Journal of Data Science 20 (2), 253–276
April 2022

DOI: 10.6339/22-JDS1044
Statistical Data Science

Inference for Optimal Differential Privacy Procedures for Frequency Tables

Chengcheng Li[1], Naisyin Wang[1], and Gongjun Xu[1,*]
[1] Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

When releasing data to the public, a vital concern is the risk of exposing personal information of the individuals who have contributed to the data set. Many mechanisms have been proposed to protect individual privacy, though less attention has been dedicated to practically conducting valid inferences on the altered privacy-protected data sets. For frequency tables, the privacy-protection-oriented perturbations often lead to negative cell counts. Releasing such tables can undermine users' confidence in the usefulness of such data sets. This paper focuses on releasing one-way frequency tables. We recommend an optimal mechanism that satisfies ϵ-differential privacy (DP) without suffering from having negative cell counts. The procedure is optimal in the sense that the expected utility is maximized under a given privacy constraint. Valid inference procedures for testing goodness-of-fit are also developed for the DP privacy-protected data. In particular, we propose a de-biased test statistic for the optimal procedure and derive its asymptotic distribution. In addition, we also introduce testing procedures for the commonly used Laplace and Gaussian mechanisms, which provide a good finite sample approximation for the null distributions. Moreover, the decaying rate requirements for the privacy regime are provided for the inference procedures to be valid. We further consider common users' practices such as merging related or neighboring cells or integrating statistical information obtained across different data sources and derive valid testing procedures when these operations occur. Simulation studies show that our inference results hold well even when the sample size is relatively small. Comparisons with the current field standards, including the Laplace, the Gaussian (both with/without post-processing of replacing negative cell counts with zeros), and the Binomial-Beta McClure-Reiter mechanisms, are carried out. In the end, we apply our method to the National Center for Early Development and Learning's (NCEDL) multi-state studies data to demonstrate its practical applicability.

Keywords *goodness-of-fit; hypothesis testing; optimality; table merging*

1 Introduction

When releasing data to the public, a critical concern is the risk of exposing individual information in the data set. Law enforcement, such as the European General Data Protection Regulation, has made de-identification compulsory before releasing data, i.e. removing personal identity from the record. However, even with such measures, data adversaries may still be able to infer about an individual's identity. Some examples on privacy breach include the Netflix prize (Narayanan and Shmatikov, 2008), the Washington State health record identification (Sweeney, 2013), recovery

[*] Corresponding author Email: gongjun@umich.edu.

Received March 28, 2022; Accepted March 30, 2022

of the anonymous location data (Golle and Partridge, 2009) and privacy loss in genomic data (Wang et al., 2009).

In the past two decades, a concept known as data differential privacy (DP) has been developed for the purpose of protecting against the risk of privacy loss. Dwork et al. (2006b) define the first formal definition of DP, which relates the risk for privacy loss to how much the answer to a query would change given the presence or absence of the most extreme person who is prone to privacy breach. Machine-learning types of development quickly adopt the DP-concept. A non-exhaustive list includes streaming (Dwork et al., 2010), data mining (Mohammed et al., 2011), dimension reduction (Chaudhuri et al., 2012), genome-wide association tests (Yu et al., 2014), Bayesian learning (Wang et al., 2015b) and recommender systems (Friedman et al., 2016). In a separate front, efforts have been made to incorporate the DP framework into the traditional disclosure-risk control data-synthesis approaches in which synthetic datasets are generated to represent the original observed data (Rubin, 1993; Little, 1993; Raghunathan et al., 2003; Reiter, 2005; Drechsler, 2011; Raab et al., 2016). In this direction, applications include creating private versions of discrete and categorical data (Charest, 2011; McClure and Reiter, 2012; Abowd and Vilhuber, 2008; Hay et al., 2016; Quick, 2019), continuous data (Wasserman and Zhou, 2010; Snoke et al., 2016), and network data (Karwa et al., 2015; Karwa and Slavković, 2016). Bowen and Liu (2020) provide a comprehensive review of the private data synthesis methods.

The research on performing statistical inference within the framework of DP has gained momentum lately and consists of diverse directions. For the traditional tests applying to normal data, Sheffet (2017) considers the DP hypothesis testing and confidence interval construction for ordinary least squares and ridge estimators in linear regression. Barrientos et al. (2019) propose a differentially private mechanism to release the test statistic and p-value from testing a regression coefficient against 0 from a linear regression model. In anomaly detection, Degue and Le Ny (2018) design a DP generalized likelihood ratio method to decide if data modeled as a sequence of independent and identically distributed (i.i.d.) Gaussian random variables has a given mean value. Ding et al. (2018) study how to conduct hypothesis tests on two population means while preserving privacy under the more restrictive requirement of local differential privacy. Campbell et al. (2018) provide a private analogue of the ANOVA test. Task and Clifton (2016) and Couch et al. (2019) study non-parametric DP rank-based tests. Generalizing a concept given by Wasserman and Zhou (2010), Liu et al. (2019) investigate the relationship between differential privacy framework and hypothesis testing with the goal of using testing analogue to further refine optimal DP regime. Ferrando et al. (2020) consider using parametric bootstrap to construct private confidence intervals and establish a consistency result for the proposed intervals.

Most of the existing DP mechanisms are developed on releasing summary statistics of the data set or responding to queries. On the inference front, the development also mostly focuses on 'perturbing' summary statistics, e.g., test statistics or sufficient statistics in parametric settings. Rogers et al. (2016) investigate the use of adaptive hypothesis testing for p-value corrections and derive valid testing procedures under the challenging (ϵ, δ)-DP scenarios. Avella-Medina (2021) extends this direction by working on the influence-function structure directly for M-estimators' robust DP inferences. When it comes to releasing the data itself, which is essential in the current data-sharing climate, data synthesis methods constructed under the Bayesian framework remain to be the dominating trend.

As for releasing tabular data, the summary cell counts are also observations. For hypothesis testing using the DP-private tabular data, there are many research issues. For example, suppose the cell counts are provided according to different genders and races but the users are only interested at the variable race when the distributions of cell counts do not differ for different genders. For the original data, users can simply combine the categories from different genders and conduct testing on the merged data directly. For the private data on the other hand, to the best of our knowledge, an investigation of testing procedures has not been presented in the literature for this simple and commonly used operation. The same statement applies when one combines the observations collected from different locations or from different years in an analysis.

There have been some works dedicated to developing hypothesis-testing procedures on private tabular data sets. When adding noise to each cell of a contingency table, Johnson and Shmatikov (2013) justify the practice of using classical statistical tests on the private tabular data theoretically by showing that the test statistic computed from a noisy table still asymptotically has the same chi-squared distribution as using the classical method. In an earlier and pioneer work, Vu and Slavkovic (2009) investigate the sample-size determination so that Chi-squared tests using either the private tabular data or the original data can achieve the same power. That is, the additional variation that is considered negligible asymptotically is not always truly negligible – a statement supported by our numerical investigation. Wang et al. (2015a) add Laplace errors to cell counts to create private tabular data and consider the additional variation in the testing procedures. They resort to Monte Carlo methods to ensure the validity of the testing methods. Besides creating their own version of Monte-Carlo based methods, Gaboardi et al. (2016) add Laplace or Gaussian errors to create private tabular data and use the structure to derive the corresponding asymptotic distributions for the test statistics. Focusing on a simple Binomial setting and using the positive count of the original data, Awan and Slavković (2018) extend the use of the Neyman-Pearson lemma to construct the most powerful test under the DP framework. For frequency tables, these existing methods cannot avoid the scenarios of producing negative cell counts.

As the cell counts in a frequency table can be presented as a one-way frequency, we concentrate on this setting and recommend an optimal mechanism that satisfies the standard ϵ-DP. Differing from most of the existing literature, the optimal procedure does not add errors with an explicit form of distribution to the test statistics or cell counts, and only allows realization from the non-negative discrete values as entries of private cell counts. The proposed procedures naturally avoids having negative cell counts without further truncating the private versions of the observations at zero, thus are not subject to the loss of utilities discussed in Rinott et al. (2018). Valid procedures for carrying out goodness-of-fit tests on the private tabular data are developed for the associated procedures. In particular, a de-biased test statistic for the optimal procedure is proposed and its asymptotic distributions are derived. Finite sample approximating distributions for the Chi-square goodness-of-fit test statistics on the commonly used Laplace and Gaussian mechanisms (with/without post-processing of converting negative cells to zero) are also provided to adjust for the additional privacy-related noises injected so that valid inference can be made in settings with relatively small sample sizes. As far as we know, our work is among the first to deal with statistical inference for DP mechanisms with such post-processing. Furthermore, we identify an explicit rate requirement for privacy regimes ϵ under which the inference procedures are valid. Moreover, we derive valid procedures for goodness-of-fit tests on private data after performing some common operations in practice, including inter-table merging, i.e., combine multiple tables, and intra-table merging, i.e. combine interior categories within a table.

We organize the remaining of the article as follows. Section 2 reviews the foundations of DP and some field standard DP mechanisms. Section 3 presents the optimal mechanism. Inference procedures for the goodness-of-fit tests are included in Section 4, where both the inter-table and intra-table merging operations are considered. Section 5 consists of simulation studies to compare the performance of the optimal mechanism with the field standards. In Section 6, we apply our proposed methods to NCEDL's multi-state study data set to demonstrate the utility of our proposed method. Section 7 concludes with some future directions. The proofs for the developed theoretical results and the simulation results for validating the inter- and intra-table inference procedures can be found in the supplementary materials.

2 Review of DP Fundamentals

In this section, we review the basics of DP and some of the most commonly used mechanisms in the literature. Differential privacy (DP) quantifies the degree of privacy protection in terms of privacy budget ϵ. Importantly, DP is a property of the algorithms that produces the privacy-protected data and such algorithms are often created according to a given utility function. Algorithms that satisfy the DP criteria are referred to as differentially private algorithms. Before giving the formal definition of DP, we first introduce some notations. We denote the original observed count data cell and its private version as X and X^* respectively. We use X' to denote the neighbor of X. Here neighboring data means X and X' only differ by one individual. We state the formal definition of ϵ-DP.

Definition 2.1 (ϵ-Differential Privacy)**.** An algorithm $\mathcal{M}$, is ϵ-DP if for all subsets $S \subset \text{Range}(M)$ and for all X, X' such that $d(X, X') = 1$, $P(\mathcal{M}(X) \in S)/P(\mathcal{M}(X') \in S) \leqslant \exp(\epsilon)$.

In the definition above, $\epsilon > 0$ is the privacy budget and $d(X, X') = 1$ means that X and X' differ by one record, making them being the so-called neighbors. One concern about algorithms that satisfy ϵ-DP is that they tend to inject large amount of noise to statistical query results for the reason of attaining a strong privacy guarantee. The practice could result in poor data utility. Several relaxations have been developed. Examples include the (ϵ, δ)-DP (Dwork et al., 2006a) and probabilistic DP (Machanavajjhala et al., 2008). These are considered as relaxations because, while still being 'formal', they offer slightly weaker privacy guarantees. Below we give the formal definition of (ϵ, δ)-DP, which is commonly used in the literature.

Definition 2.2 ((ϵ, δ)-Differential Privacy)**.** An algorithm $\mathcal{M}$, is (ϵ, δ)-DP if for all subsets $S \subset \text{Range}(M)$ and for all X, X' such that $d(X, X') = 1$, $P(\mathcal{M}(X) \in S) \leqslant \exp(\epsilon) P(\mathcal{M}(X') \in S) + \delta$, where $\delta \in [0, 1]$.

Note ϵ-DP is a special case of (ϵ, δ)-DP when $\delta = 0$. The parameter δ adds a small probability when the bound given in Definition 2.1 does not hold. Next we review some DP mechanisms that are considered as field standards, for the purpose of comparison. Since we focus on studying frequency tables, we formalize these algorithms using one-way frequency tables as observed data set $D = (X_1, \ldots, X_K)$. We assume $D \sim \text{Multinomial}(n, P_1, P_2, \ldots, P_K)$ here.

Laplace Mechanism (Lap). The Laplace mechanism can be applied to each of the K cells independently via $X_k^* \sim \text{Lap}(x_k, 1/\epsilon)$ independently for $k = 1, \ldots, K$, where $X_k = x_k$ is the actual observation made. Since we are dealing with count data, we use a discretized version of Laplace distribution with probability mass function, $P(X_k^* = x_k^* \mid X_k = x_k) = (1/C_1) \exp(-\epsilon |x_k - x_k^*|)$, for any integer x_k^*, where $C_1 = \sum_{l \in \mathbb{Z}} \exp\{-\epsilon |x_k - l|\} = 1 + 2\exp(-\epsilon)/\{1 - \exp(\epsilon)\}$.

It is well-known that the above procedure satisfies ϵ-DP. Since frequency tables contain non-negative cell counts only, it is natural to perform post-processing to ensure all the private data entries are non-negative. Here we denote the Lap procedure with the post-processing of converting all negative private entries to zero as the truncated Laplace (TLap) mechanism.

Gaussian Mechanism (GDP). Similar to the Laplace mechanism, the Gaussian mechanism perturbs each of the K cells independently via $X_k^* \sim \mathrm{N}(x_k, \sigma^2 = 2\log(1.25/\delta)/\epsilon^2)$ for $k = 1, \ldots, K$. It has been shown in Dwork and Roth (2014) that this procedure satisfies (ϵ, δ)-DP whenever $0 < \epsilon, \delta < 1$. Again, we use a discretized version here with $P(X_k^* = x_k^* \mid X_k = x_k) = (1/C_2)\exp\{-(x_k - x_k^*)^2/(2\sigma^2)\}$, for any integer x_k^*, where $C_2 = \sum_{m\in\mathbb{Z}} \exp\{-(x_k - m)^2/(2\sigma^2)\}$.

It has been shown in Canonne and Steinke T (2020) that this discrete version has approximately the same privacy guarantee as the continuous Gaussian mechanism. For comparison purpose only, we adopt this discrete version of the Gaussian mechanism. Similarly, we consider the GDP with the post-processing of converting all negative private entries to zero as the truncated GDP (TGDP) mechanism.

Binomial-Beta McClure-Reiter Mechanism (MR). McClure and Reiter (2012) propose an approach to synthesize count data using $X_k^* \mid D = (x_1, \ldots, x_K) \sim \mathrm{Bin}(n, (X_k + \alpha_k)/(n + \alpha_k + \beta_k))$ independently for each cell X_k, where $\alpha_k = \beta_k = 1/\{\exp(\epsilon/n) - 1\}$ makes this procedure satisfy ϵ-DP.

This is among the most commonly used data synthesis method, adapted to satisfy the DP requirement. Its advantage is that it preserves the underlying data structure in that, marginally, each X_k follows a Binomial distribution. However, this procedure completely ruins the cell-wise information due to large α and β values and will in general yield deteriorated utilities.

3 Optimal Mechanism

From the previous section, we know some existing privacy mechanisms have been developed for releasing the frequency table data. However, these mechanisms are not optimal and have other shortcomings. Take the most commonly used Laplace and Gaussian mechanisms as examples; one of the concerns is that negative count data are easily generated, which does not make any practical sense in the frequency table setting. The popular existing methods that overcome this shortcoming often bear a large amount of variation. Taking the MR mechanism as an example, we note that, although the negative count issue is overcome, the damages to the utilities are not always well controlled under the targeted privacy constraints. Furthermore, in real applications, practitioners may face too many choices of mechanisms, often making it difficult for them to pick the "best" one to use. We also note that optimality of DP algorithms in terms of utility maximization have been discussed by several authors. For example, Ghosh et al. (2012) study the optimality of ϵ-differentially private mechanisms under a Bayesian framework. Geng and Viswanath (2015) derive that the optimal ϵ-differentially private mechanism for real-valued query functions takes the staircase-shaped probability densities that are geometrically decaying. While Kairouz et al. (2016) prove the optimality of the randomized aggregatable privacy-preserving ordinal response algorithm and the k-ary randomized response algorithm, under the local differential privacy framework. In this section, we seek to extend the universal optimality idea from Ghosh et al. (2012), in which the mechanism allows a flexible design of loss functions to measure utility, and the corresponding expected utilities are maximized under any given privacy requirements, and we recommend an optimal mechanism for the practitioners when applied to releasing the frequency-table type of data.

Before introducing the optimal mechanism, we first define some notations. Denote the observed data list as $D = (x_1, \ldots, x_K)$ which is generated from Multinomial$(n, P_1, \ldots, P_K)$. The corresponding private data after DP procedures is denoted as $D^* = (x_1^*, \ldots, x_K^*)$. For notation simplicity, we denote $i \in \{0, 1, \ldots, n\}$ as inputs (i.e. the values for x_k). Furthermore, we denote r as the private responses (i.e. the values for x_k^*) where $r \in \{0, 1, \dots, n\}$. Let $p = \{p_{ir} : i = 0, 1, \dots, n, r = 0, 1, \dots, n\} \in \mathbb{R}^{(n+1)\times(n+1)}$ with p_{ir} denoting the probability of mapping an input i to r. Then the optimal p, denoted as $p^* \in \mathbb{R}^{(n+1)\times(n+1)}$, minimizes the expected loss (i.e. maximizes the expected utility), such that

$$p^* = \arg\min_{p} \sum_{i=0}^{n} \sum_{r=0}^{n} p_{ir} L(i, r), \tag{31}$$

where $L(i, r)$ can be any arbitrary loss function, subject only to the constraints that $L(i, r)$ are non-negative, and non-decreasing in $|i - r|$ for each fixed $i = 0, \dots, n$. Note that $p^* = (p_{ir}^*)$ defines a stochastic mechanism that maps an input $i = 0, \dots, n$ to an output $r = 0, \dots, n$. The commonly used loss functions include L_1 and L_2 losses.

The optimal mechanism is detailed in Algorithm 1. Step 1 in Algorithm 1 evaluates a perturbation matrix g corresponding exactly to the discretized Laplace Mechanism, but truncated at 0 and n. The tail probabilities beyond 0 and n are all accumulated as the boundary probabilities. Its optimality has been demonstrated in Geng and Viswanath (2014) and Ghosh et al. (2012). If we fix response r in Step 2, we note that the vector $h_{0r}, \dots, h_{nr}$ can be interpreted as a list of posterior probabilities conditioned on the response r with a uniform prior imposed on the inputs $i = 0, \dots, n$. We will use h to evaluate an optimal remap specific to the loss function L, which is presented in Step 3. Its main goal is to achieve the best balance between the bias and variance so that the expected loss can be minimized. For a general loss function L, step 3 seeks to find an optimal remap index r^*, for each response $r = 0, \dots, n$, such that

$$r^* = \arg \min_{j\in\{0,\dots,n\}} \sum_{i=1}^{n} h_{ir} L(i, j).$$

Note that it is computed as the minimizer of the weighted expected loss. In reality, this optimal remap often brings in some bias into the random error added, but the output variance is significantly reduced which more than compensates for the bias. Then the optimal remapping matrix $y \in \mathbb{R}^{(n+1)\times(n+1)}$ is set to be $y_{rk} = 1$ if $k = r^*$ and $y_{rk} = 0$ if $k \neq r^*$. Finally in step 4, the optimal perturbation matrix p^* can be obtained by combining g and the optimal remap matrix y with $p^* = g \times y$, where × here denotes the matrix multiplication. Lastly, to find the private data cell x_k^*, we can simply sample using $r \in \{0, 1, \dots, n\}$ with probability distribution $\{p_{x_k r}^* : r = 0, \dots, n\}$.

Remark 1. In step 3, when $L(i, r) = |i - r|$ is the L_1 loss, optimal remap index is simply r^*=min$\{k = 0, 1, \dots, n : \sum_{i=0}^{k} h_{ir} \geqslant 0.5\}$. Step 3 in Algorithm 1 simply returns r^* as the ceiling function of the conditional median of $\{0, \dots, n\}$ with probabilities h_{ir} for $i = 0, \dots, n$ in this case. Note that if we take $L(i, j) = (i - j)^2$ as the squared loss, then optimal r^* can be evaluated as the ceiling function of the conditional mean of $\{0, \dots, n\}$ with probabilities h_{ir} for $i = 0, \dots, n$.

Remark 2. Algorithm 1 has time complexity of $O(n^3)$ and space complexity of $O(n^2)$ in general. Note that the time complexity is dominated by Step 3. In the most commonly used L_1 and L_2 losses, short-cuts in Remark 1 can be used, in which cases the time complexity can be reduced to $O(n^2)$.

Following from Theorem 3.1 in Ghosh et al. (2012) by taking a uniform prior on the input $\{i = 0, 1, \dots, n\}$ with probability mass function $P(i) = 1/(1 + n)$, it can be shown that the p^* obtained from the above steps solves the objective function (31) while satisfying the ϵ-DP framework. This is formalized in the proposition below.

```
Algorithm 1: Optimal Mechanism.
  Input: Observed data $D = (x_1, \dots, x_K)$, privacy regime $\epsilon$.
  Output: Optimal perturbation matrix $p^*$, private data $D^*$.
  Set $\alpha = \exp(-\epsilon)$; Initialize $g, h, y \in \mathbb{R}^{(n+1)\times(n+1)}$.
  Step 1. Evaluate $g$ corresponding to the truncated and discretized Laplace Mechanism.
  for $i = 0, 1, 2, \dots, n$ do
      for $r = 0, 1, 2, \dots, n$ do
          if $r = 0$ or $1$ then
              $g_{ir} = \alpha^{|i-r|}/(1+\alpha)$;
          else
              $g_{ir} = \alpha^{|i-r|}(1-\alpha)/(1+\alpha)$.
      end
  end
  Step 2. Evaluate the “posterior” probabilities $h$ as if using uniform prior on $i$.
  for $r = 0, 1, 2, \dots, n$ do
      $s = \left(\sum_{i'=0}^{n} g_{i'r}\right)$;
      for $i = 0, 1, 2, \dots, n$ do
          $h_{ir} = g_{ir}/s$.
      end
  end
  Step 3. Compute the optimal remap matrix $y$.
  for $r = 0, 1, 2, \dots, n$ do
      $r^* = \arg\min_{j\in\{0,\dots,n\}} \sum_{i=1}^{n} h_{ir} L(i, j)$.
      for $k = 0, 1, \dots, n$ do
          if $k = r^*$ then
              $y_{rk} = 1$.
          else
              $y_{rk} = 0$.
      end
  end
  Step 4. Evaluate the optimal perturbation matrix $p^*$.
  for $i = 0, 1, 2, \dots, n$ do
      for $r = 0, 1, 2, \dots, n$ do
          $p^*_{ir} = \sum_{r'=0}^{n} g_{ir'} y_{r'r}$.
      end
  end
  Step 5. Generate private frequency table.
  for $k = 1, 2, \dots, K$ do
      Sample $x^*_k \sim \{0, 1, \dots, n\}$ according to $\{p^*_{x_k r} : r = 0, 1, \dots, n\}$;
      $D^*[k] = x^*_k$.
  end
```

Proposition 1. *The perturbation matrix $p^* \in \mathbb{R}^{(n+1)\times(n+1)}$ obtained through Steps 1 to 4 in Algorithm 1 solves the problem (31) with loss function $L(i, r)$ that is non-negative and non-decreasing in $|i - r|$, satisfying the following constraints: for any $0 < \epsilon < \infty$, (1) $p^*_{ir} - \exp(\epsilon) p^*_{(i+1)r} \geqslant 0$ for $i = 0, \dots, n-1, r = 0, \dots, n$ and (2) $\exp(\epsilon) p^*_{ir} - p^*_{(i+1)r} \leqslant 0$ for $i = 0, \dots, n-1, r = 0, \dots, n$.*

Therefore, the mechanism described in Algorithm 1 satisfies ϵ-DP.

Note that here p^* gives a perturbation matrix that is optimal in that it minimizes the overall expected losses while satisfying the ϵ-DP framework. In the following sections, we will work with the optimal mechanism that minimizes the most commonly used expected L_1 loss and develop inference procedures for it. At the same time, the derivation applies to other losses as well.

4 Goodness-of-Fit Test

In this section, we develop procedures for conducting goodness-of-fit tests on private data. Furthermore, we also consider common operations including both inter- and intra-table mergings.

We assume the true frequency data is $D = (X_1, X_2, \ldots, X_K) \sim \text{Multinomial}(n, P_1, P_2, \ldots, P_K)$. Following a common practice, we release both $D^* = \left(X_1^*, X_2^*, \ldots, X_K^*\right)$, the private tabular data, and the private mechanism used to generate D^*. Suppose we are interested in the goodness-of-fit test $H_0 : P_1 = p_1, P_2 = p_2, \ldots, P_K = p_K$ against $H_1 : P_1 \neq p_1$ or $P_2 \neq p_2$,. . . , or $P_K \neq p_k$. Note that unlike the Gaussian or the Laplace mechanisms that inject a mean zero noise into each tabular cell, the boundary truncation and the optimal remapping step in the optimal mechanism will introduce some biases into the outputs to reach optimality. We propose a de-biased goodness-of-fit test statistic on the private data generated from the optimal procedures described in Section 3. Consider the test statistic T^*_{opt} with

$$T^*_{opt} = \sum_{k=1}^{K} \left(\frac{x_k^* - np_k - b(x_k^*)}{\sqrt{np_k}} \right)^2 = \sum_{k=1}^{K} T_k^{'\ 2},$$

where $b(x_k^*)$ is the bias estimate stemming from the injected noise which can be evaluated using Algorithm 2 below.

Remark 3. Step 1 of Algorithm 2 seeks to find a list of probabilities of input values (denoted as $f_{x_k^*}$) from which the observed private x_k^* is likely to be sampled from. While step 2 computes the list of expected biases if the input values are $0, 1, \ldots, n$. Step 3 computes a weighted average of the expected biases to give the final bias estimate at x_k^*.

In order to take the second moment of the injected noise into account, we give an estimate for the variance, $v(x_k^*)$, to approximate the variance of the injected noise added to x_k using Algorithm 3 below.

```
Algorithm 2: Evaluation of Bias.
  Input: Private data $D^*$ and optimal perturbation matrix $p^*$.
  Output: Bias terms $b(x_k^*)$ for $k = 1, \ldots, K$.
  for $k = 1, \ldots, K$ do
      1. $f_{x_k^*} = (p^*_{0x_k^*}, p^*_{1x_k^*}, \ldots, p^*_{nx_k^*})^T / \left( \sum_{i=0}^n p^*_{ix_k^*} \right)$;
      2. Evaluate $b = (b_0, b_1, \ldots, b_n)$:
      for $i = 0, 1, \ldots, n$ do
          $b_i = \sum_{j=0}^n p^*_{ij}(j - i)$;
      end
      3. $b(x_k^*) = \sum_{i=0}^n f_{ix_k^*} b_i$.
  end
```

```
Algorithm 3: Evaluation of Variance.
  Input: Private data $D^*$ and optimal perturbation matrix $p^*$.
  Output: Variance terms $v(x_k^*)$ for $k = 1, \dots, K$.
  for $k = 1, \dots, K$ do
      1. $f_{x_k^*} = (p^*_{0x_k^*}, p^*_{1x_k^*}, \dots, p^*_{nx_k^*})^T / \Big(\sum_{i=0}^n p^*_{ix_k^*}\Big)$;
      2. Evaluate $v = (v_0, v_1, \dots, v_n)$:
      for $i = 0, 1, \dots, n$ do
          $v_i = \sum_{j=0}^n p^*_{ij}(j - \sum_{j=0}^n jp^*_{ij})^2$;
      end
      3. $v(x_k^*) = \sum_{i=1}^n f_{ix_k^*} v_i$.
  end
```

Remark 4. Step 1 of Algorithm 3 is exactly the same as in Algorithm 2. Step 2 of Algorithm 3 computes a list of expected variances given the possible original observations of $0, 1, \dots, n$ (we denote it as $v = (v_0, v_1, \dots, v_n)$). Step 3 computes a weighted average of the expected variances to give the final estimate for the variance term at x_k^*.

We characterize the asymptotic null distribution of T^*_{opt} in the following theorem.

Theorem 1. *Assume the private data are generated from the optimal procedure with privacy regime ϵ_n satisfying $\epsilon_n^{-1} n^{-1/2} \to 0$ as $n \to \infty$. Then under the null hypothesis $H_0 : P_1 = p_1, \dots, P_K = p_K$, for some $1 < K < \infty$, $T^*_{opt} \to \sum_{k=1}^K \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K \times K}$ where $\Sigma_{kk} = 1 - p_k + v(x_k^*)/(np_k)$ for $k = 1, \dots, K$ and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K$.*

Remark 5. We can decompose $x_k^* = x_k + err_k$, and Theorem 1 takes the second moment of err_k into account so that the asymptotic null distribution can have better finite sample properties when the sample size n is small and the privacy-protection requirement is high (small ϵ_n). Furthermore, we state that the rate of decrease of privacy regime ϵ_n cannot be faster than $n^{-1/2}$ for the asymptotics to work. For much perturbed outputs with small n and ϵ_n, inference procedures have low powers in general. Under such scenarios, as shown in the numerical outcomes, our proposed optimal procedure outperforms others. When an even smaller ϵ_n is required so that the asymptotic fails, one perhaps should carefully consider whether it is meaningful to release such a deteriorated data set.

Theorem 1 can be generalized easily to any DP mechanisms whose injected noises are additive to the true cell counts. Below we take the most commonly used Laplace and Gaussian mechanisms (and their corresponding post-processing versions, TLap and TGDP) as examples and derive their asymptotic distributions. We use a standard Pearson Chi-square test statistic in the literature which is given as follows,

$$T^* = \sum_{k=1}^K \frac{(x_k^* - np_k)^2}{np_k} = \sum_{k=1}^K \Big(\frac{x_k^* - np_k}{\sqrt{np_k}}\Big)^2 = \sum_{k=1}^K T_k^2.$$

Theorem 2. *Assume the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$. Under the null hypothesis $H_0 : P_1 = p_1, \dots, P_K = p_K$, for some $1 < K < \infty$, the following results hold.*

(a). When the private data are generated from the ϵ_n-DP Laplace mechanism or the ϵ_n-DP truncated Laplace mechanism (at zero). $T^* \to \sum_{k=1}^K \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2

random variables and Λ_k are eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K\times K}$ where $\Sigma_{kk} = 1 - p_k + 2/(np_k\epsilon_n^2)$ for $k = 1, \dots, K$ and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K$.

(b). When the private data are generated from the (ϵ_n, δ)-DP Gaussian mechanism or the truncated (ϵ_n, δ)-DP Gaussian mechanism (at zero) for some $0 < \delta < 1$. $T^* \to \sum_{k=1}^K \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K\times K}$ where $\Sigma_{kk} = 1 - p_k + (2\log(1.25/\delta) - 1)/(np_k\epsilon_n^2)$ for $k = 1, \dots, K$ and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K$.

4.1 Merging Multiple Frequency Tables

The data users may often encounter the need to merge different private tabular data sets. For example, the users may want to merge multiple data sets across different time-periods or regions before performing statistical analysis. Merging multiple frequency lists can increase sample size and therefore improve confidence when performing statistical inference. In this section, we develop inference procedures that can be applied to the merged private frequency tables.

Suppose the users are interested in merging C data lists $j = 1, \dots, C$ together, with the j'th private data list denoted as $D_j^* = \{X_{j1}^*, X_{j2}^*, \dots, X_{jK}^*\}$ with sample size n_j. Further assume $n = \sum_{j=1}^C n_j$. The merged data set can then be denoted as $D_m^* = \{\sum_{j=1}^C X_{j1}^*, \dots, \sum_{j=1}^C X_{jK}^*\}$. Furthermore, suppose the user knows the DP procedure used to create each of the private data lists D_j^*. To test $H_0 : P_1 = p_1, P_2 = p_2, \dots, P_K = p_K$ against H_1: H_0 does not hold on the merged data, We consider the following test statistic

$$T_M^* = \sum_{k=1}^K \Big(\frac{\sum_{j=1}^C X_{jk}^* - np_k - b_M(\{x_{jk}^*\}_{j=1}^C)}{\sqrt{np_k}} \Big)^2 = \sum_{k=1}^K T_{mk}^2,$$

where $b_M(\{x_{jk}^*\}_{j=1}^C) = \sum_{j=1}^C b(x_{jk}^*)$. Theorem 3 characterizes the asymptotics of the T_M^* under the null hypothesis. We give the results for both the recommended optimal procedure and the commonly used mechanisms in the literature.

Theorem 3. *Assume $\epsilon_n^{-1} n^{-1/2} \to 0$ as $n \to \infty$. Under the null hypothesis $H_0 : P_1 = p_1, \dots, P_K = p_K$, for some $1 < K, C < \infty$, the following results hold.*

(a) If D_j^* are obtained from the optimal procedure with privacy regime ϵ_n, then $T_M^* \to \sum_{k=1}^K \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K\times K}$ where $\Sigma_{kk} = 1 - p_k + v_M(\{x_{jk}\}_{j=1}^C)/(np_k)$ for $k = 1, \dots, K$ with $v_M(\{x_{jk}\}_{j=1}^C) = \sum_{j=1}^C v(x_{jk}^*)$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K$.

(b) If D_j^* are obtained from the ϵ_n-DP Laplace mechanism or the truncated ϵ_n-DP Laplace mechanism (at zero), then we set $b_M(\{x_{jk}^*\}_{j=1}^C) = 0$ in T_M^*. We have $T_M^* \to \sum_{k=1}^K \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K\times K}$ where $\Sigma_{kk} = 1 - p_k + 2C/(\epsilon_n^2 np_k)$ for $k = 1, \dots, K$ and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K$.

(c) If D_j^* are obtained from the (ϵ_n, δ)-Gaussian mechanism or the truncated (ϵ_n, δ)-Gaussian mechanism (at zero) for some $0 < \delta < 1$, then we set $b_M(\{x_{jk}^*\}_{j=1}^C) = 0$ in T_M^*. We have $T_M^* \to \sum_{k=1}^K \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K\times K}$ where $\Sigma_{kk} = 1 - p_k + Cv_M/np_k$ for $k = 1, \dots, K$ with $v_M = (2\log(1.25/\delta) - 1)/\epsilon_n^2$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K$.

4.2 Merging Cells Within a Frequency Table

Data users may also be interested in combining entries within a frequency table, either because they are interested in a more general group of classes or because the sample sizes of some cells are too small to carry out valid analysis. Similar to inter-table merging, intra-table merging directly on the private tabular data accumulates random noises in the merged cells, resulting in invalid analysis results if these noises are not taken into account separately. In this section, we provide goodness-of-fit test procedures that can be applied to the intra-table merged private data sets.

Without loss of generality, suppose the users are interested in merging the first M cells of the private list $D^* = \{X_1^*, X_2^*, \dots, X_K^*\}$ for some $M < K$. Denote the resulting merged data set as $D_m^* = \left\{ \sum_{k=1}^M X_k^*, X_{M+1}^*, \dots, X_K^* \right\} = \{X_{m1}^*, X_{m2}^*, \dots, X_{m(K-M+1)}^*\}$. To test $H_0 : P_{m1} = p_1, P_{m2} = p_2, \dots, P_{m(K-M+1)} = p_{K-M+1}$ against H_1: H_0 does not hold on the merged data set D_m^*. We consider

$$T_M^* = \sum_{k=1}^{K-M+1} \left(\frac{X_{mk}^* - np_k - b_M(x_{mk}^*)}{\sqrt{np_k}} \right)^2 = \sum_{k=1}^{K-M+1} T_{mk}^2,$$

where $b_M(x_{m1}^*) = \sum_{i=1}^M b(x_i^*)$ and $b_M(x_{mk}^*) = b(x_{M+k-1}^*)$ for $k = 2, \dots, K-M+1$. The following theorem characterizes the asymptotic null distribution of T_M^*. Again, we give the results for both the recommended optimal procedure and the commonly used mechanisms in the literature.

Theorem 4. *Assume the privacy regime ϵ_n satisfies $\epsilon_n^{-1} n^{-1/2} \to 0$ as $n \to \infty$. Under the null hypothesis $P_{m1} = p_1, P_{m2} = p_2, \dots, P_{m(K-M+1)} = p_{K-M+1}$, for some $1 < K < \infty$ and $1 \leqslant M < K$, the following results hold.*

(a) If D^* are from the ϵ_n-DP optimal procedure, then $T_M^* \to \sum_{k=1}^{K-M+1} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the matrix $\Sigma \in \mathbb{R}^{(K-M+1)\times(K-M+1)}$ where $\Sigma_{kk} = 1 - p_k + v_M(x_{mk}^*)/(np_k)$ for $k = 1, \dots, K-M+1$, with $v_M(x_{m1}^*) = \sum_{i=1}^M v(x_i^*)$ and $v_M(x_{mk}^*) = v(x_{M+k-1}^*)$ for $k = 2, \dots, K-M+1$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K-M+1$.

(b) If D^* are from the ϵ_n-DP Laplace mechanism or the truncated ϵ_n-DP Laplace mechanism (at zero), we set $b_M(x_{mk}^*) = 0$ in T_M^*. then $T_M^* \to \sum_{k=1}^{K-M+1} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{(K-M+1)\times(K-M+1)}$ where $\Sigma_{11} = 1 - p_1 + Mv_M/np_k$ for $k = 1, \dots, K$ and $\Sigma_{kk} = 1 - p_k + v_M/np_k$ for $k = 2, \dots, K-M+1$, with $v_M = 2/\epsilon_n^2$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K-M+1$.

(c) If D^* are from the (ϵ_n, δ)-DP Gaussian mechanism or the truncated (ϵ_n, δ)-DP Gaussian mechanism (at zero), we set $b_M(x_{mk}^*) = 0$ in T_M^*. Then we have $T_M^* \to \sum_{k=1}^{K-M+1} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{(K-M+1)\times(K-M+1)}$ where $\Sigma_{11} = 1-p_1+Mv_M/np_k$, $\Sigma_{kk} = 1-p_k+v_M/np_k$ for $k = 2, \dots, K-M+1$ with $v_M = (2\log(1.25/\delta) - 1)/\epsilon_n^2$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leqslant k \neq j \leqslant K-M+1$.

5 Simulation Studies

Various simulation studies are designed to examine and compare the effectiveness of the recommended Opt procedure in Section 3 with the five methods, Lap, TLap, GDP, TGDP and MR reviewed in Section 2; the latter five are commonly used algorithms in the literature. Throughout the section, we consider three targeted privacy regimes of $\epsilon = 0.25, 0.5, 0.75$. L_1 loss is used to compare the performance of different procedures. We also numerically validate the inference procedures given in Section 4.

5.1 Utility

Here, sample size $n = 500$ and observed counts of $i = 5, 200, 450$ are considered. 5000 Monte Carlo samples are generated for each setting and L_1 losses are evaluated. For the MR mechanism, $\alpha = \beta = 1999.5, 999.5, 666.2$ are required to achieve $\epsilon = 0.25, 0.5, 0.75$ respectively. For the GDP/TGDP, since the ϵ-DP does not exist, we relax it to (ϵ, δ)-DP instead. It is a common practice to take $\delta \leqslant 1/n$ and we set $\delta = 1/500 = 0.002$ in this case. The distributions of the losses across different settings are summarized and compared using box-plots in Figure 1. From Figure 1, we note that the Opt mechanism improves utilities significantly in comparison to the traditional data synthesis MR mechanism. We also observe that the MR method's utility varies with the observed values of i (5, 200, 450). The closer the observed i-value is to $n/2 = 250$, the better the MR method's utility. Regardless of the values of i, MR mechanism's performance is much worse than the other procedures. GDP/TGDP, on the other hand, though requiring further relaxation on privacy regime with an additional δ of 0.002, its utility is still worse than that of the Lap/TLap and the Opt across all scenarios. The truncated mechanisms have similar performances as their counterparts, though they tend to give results with lower variations as indicated by slightly smaller box sizes.

Overall, the Opt mechanism appears to achieve comparable utilities to the Lap/TLap mechanisms. To examine closely, we present the Monte Carlo means and variances of the L_1 losses in Table 1. Compared to the Lap/TLap mechanisms, we observe that the Opt mechanism achieves smaller Monte Carlo means and similar variances of the L_1 losses under the same privacy constraints. The major improvement of Opt is achieved under the most restrictive privacy regime, $\epsilon = .25$ and for the observed $i = 5$, most distant from the center $n/2$ and therefore more prone to privacy risks. Moreover, we also point out that the truncated mechanisms, including the TLap and the TGDP, are expected to perform no worse than their non-truncated counterparts because converting negative cells to zero will produce private data closer to their true underlying values which are always non-negative. Indeed, when the true value is close to zero at $i = 5$ and when the privacy regime is high at $\epsilon = 0.25$ (so the injected noises are large and more negative cells would be converted to zero), we observe significant improvements on the mean and variance of L_1 losses for the truncated versions over their non-truncated counterparts. While when the underlying value is large and privacy regime is small, the improvement is not so obvious. Some counter-intuitive observations are due to Monte Carlo errors. In the last part of Table 1 under 'Negative Proportion', we report the proportion of Monte Carlo samples with at least one negative count. In reality, a negative cell count will not be observed. Releasing a private table with a negative cell count will likely reduce the users' confidence in the quality of the released tables. We notice that when $\epsilon = 0.25$, the proportion of negative counts yielded from the Lap and GDP versions of private tables are 12.3% and 36.8% respectively. In contrast, the Opt mechanism ensures that no negative count will be generated. This characteristic demonstrates the benefits of releasing Opt versions of private frequency tables as compared to the traditional Lap or GDP mechanism without any post-processing procedures.

5.2 Goodness-of-fit Test

In this sub-section, we seek to numerically validate the inference procedures in Section 4. Considering the frequency data $D \sim \text{Multinomial}(n, P_1 = 0.1, P_2 = 0.1, P_3 = 0.8)$, we are interested in testing $H_0 : P_1 = 0.1, P_2 = 0.1, P_3 = 0.8$. We consider different sample sizes $n = 100, 1000$, under three privacy targets $\epsilon = 0.25, 0.5, 0.75$. Setting the significance level to be 0.05, we evaluate

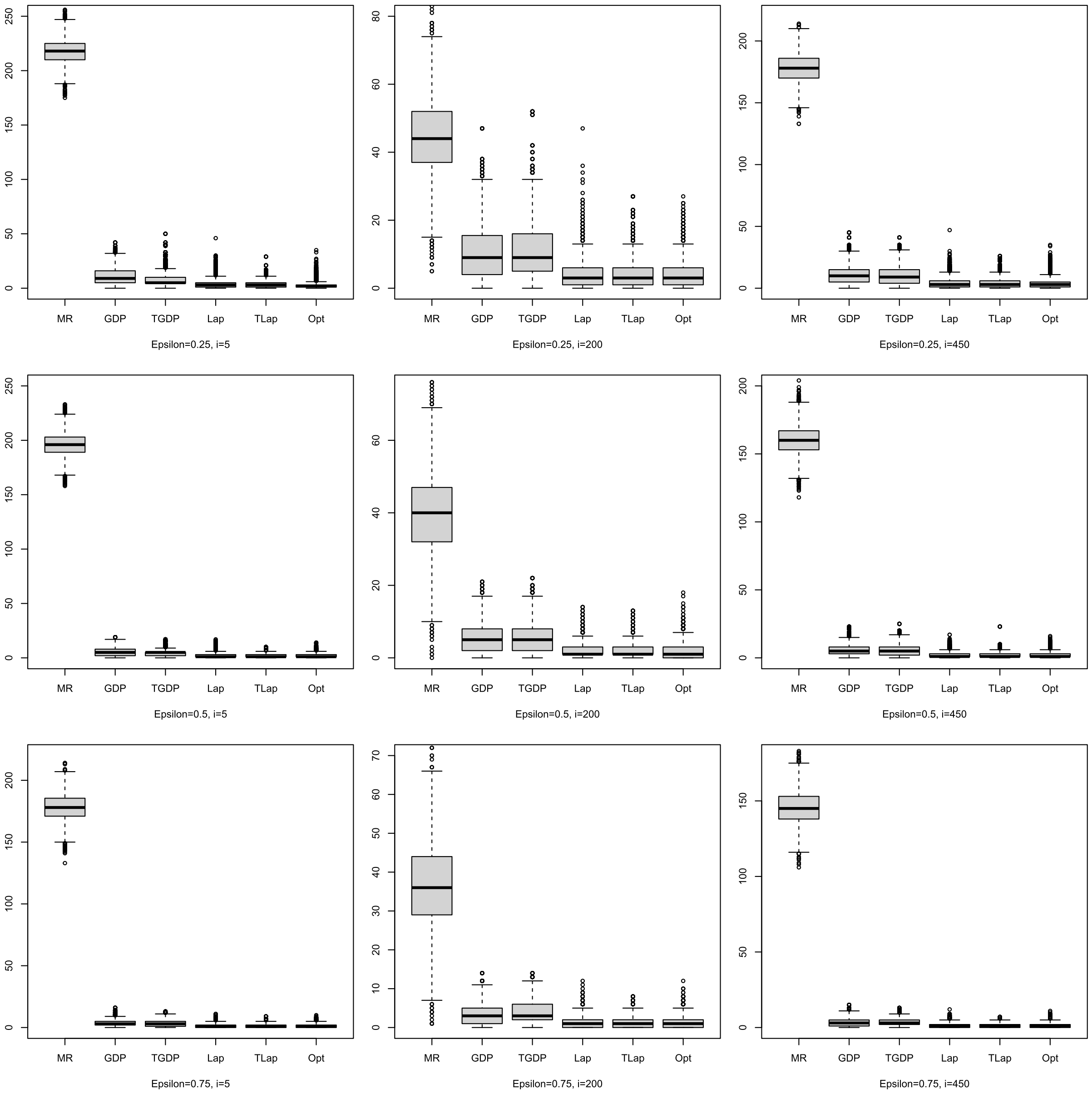

Figure 1: Utility comparison (L_1 loss) amongst the Opt, Lap, TLap (truncated Laplace), GDP, TGDP (truncated GDP) and MR mechanisms across different privacy regimes $\epsilon = 0.25, 0.5, 0.75$ and observed data counts $i = 5, 200, 450$.

500 empirical test statistics. First, we examine the use of the traditional Chi-square distribution with $K - 1$ degrees of freedom as if the data is not perturbed. We check whether the empirical type I errors could be controlled using this naive asymptotic null distribution for the private data sets produced by the five mechanisms, Opt, Lap, TLap, GDP and TGDP. The resulting average empirical type I errors are provided under the "Naive Method" scenario in Table 2. Except for the setups when the sample size is large at $n = 1000$ and the privacy control is not strict with $\epsilon > 0.5$, the average empirical type I error rates are way above the targeted value of

Table 1: Mean, variance of the L_1 losses and proportion of negative counts out of 5000 Monte Carlo samples for different privacy regimes, $\epsilon = 0.25, 0.5, 0.75$, and observed counts, $i = 5, 200, 450$.

	$\epsilon = 0.25$			$\epsilon = 0.5$			$\epsilon = 0.75$		
	i=5	i=200	i=450	i=5	i=200	i=450	i=5	i=200	i=450
					Mean				
Opt	2.94	3.98	3.93	1.79	1.90	1.88	1.23	1.20	1.24
Lap	3.97	4.04	4.08	2.03	1.96	1.96	1.31	1.27	1.33
TLap	3.44	3.99	4.15	1.84	2.04	1.94	1.31	1.31	1.34
GDP	11.08	11.04	10.98	5.41	5.61	5.76	3.75	3.42	3.63
TGDP	7.82	11.19	10.40	4.47	5.64	5.67	3.49	3.74	3.67
MR	217.55	44.20	177.89	195.79	39.80	160.05	178.20	36.45	145.29
					Variance				
Opt	10.03	15.53	15.60	3.06	4.11	3.94	1.84	1.86	1.96
Lap	16.68	16.67	16.60	4.58	3.97	4.17	1.96	1.83	1.95
TLap	9.69	17.29	17.26	2.98	4.04	4.35	1.66	1.80	1.87
GDP	68.61	71.93	69.33	15.87	18.90	17.84	8.40	7.73	7.56
TGDP	52.56	72.54	63.63	9.41	16.87	18.64	6.03	7.38	7.33
MR	121.85	126.44	125.23	118.62	129.83	123.29	115.95	122.04	118.34
					Negative Proportion				
Opt		0.000			0.000			0.000	
Lap		0.123			0.037			0.009	
TLap		0.000			0.000			0.000	
GDP		0.368			0.210			0.104	
TGDP		0.000			0.000			0.000	
MR		0.000			0.000			0.000	

0.05 for all the mechanisms. That is, the use of the naive χ^2_{K-1} null distribution cannot control the type I error in such cases. The situation is worse when the GDP or TGDP mechanisms are used. When sample sizes are small at $n = 100$ or when the privacy regime is strict with $\epsilon = 0.25$, the naive method performs poorly and valid inference is impossible.

Next, we test the results in Theorems 1 and 2 under the same simulation settings as above. The goal is to check whether the new test statistics and null distributions derived can control the type I error rate well so that valid hypothesis testing can be implemented. Under the "Proposed Method" scenario of Table 2, we report the mean empirical type I errors obtained from the 500 simulated samples. In contrast to the "Naive Method" scenario, the empirical type I errors are controlled fairly well at around 5%. We point out that truncation tends to reduce the type I error rate, especially when sample size is small and privacy regime is high.

To compare the statistical powers, we consider the alternative hypotheses $H_1 : P_1 = p_1, P_2 = p_1, P_3 = 1 - 2p_1$, where we explore the cases with $p_1 = 0.1, 0.11, 0.12, \dots, 0.25$ for $n = 100$ and $p_1 = 0.1000, 0.1025, 0.1050, \dots, 0.1350$ for $n = 1000$. The results are summarized in Figure 2. In Figure 2, as the H_1 hypothesized values depart further from that in the H_0, all mechanisms have the powers rising to 1. The Lap/TLap mechanisms and the Opt procedures have significantly higher power than the GDP/TGDP mechanisms, especially when the sample size and the ϵ is small ($n = 100$ or $\epsilon = 0.25$). The Lap/TLap and the Opt procedures perform similarly with the Opt outperforming Lap/TLap slightly when the sample size is small. As n and ϵ get larger, the differences in power amongst the mechanisms diminish. This is reasonable because the random noise injected is inversely proportional to ϵ and is scaled by $\sqrt{n}$ in the test statistic. So, when the sample size increases, the noises become increasingly more negligible. That makes the statistical

Table 2: Mean empirical type I errors out of 500 simulated samples under two study scenarios across different sample sizes $n = 100, 1000$ and different privacy regimes $\epsilon = 0.25, 0.5, 0.75$. The two scenarios are "Naive Method": Chi-square tests for original data, "Proposed Method": procedure proposed in Section 4.

		$n = 100$			$n = 1000$		
		$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
Naive Method	Opt	0.392	0.180	0.108	0.102	0.063	0.054
	Lap	0.449	0.188	0.112	0.102	0.064	0.056
	TLap	0.448	0.187	0.113	0.101	0.062	0.055
	GDP	0.865	0.549	0.327	0.401	0.141	0.088
	TGDP	0.865	0.549	0.327	0.401	0.141	0.088
Proposed Method	Opt	0.037	0.047	0.054	0.052	0.050	0.052
	Lap	0.066	0.058	0.055	0.055	0.051	0.051
	TLap	0.034	0.053	0.055	0.055	0.051	0.051
	GDP	0.052	0.051	0.051	0.051	0.052	0.051
	TGDP	0.023	0.028	0.038	0.051	0.052	0.051

powers for the mechanisms converge to the same performance level when the sample size becomes large. We also note that post-processing of truncating at zero does not have much impact on the statistical power.

Furthermore, we compare the power performance of the non-debiased and the de-biased test statistics for the Opt under the same setting as above. The results are summarized in Figure 3. We observe that the bias correction step can help improve statistical power slightly when ϵ and sample size are small. On the other hand, the simpler version without the bias-correction component is as competitive in all settings.

Moreover, we also conduct simulation studies to validate the results given in Theorems 3 and 4 for inter- and intra-table merging. Simulation results show that empirical type I errors can be controlled well using the approximate null distribution developed, suggesting Theorems 3 and 4 provide valid inference procedures. Our simulation results also show that the recommended Opt mechanism perform the best in terms of the empirical powers. Due to space limit, we include these additional simulation results in the supplementary materials.

6 Analysis of Children's Early Development and Learning Data

In this section, we consider an application to the data from the NCEDL's multi-state study (M. Clifford et al., 2017). The data set consists of 2982 records of 308 variables collected from pre-kindergarten children in 11 states of the United States of America. We focus on one-way frequency tables and select two categorical variables *household type* and *family income* to investigate the differences between the three east coast states, Massachusetts, New York and New Jersey, and the other states. We start with the variable, *household type.* This is a variable with five categories describing different household types: (I) single mom or dad, (II) mom and dad both in home, (III) w/o dad, (IV) multiple adults, but parents/step-parents not both in home, and (V) single adult, not mom or dad. After removing the instances with missing outcomes and

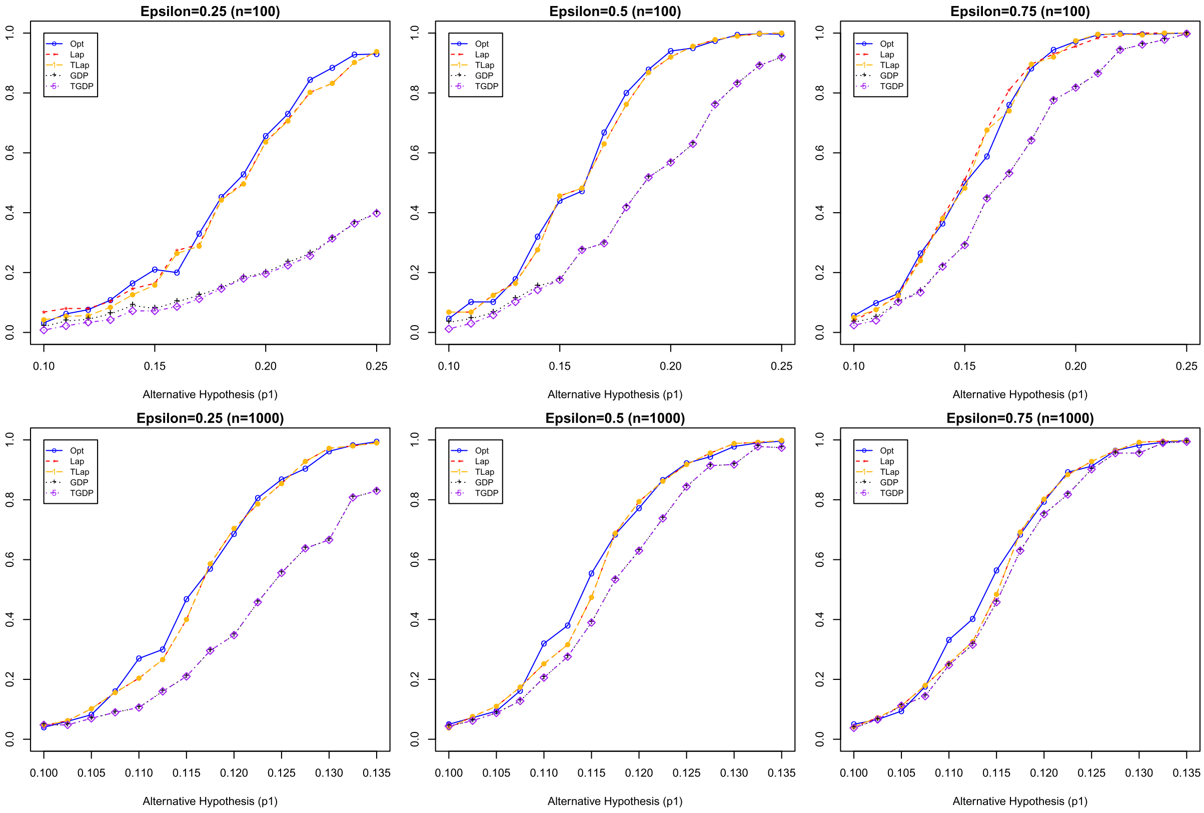

Figure 2: Empirical power comparison for five privacy procedures: Opt, Lap, TLap, GDP and TGDP.

Table 3: One-way frequency table of the variable, *household type,* in Massachusetts, New York, New Jersey and the other 8 states (denoted as Others) in the NCEDL study.

	(I)	(II)	(III)	(IV)	(V)	n
Massachusetts	85	237	9	36	5	372
New York	48	83	4	24	3	162
New Jersey	66	174	18	51	4	313
Others	403	1241	143	251	21	2059

collapsing into frequency tables, the values, including those from the three east coast states and the remaining eight states, are summarized in Table 3.

State by state for the three east coast states, we apply the three mechanisms considered in Section 5 with $\epsilon = 0.25, 0.5, 0.75$, and $\delta = 1/n$ (only for GDP and TGDP) to these frequency tables. In particular, the Opt mechanism used here optimizes against L_1 losses and is implemented according to Algorithm 1. For each case, 500 private samples are simulated. Using the New York state as an example, we report the summarized utilities for the five DP-mechanisms. In Table 4, we present the mean values of the private entries under each of the five categories, the average entry-wise Monte Carlo standard errors (Ave. SD), the average entry-wise L_1 loss (Mean L_1 loss), and the proportions of private datasets generated with at least one negative entry out of

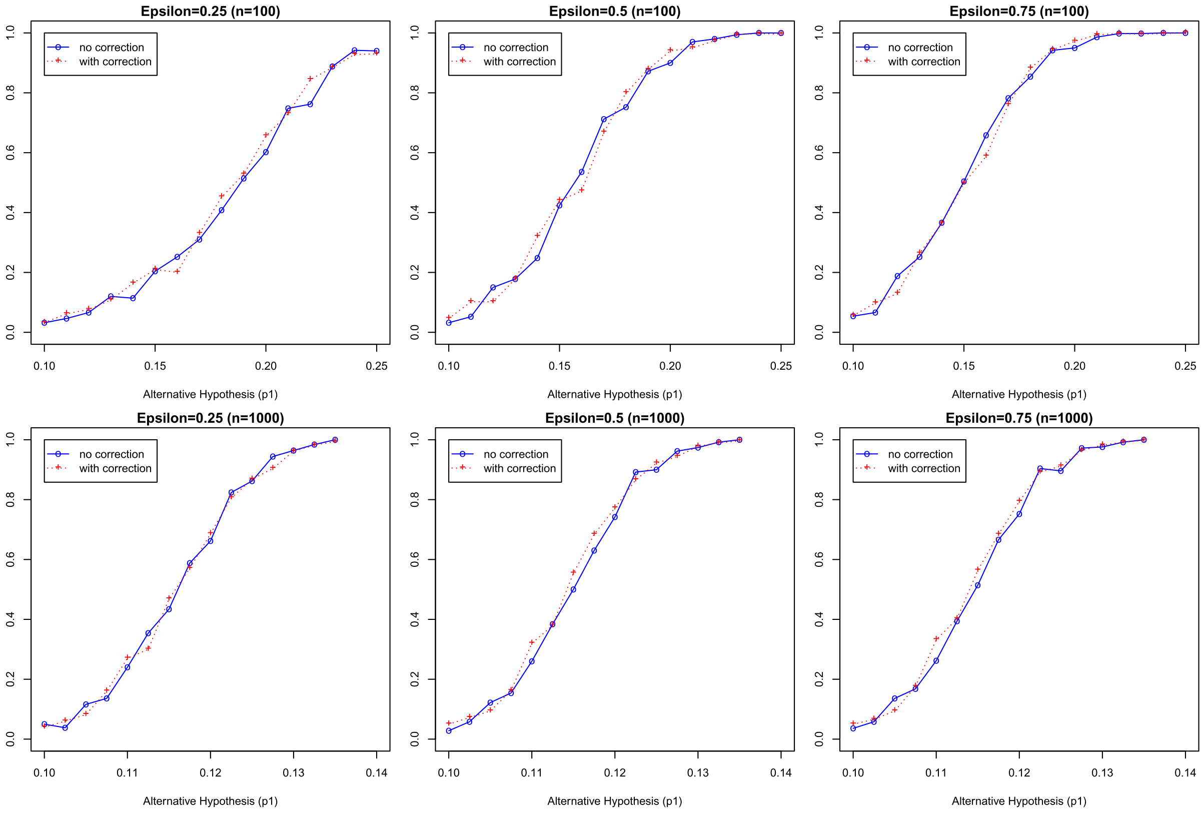

Figure 3: Empirical power comparisons for the Opt procedures with and without bias-correction.

500 simulated samples (‘<0’ Proportion). We note that the Opt procedure will always produce private data with the smallest mean L_1 losses as compared to the other methods, in line with the theoretical results in Proposition 1. The results also suggest that the improvement in the utility is a result of the reduced uncertainty as illustrated by the smallest average Monte Carlo SD’s achieved by the Opt mechanism. Furthermore, we observe that categories (III) and (V) have close to 0 entries; thus their corresponding private versions for Lap and GDP procedures might be negative, especially in high privacy regimes where noise injected is large. Indeed, when $\epsilon = 0.25$, the proportion of private frequency tables produced by the Lap and GDP mechanisms with negative entries are quite high. In reality, releasing such frequency tables could cause confusion and doubts amongst users about the usefulness of the data sets.

For the variable *household type* and using the entries from the other eight states, we obtain the proportions under the five categories, $p_0 = (p_{01}, \dots, p_{05})^T$, where $p_{01} = 0.196,\ p_{02} = 0.603,\ p_{03} = 0.069,\ p_{04} = 0.122,\ p_{05} = 0.010$ and use them as the null-hypothesized values. Let $P_{MA,\ell}$, $P_{NY,\ell}$, and $P_{NJ,\ell}$, $\ell = 1, 2, \dots, 5$, denote the population proportions of the five *household type* categories for the states of Massachusetts, New York and New Jersey respectively. Here, we are interested in testing whether the distribution(s) of the variable *household type* for the three east coast states equal the null hypothesized value p_0. Specifically, we have the null hypothesis $H_0 : P_{MA,\ell} = p_{0\ell}, P_{NY,\ell} = p_{0\ell}, P_{NJ,\ell} = p_{0\ell}$, for $\ell = 1, \dots, 5$. The goodness-of-fit testing procedures proposed in Section 4 are directly applicable here for individual state and for the complex H_0 for all three states as stated above. We apply the test statistics constructed using results from

Table 4: Properties of private one-way frequency tables of the variable *household type* in New York state. The summary statistics reported are: mean values of the private entries under each of the five categories (I) to (V), average Monte Carlo standard deviations of all the private data entries (Ave. SD), mean L_1 loss with respect to true values in Table 3 (Ave. L_1 loss) and the proportion of 500 private tables with at least one negative entry (‘<0’ Proportion).

		(I)	(II)	(III)	(IV)	(V)	Ave. SD	Ave. L_1 loss	‘<0’ Proportion
$\epsilon = 0.25$	OPT	48.00	82.70	5.37	24.12	4.89	4.87	3.44	0.00
	Lap	47.59	82.99	3.76	23.88	3.07	5.74	4.01	0.35
	TLap	47.87	83.10	4.76	24.13	3.96	5.23	3.70	0.00
	GDP	48.29	82.45	3.96	23.64	3.12	12.87	10.26	0.63
	TGDP	48.29	82.45	7.16	23.84	6.79	11.20	8.84	0.00
$\epsilon = 0.5$	OPT	47.79	82.92	4.25	23.77	3.32	2.59	1.79	0.00
	Lap	47.79	82.98	3.87	23.95	3.01	2.88	1.99	0.12
	TLap	47.79	82.98	4.04	23.95	3.21	2.72	1.92	0.00
	GDP	48.14	82.69	4.01	23.84	3.05	6.45	5.12	0.45
	TGDP	48.14	82.69	4.95	23.84	4.35	5.86	4.67	0.00
$\epsilon = 0.75$	OPT	48.02	83.04	3.94	23.91	2.98	1.80	1.20	0.00
	Lap	47.88	83.01	3.92	23.96	3.00	1.93	1.31	0.06
	TLap	47.88	83.01	3.97	23.96	3.08	1.86	1.28	0.00
	GDP	48.10	82.82	3.99	23.87	3.05	4.30	3.42	0.31
	TGDP	48.10	82.82	4.33	23.87	3.63	4.03	3.23	0.00

Theorems 1 and 2. For the Opt mechanism, bias correction using Algorithm 2 is applied. For all the private mechanisms tested, we add the three test statistics constructed using the private data for each of the east coast states. From Theorems 1 and 2, we know that asymptotically this combined test statistic is equivalent to the sum of 15 weighted Chi-square random variables with one degree of freedom, where the weights can be evaluated according to our results. To verify the effectiveness of our proposed procedure under this finite-sample setting, we conduct a parallel simulation study. First, we generate new data sets assuming that the data generating process is as in the H_0 and the sample sizes are the same as those of the three east coast states. To explore the statistical powers, we simulate data sets according to the alternative hypotheses H_1: $P_{NJ,2} = 0.603, 0.603 + 0.005, \ldots, 0.603 + 0.05$, $P_{MA,1} = 0.196, 0.196 - 0.005, \ldots, 0.196 - 0.05$, and all the other terms in the H_1 are kept to be the same as in the H_0. We set the level of significance to be 0.05. The empirical type I errors and powers are summarized in Table 5 and in Figure 4 respectively. We observe that the empirical type I errors are well controlled across all three mechanisms and for all privacy regimes. The Opt mechanism attains the highest power compared to the other methods, especially when the level of privacy-protection is high.

Next we compare the p-values obtained using the true data and using the private data. In the simulation, p-values are evaluated on each of the 500 simulated private data tables and the average is reported in Table 6. Using the true data, we obtain a p-value of 0.0006, suggesting that the distributions of the variable *household type* differ between the three east coast states and the other eight states considered in the NCEDL study. However, in Table 6, we observe that all the p-values yielded from the private data are inflated to some extent, due to the

Table 5: Mean empirical type I errors of the goodness-of-fit test using the private data sets generated when H_0 is true. The reported values are calculated using 500 simulated samples generated according to the NCEDL's settings. Two variables considered are *household type* and *income level*.

		$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
	Opt	0.055	0.052	0.053
	Lap	0.062	0.055	0.054
Household type	TLap	0.058	0.055	0.054
	GDP	0.050	0.051	0.053
	TGDP	0.033	0.050	0.053
	Opt	0.052	0.051	0.050
	Lap	0.052	0.050	0.052
Income level	TLap	0.051	0.050	0.052
	GDP	0.051	0.051	0.052
	TGDP	0.044	0.050	0.051

Table 6: Average p-values of the goodness-of-fit tests using the private data sets in the NCEDL studies. The reported values are calculated using 500 simulated samples. Two variables considered are *household type* and *income level*.

		$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
	Opt	0.372	0.042	0.006
	Lap	0.326	0.068	0.006
Household type	TLap	0.394	0.054	0.010
	GDP	0.450	0.297	0.157
	TGDP	0.673	0.422	0.189
	Opt	0.000	0.000	0.000
	Lap	0.000	0.000	0.000
Income level	TLap	0.000	0.000	0.000
	GDP	0.038	0.000	0.000
	TGDP	0.022	0.000	0.000

information loss as a result of the random noises injected. When the privacy requirement is high at $\epsilon = 0.25$, none of the methods correctly rejects the potentially wrong H_0. When $\epsilon = 0.5$ and if the level of significance is set to be 0.05, only the private data yielded from the Opt mechanism can correctly reject H_0. When the privacy regime is set at $\epsilon = 0.75$, the Lap, TLap and Opt procedures produce satisfactory p-values with the Lap and Opt mechanisms yielding a slightly lower average p-value than TLap. All the p-values yielded from private data generated from GDP or TGDP mechanisms do not give correct inference results. The numerical results here suggest that the chances of the testing signals being undetermined increase with the levels of privacy requirements (decrease with ϵ). The Opt mechanism tend to give the smaller deviations from the truth, thus is more preferred to conduct private inferences.

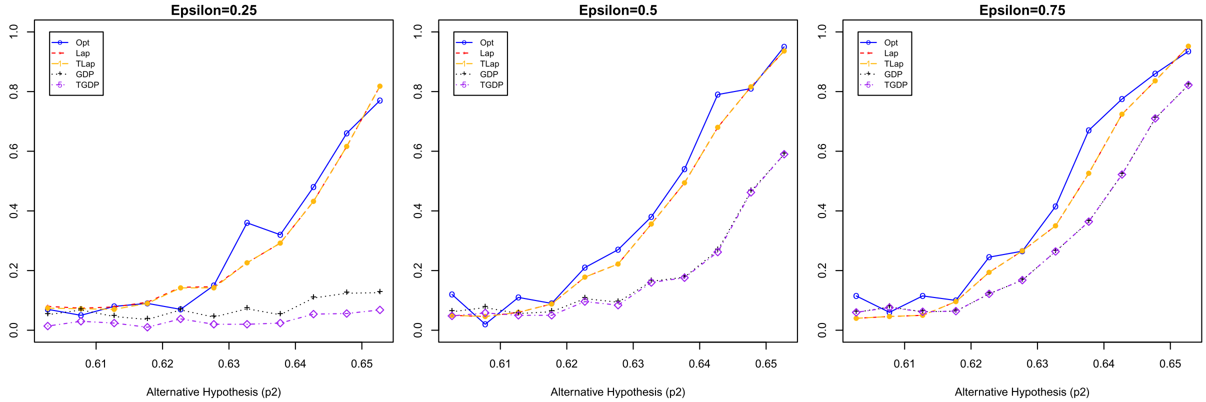

Figure 4: Empirical power comparisons for five privacy procedures: Opt, Lap, TLap, GDP and TGDP on household type data.

Hereafter, we consider another categorical variable *family income* which has 18 categories describing the income levels of the population. The categories are (1) $\leqslant \$5,000$, (2) \$5,001 - \$10,000, (3) \$10,001 - \$15,000,..., (18) $\geqslant \$85,001$. There are 354, 155 and 286 observations for the three east coast states, Massachusetts, New York and New Jersey, respectively. It is a common operation to reduce the total number of categories in a variable when the number of categories is large. The results reported in Theorem 4 provide theoretical support for reducing the number of categories via combining multiple cells into one. Here, for each state, the data is collapsed into frequency tables with three categories: low income ($\leqslant \$20,000$), middle income ($\$20,001 - \$50,000$) and high income ($\geqslant \$50,001$), and denote them as D_{MA}, D_{NY} and D_{NJ} respectively. The three *family income* categories are constructed by respectively merging 4, 6, and 8 out of the original 18 categories into three based on the corresponding income levels.

Goodness-of-fit tests are conducted to check whether the distributions of the variable *family income* for the three east coast states differ from the targeted distribution built from the other states. Following a similar operation as earlier for the variable *household type,* for the 18 categories from the eight other states, the corresponding proportions are $p_0 = (p_{0,1}, \ldots, p_{0,18})^T$, with the value of $p_0 = (0.099, 0.101, 0.119, 0.118, 0.106, 0.103, 0.074, 0.043, 0.036, 0.031, 0.020, 0.026, 0.022, 0.013, 0.012, 0.014, 0.013, 0.050)^T$. The act of combining categories leads to the null-hypothesized value of $p_0^c = (0.437, 0.393, 0.170)^T$ and the composite $H_0 : P_{MA,\ell} = p_{0,\ell}^c, P_{NY,\ell} = p_{0,\ell}^c, P_{NJ,\ell} = p_{0,\ell}^c$, for $\ell = 1, 2, 3$. For the power evaluation, we construct the alternatives from the original 18 categories. Specifically, we set $p_{1,6} = p_{0,6} - k\triangle$, and $p_{1,4} = p_{0,4} + k\triangle$, where $\triangle = 0.0025$ and $k = 0, 1, \ldots, 14$, while keeping the rest of the H_1 the same as the H_0. Note that $p_{0,4}$ has the largest value amongst the 15 cell probabilities. The significance level is set to be 0.05. The empirical type I errors under the H_0 and the statistical powers are shown in Table 5 and Figure 5, respectively. We observe that the empirical type I errors are well controlled across all different mechanisms and for all privacy regimes. The Opt mechanism attains comparable power to the Lap/TLap method, but much larger than those from the GDP/TGDP mechanisms.

Furthermore, we compare the p-values obtained from the true observations and from the private data sets. A p-value of zero is obtained using the true data, suggesting that the distribution for the variable *income level* of the east coast states differs from that of the other eight states in the NCEDL study. For the private data, p-values are evaluated on each of the 500 Monte Carlo samples and the averages are reported in Table 6. In this case, the signal of true

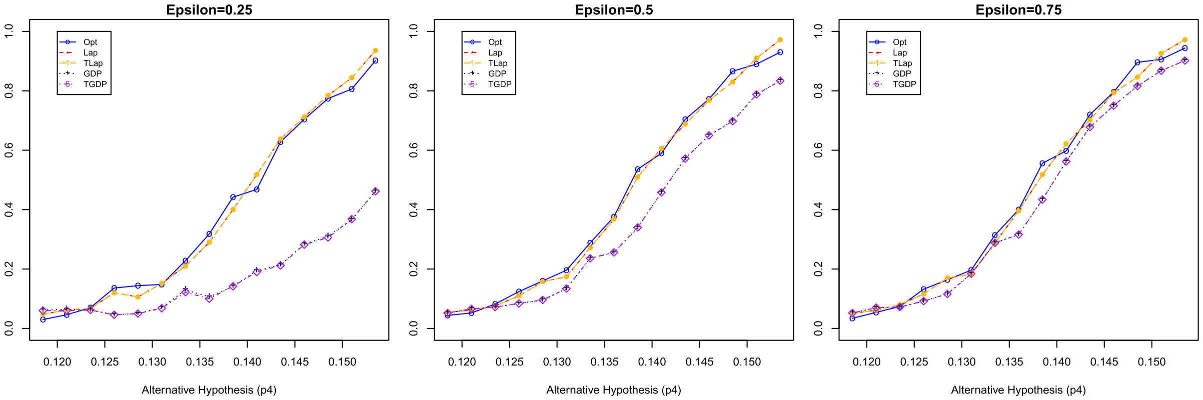

Figure 5: Empirical power comparisons for five privacy procedures: Opt, Lap, TLap, GDP and TGDP on the intra-table merged income level data.

H_1 is strong enough to be detected for all the mechanisms using a level of significance of 0.05, and the Opt approach attains one of the smallest p-values amongst the five mechanisms.

7 Conclusion

In this paper, we recommend an optimal mechanism satisfying ϵ-DP, specifically applicable to one-way frequency tables, in the sense that the expected losses are minimized under a given privacy constraint, where the losses are flexible. Furthermore, we develop valid inference procedures for goodness-of-fit tests for the private data, not only for the optimal mechanism, but also for the Laplace mechanism and the Gaussian mechanism (with/without post-processing of converting negative cells to zero). In fact, the inference procedures developed work for general mechanisms with additive noises. Everyday operations, including merging multiple frequency tables and combining categories within a table, are also considered. The valid inference procedures applicable to the private frequency tables are derived. However, ϵ-DP procedures can be too noisy in practice and it might be desirable to extend the current results to the (ϵ, δ)-DP framework. Currently, the developments of the (ϵ, δ)-DP mainly focus on the Gaussian mechanism, under which the numerical properties are vastly inferior to other mechanisms. The investigation and developments of alternative (ϵ, δ)-DP mechanisms with satisfactory inference characteristics are left as future work to explore.

Supplementary Material

Supplementary Material available online includes proofs of theoretical results and additional simulation study results on inter- and intra-table merging.

Acknowledgement

The authors are grateful to Professor Jun Yan for valuable comments and suggestions. The National Center for Early Development and Learning's (NCEDL) multi-state studies data (M. Clifford et al., 2017) used in the real data application section is subject to the terms and conditions set by the inter-university consortium for political and social research (ICPSR). It shall not be re-distributed without ICPSR's consent. We acknowledge that the data set is used solely for research or statistical purposes and not for investigation of specific subjects in the data.

Funding

This research was partially supported by NSF SES-1846747.

References

Abowd JM, Vilhuber L (2008). How protective are synthetic data? In: *International Conference on Privacy in Statistical Databases*, 239–246. Springer, New York, U.S.

Avella-Medina M (2021). Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116(534): 969–983.

Awan J, Slavković A (2018). Differentially private uniformly most powerful tests for binomial data. In: *Advances in Neural Information Processing Systems* (S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, eds.), volume 31 of *Curran Associates*. Inc., New York, U.S.

Barrientos AF, Reiter JP, Machanavajjhala A, Chen Y (2019). Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28(2): 440–453.

Bowen CM, Liu F (2020). Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2): 280–307.

Campbell Z, Bray A, Ritz A, Groce A (2018). Differentially private ANOVA testing. In: *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, 281–285.

Canonne CL, Kamath G, Steinke T (2020). The discrete gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33: 15676–15688.

Charest AS (2011). How can we analyze differentially private synthetic datasets? *Journal of Privacy and Confidentiality*, 2(2).

Chaudhuri K, Sarwate A, Sinha K (2012). Near-optimal differentially private principal components. In: *Advances in Neural Information Processing Systems*, 989–997.

Couch S, Kazan Z, Shi K, Bray A, Groce A (2019). Differentially private nonparametric hypothesis testing. In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 737–751.

Degue KH, Le Ny J (2018). On differentially private Gaussian hypothesis testing. In: *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 842–847. IEEE.

Ding B, Nori H, Li P, Allen J (2018). Comparing population means under local differential privacy: with significance and power. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Drechsler J (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, volume 201. Springer Science & Business Media, Berlin, Germany.

Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M (2006a). Our data, ourselves: privacy via distributed noise generation. In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 486–503. Springer.

Dwork C, McSherry F, Nissim K, Smith A (2006b). Calibrating noise to sensitivity in private data analysis. In: *Theory of Cryptography Conference*, 265–284. Springer.

Dwork C, Naor M, Pitassi T, Rothblum GN, Yekhanin S (2010). Pan-private streaming algorithms. In: *ICS*, 66–80.

Dwork C, Roth A (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4): 211–407.

Ferrando C, Wang S, Sheldon D (2020). Parametric bootstrap for differentially private confidence intervals. arXiv preprint: https://arxiv.org/abs/2006.07749.

Friedman A, Berkovsky S, Kaafar MA (2016). A differential privacy framework for matrix factorization recommender systems. *User Modeling and User-Adapted Interaction*, 26(5): 425–458.

Gaboardi M, Lim H, Rogers R, Vadhan S (2016). Differentially private Chi-squared hypothesis testing: Goodness of fit and independence testing. In: *International Conference on Machine Learning*, 2111–2120. PMLR.

Geng Q, Viswanath P (2014). The optimal mechanism in differential privacy. In: *2014 IEEE International Symposium on Information Theory*, 2371–2375. IEEE.

Geng Q, Viswanath P (2015). The optimal noise-adding mechanism in differential privacy. *IEEE Transactions on Information Theory*, 62(2): 925–951.

Ghosh A, Roughgarden T, Sundararajan M (2012). Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6): 1673–1693.

Golle P, Partridge K (2009). On the anonymity of home/work location pairs. In: *Pervasive Computing*, 390–397. Springer, Berlin Heidelberg, Berlin, Heidelberg.

Hay M, Machanavajjhala A, Miklau G, Chen Y, Zhang D (2016). Principled evaluation of differentially private algorithms using dpbench. In: *Proceedings of the 2016 International Conference on Management of Data*, 139–154.

Johnson A, Shmatikov V (2013). Privacy-preserving data exploration in genome-wide association studies. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1079–1087.

Kairouz P, Bonawitz K, Ramage D (2016). Discrete distribution estimation under local privacy. In: *International Conference on Machine Learning*, 2436–2444. PMLR.

Karwa V, Krivitsky PN, Slavković AB (2015). Sharing social network data: differentially private estimation of exponential family random graph models. arXiv preprint: https://arxiv.org/abs/1511.02930.

Karwa V, Slavković A (2016). Inference using noisy degrees: differentially private model and synthetic graphs. *The Annals of Statistics*, 44(1): 87–112.

Little RJ (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9(2): 407–426.

Liu C, He X, Chanyaswad T, Wang S, Mittal P (2019). Investigating statistical privacy frameworks from the perspective of hypothesis testing. *Proceedings on Privacy Enhancing Technologies*, 3: 233–254.

Clifford R M, Bryant D, Burchinal M, Barbarin O, Early D, Howes C, et al. (2017). National Center for Early Development and Learning Multistate Study of Pre-Kindergarten. Inter-university Consortium for Political and Social Research [distributor].

Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L (2008). Privacy: Theory meets practice on the map. In: *2008 IEEE 24th International Conference on Data Engineering*, 277–286. IEEE.

McClure D, Reiter JP (2012). Differential privacy and statistical disclosure risk measures: an investigation with binary synthetic data. *Trans. Data Priv.*, 5(3): 535–552.

Mohammed N, Chen R, Fung BC, Yu PS (2011). Differentially private data release for data mining. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 493–501.

Narayanan A, Shmatikov V (2008). Robust de-anonymization of large sparse datasets. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111–125.

Quick H (2019). Generating Poisson-distributed differentially private synthetic data. arXiv preprint: https://arxiv.org/abs/1906.00455.

Raab GM, Nowok B, Dibben C (2016). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3): 67–97.

Raghunathan TE, Reiter JP, Rubin DB (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1): 1–16.

Reiter JP (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21(3): 441–462.

Rinott Y, O'Keefe CM , Shlomo N, Skinner C (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science*, 33(3): 358–385.

Rogers R, Roth A, Smith A, Thakkar O (2016). Max-information, differential privacy, and post-selection hypothesis testing. In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 487–494. IEEE.

Rubin DB (1993). Statistical disclosure limitation. *Journal of Official Statistics*, 9(2): 461–468.

Sheffet O (2017). Differentially private ordinary least squares. In: *International Conference on Machine Learning*, 3105–3114. PMLR.

Snoke J, Raab G, Nowok B, Dibben C, Slavkovic A (2016). General and specific utility measures for synthetic data. arXiv preprint: https://arxiv.org/abs/1604.06651.

Sweeney L (2013). Matching known patients to health records in Washington state data. arXiv preprint: https://arxiv.org/abs/1307.1370.

Task C, Clifton C (2016). Differentially private significance testing on paired-sample data. In: *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)*, 153–161.

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

Wang R, Li YF, Wang X, Tang H, Zhou X (2009). Learning your identity and disease from research papers: information leaks in genome wide association study. In: *Proceedings of the 16th ACM Conference on Computer and Communications Security*, 534–544.

Wang Y, Lee J, Kifer D (2015a). Revisiting differentially private hypothesis tests for categorical data. arXiv preprint: https://arxiv.org/abs/1511.03376.

Wang YX, Fienberg S, Smola A (2015b). Privacy for free: posterior sampling and stochastic gradient monte carlo. In: *International Conference on Machine Learning*, 2493–2502.

Wasserman L, Zhou S (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489): 375–389.

Yu F, Fienberg SE, Slavković AB, Uhler C (2014). Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50: 133–141.