^2}\right) = o_p\left(\frac{1}{n_h}\right)$. □

Acknowledgments

We are grateful for helpful conversations with and comments from (in no particular order) Rolando Rodriguez, Brian Finley, Jörg Drechsler, Gary Benedetto, Michael Freiman, and Justin Doty.

Funding

The research presented in this paper was supported by the U.S. Census Bureau Cooperative Agreement CB20ADR0160001.

References

[1] Abowd., J. M. (2016). The challenge of scientific reproducibility and privacy protection for statistical agencies.
[2] Bickel, P. J. and Freedman, D. A. (1984). Asymptotic Normality and the Bootstrap in Stratified Sampling. *The Annals of Statistics* **12** 470 – 482. MR0740906
[3] Brawner, T. and Honaker, J. (2018). Bootstrap Inference and Differential Privacy: Standard Errors for Free. *Unpublished Manuscript*.
[4] Bun, M. and Steinke, T. (2016). Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. *ArXiv arXiv:1605.02065*. MR3591832
[5] Bureau, U. S. C. (2020). Disclosure Avoidance for the 2020 Census: An Introduction.
[6] Cai, T. T., Wang, Y. and Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics* **49** 2825–2850. MR4338894
[7] Covington, C., He, X., Honaker, J. and Kamath, G. (2021). Unbiased Statistical Estimation and Valid Confidence Intervals Under Differential Privacy. *ArXiv* **abs/2110.14465**.
[8] Ding, B., Kulkarni, J. and Yekhanin, S. (2017). Collecting Telemetry Data Privately. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* 3571–3580.
[9] D'Orazio, V., Honaker, J. and King, G. (2015). Differential Privacy for Social Science Inference. *ERN: Other Econometrics: Econometric & Statistical Methods (Topic)*.
[10] Drechsler, J., Globus-Harris, I., Mcmillan, A., Sarathy, J. and Smith, A. (2022). Nonparametric Differentially Private Confidence Intervals for the Median. *Journal of Survey Statistics and Methodology* **10** 804-829. https://doi.org/10.1093/jssam/smac021.
[11] Du, W., Foot, C., Moniot, M., Bray, A. and Groce, A. (2020). Differentially Private Confidence Intervals. *ArXiv arXiv:2001.02285*.
[12] DuMouchel, D. G. J. William H. (1983). Using Sample Survey Weights

in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association* **78** 535–543.

[13] Dwork, C. and Lei, J. (2009). Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing* 371–380. MR2780083

[14] Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* 265–284. Springer. MR2241676

[15] Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* **9** 211–407. MR3254020

[16] Dwork, C. and Rothblum, G. N. (2016). Concentrated Differential Privacy. *CoRR arXiv:1603.01887*.

[17] Dwork, C. and Smith, A. (2009). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality* **1**. MR2472670

[18] Díaz-Francés, E. and Rubio, F. (2013). On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. *Statistical Papers* **54**. https://doi.org/10.1007/s00362-012-0429-2. MR3043290

[19] Erdös, P. and Rényi, A. (1959). On the Central Limit Theorem for Samples From a Finite Population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **4** 49–61. MR0107294

[20] Erlingsson, U., Pihur, V. and Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. CCS'14* 1054–1067. Association for Computing Machinery, New York, NY, USA.

[21] Ferrando, C., Wang, S. and Sheldon, D. (2022). Parametric Bootstrap for Differentially Private Confidence Intervals. In *AISTATS*.

[22] Franco, C., Little, R. J. A., Louis, T. A. and Slud, E. V. (2019). Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys. *Journal of survey statistics and methodology* **7 3** 334-364.

[23] Gaboardi, M., Rogers, R. M. and Sheffet, O. (2019). Locally Private Mean Estimation: Z-test and Tight Confidence Intervals. *ArXiv arXiv:1810.08054*.

[24] Groshen, E. L. and Goroff, D. (2022). Disclosure Avoidance and the 2020 Census: What Do Researchers Need to Know? *Harvard Data Science Review* **Special Issue 2**.

[25] Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** 361–374. MR0125612

[26] Hayya, J., Armstrong, D. and Gressis, N. (1975). A Note on the Ratio of Two Normally Distributed Variables. *Management Science* **21** 1338-1341.

[27] Karwa, V. and Vadhan, S. (2017). Finite sample differentially private

confidence intervals. *CoRR arXiv:1711.03908*. MR3761780
[28] Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *ACM SIGMOD Conference*.
[29] Kroll, M. (2021). On density estimation at a fixed point under local differential privacy. *Electronic Journal of Statistics* **15** 1783 – 1813. https://doi.org/10.1214/21-EJS1830. MR4255316
[30] Kuethe, D., Caprihan, A., Gach, H., Lowe, I. and Fukushima, E. (2000). Imaging obstructed ventilation with NMR using inert fluorinated gases. *Journal of applied physiology (Bethesda, Md.: 1985)* **88** 2279-86.
[31] Lam-Weil, J., Laurent, B. and Loubes, J.-M. (2022). Minimax optimal goodness-of-fit testing for densities and multinomials under a local differential privacy constraint. *Bernoulli* **28** 579 – 600. https://doi.org/10.3150/21-BEJ1358. MR4337717
[32] Lehmann, E. L., ed. (1999). *Elements of Large-Sample Theory*. Springer New York, New York, NY. MR1663158
[33] Li, X. and Ding, P. (2017). General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference. *Journal of the American Statistical Association* **112** 1759-1769. https://doi.org/10.1080/01621459.2017.1295865. MR3750897
[34] Marsaglia, G. (2006). Ratios of Normal Variables. *Journal of Statistical Software* **16** 1–10.
[35] Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review* **61** 317–37.
[36] Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J. and Sobek, M. (2018). IPUMS USA: Version 8.0 Extract of 1940 Census for U.S. Census Bureau Disclosure Avoidance Research [dataset]. *Minneapolis, MN: IPUMS*.
[37] Smith, A. D. (2009). Asymptotically Optimal and Private Statistical Estimation. In *Cryptology and Network Security, 8th International Conference, CANS 2009, Kanazawa, Japan, December 12-14, 2009. Proceedings* (J. A. Garay, A. Miyaji and A. Otsuka, eds.). *Lecture Notes in Computer Science* **5888** 53–57. Springer. https://doi.org/10.1007/978-3-642-10433-6_4.
[38] Team, A. D. P. (2017). Learning with Privacy at Scale.
[39] Thompson, M. E. (1997). *Theory of Sample Surveys. Monographs on Statistics and Applied Probability*. Springer US. MR1462619
[40] van Erven, T. and Harremos, P. (2014). Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory* **60** 3797-3820. https://doi.org/10.1109/TIT.2014.2320500. MR3225930
[41] Wang, Y., Kifer, D. and Lee, J. (2019). Differentially Private Confidence Intervals for Empirical Risk Minimization. *J. Priv. Confidentiality* **9**.
[42] Wasserman, L. and Zhou, S. (2010). A Statistical Framework for Differential Privacy. *Journal of the American Statistical Association* **105** 375-389. MR2656057