The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Comparing Population Means under Local Differential Privacy: with Significance and Power

Bolin Ding, Harsha Nori, Paul Li, Joshua Allen
{bolind, hanori, paul.li, joshuaa}@microsoft.com
Microsoft, One Microsoft Way, Redmond, WA 98052

Abstract

A statistical hypothesis test determines whether a hypothesis should be rejected based on samples from populations. In particular, randomized controlled experiments (or A/B testing) that compare population means using, e.g., t-tests, have been widely deployed in technology companies to aid in making data-driven decisions. Samples used in these tests are collected from users and may contain sensitive information. Both the data collection and the testing process may compromise individuals' privacy. In this paper, we study how to conduct hypothesis tests to compare population means while preserving privacy. We use the notation of *local differential privacy* (LDP), which has recently emerged as the main tool to ensure each individual's privacy without the need of a *trusted data collector*. We propose LDP tests that inject noise into every user's data in the samples before collecting them (so users do not need to trust the data collector), and draw conclusions with bounded type-I (significance level) and type-II errors ($1-$ power). Our approaches can be extended to the scenario where some users require LDP while some are willing to provide exact data. We report experimental results on real-world datasets to verify the effectiveness of our approaches.

Introduction

Randomized controlled experiments (or A/B testing) and hypothesis tests are used by many companies, e.g., Google, Facebook, Amazon, and Microsoft, to design and improve their products and services (Tang et al. 2010; Panger 2016; Kohavi and Round 2004; Kohavi et al. 2012). These statistical techniques base business decisions on samples of actual customer data collected during experiments to draw more informed conclusions. However, such data samples usually contain sensitive information, e.g., usage statistics of certain apps or services; in order to meet users' privacy expectations and tightening privacy regulations (e.g., European GDPR law), ensuring that these experiments and tests do not breach the privacy of individuals is an important problem.

Differential privacy (DP) (Dwork et al. 2006) has emerged as a standard for the privacy guarantees, and been used by, e.g., Apple (Apple 2017), Google (Erlingsson, Pihur, and Korolova 2014), and Uber (Greenberg 2017). In a well-studied DP model used by, e.g., (Gaboardi et al. 2016), users trust and send exact data to a data collector, who then injects noise in the testing process to ensure DP. However, this model is not applicable in our setup, as users may not trust the data collector (e.g., a tech company) due to potential hacks and leaks (Hackett 2015), and prefer not to have unprivatized data leave their devices. Therefore, we adopt the *local model of differential privacy* (LDP) (Duchi, Jordan, and Wainwright 2013). Under LDP, users do not need to trust the data collector. Before sent to the data collector, each user's data is privatized by a randomized algorithm with the property that the likelihood of any specific output of the algorithm varies little with the input, i.e., the exact data.

In this paper, we study how to conduct hypothesis tests to compare population means (e.g., in A/B testing), while ensuring LDP for each user. We focus on the class of t-tests when presenting our solutions – they can be easily extended for Z-tests if populations follow Normal distributions.

An A/B test splits users randomly into two populations, to give them two different experiences, a *control* and a *treatment*, respectively, and then tests for differences between the two population means in a measure of interest (clicks, usage, monetization, etc.). A *null hypothesis* H_0 is that the two population means are equal or differ by a fixed constant. Statistical tests (e.g., t-tests) are used to determine whether the null hypothesis should be rejected based on random samples from the populations. To measure errors in the conclusions, *type-I error* is the probability of falsely rejecting H_0 when it is true, and *type-II error*, or complement of *statistical power*, is the probability of failing to reject H_0 when it is false. We want a test to have type-I error bounded by a pre-specified threshold, called *significance level*, and have high power.

A typical test has three common steps: 1) compute the observed value of a *test statistic* from samples; 2) calculate the *p-value*, i.e., the probability, under H_0, of a test statistic being more extreme than the observed one; 3) reject H_0 if and only if the p-value is less than the significance level.

Challenges and our contributions. In t-tests (as well as Z-tests when population variances are unknown) that compare population means, sample means and sample variances are the essential terms in the test statistic to be computed in step 1). While there are several approaches, e.g., (Duchi, Jordan, and Wainwright 2013), to estimate means, there is no known technique to estimate sample variances under LDP. In fact, we show that *any* estimator to variances based on a previous

LDP mechanism (Ding, Kulkarni, and Yekhanin 2017) has a very large worst-case error (Proposition 3).

Our first approach, called *estimation-based LDP test*, is based on a seemingly direct idea. We propose a new LDP mechanism to estimate sample variances, using which we obtain an estimation of the observed test statistic in step 1) to calculate the p-value and then draw the conclusion.

One of the most important goals of hypothesis testing is to control the probability of drawing a false conclusion in terms of type-I and type-II errors. While our new LDP variance estimator is of independent interest, the first approach is unsatisfying in achieving this goal, especially when the size of the data domain is large. As errors in both the estimator to sample means and the one to variances are proportional to the domain size, it is hard to bound the error in estimating the test statistic in step 1), so there is no theoretical guarantee on type-I and type-II errors in our first approach.

The second approach we propose, called *transformation-based LDP test*, aims to provide *an upper bound of type-II error* at a pre-specified significance level, i.e., *a hard constraint of type-I error*. The main idea is to look into the relationship between the original distribution of a population and the distribution on the outputs of the LDP data-collection algorithm on users' data, called *transformed distribution*. Instead of estimating sample means and sample variances under LDP, we directly conduct tests based on LDP samples from the transformed distributions – the conclusion can then be translated into a conclusion of the test on the original population (i.e., rejecting or accepting H_0). The upper bound on type-II error during A/B testing is critical in estimating the number of users needed in the samples to detect significant differences between the control and the treatment populations. We derive such an estimation of sample sizes needed to reduce type-II error below a threshold at the specified significance level. This approach can be extended to a hybrid-privacy scenario where some users require LDP while some are willing to provide exact data.

Experiments are conducted on real datasets to verify our theoretical results and the effectiveness of our approaches.

Related work. There are a long line of works on hypothesis testing under the DP model *with* a *trusted* data collector, with genome-wide association studies as a primary application. In this setup, the data collector receives exact samples from users. The first type of approaches inject noise into aggregates (or marginal tables) of data to ensure DP, and compute or estimate the test statistic in step 1) from these noisy aggregates (Fienberg, Rinaldo, and Yang 2010; Karwa and Slavković 2012; Johnson and Shmatikov 2013; Karwa and Slavković 2016). The intuition is that the impact of the DP noise is small when the sample size is large enough (Vu and Slavkovic 2009). However, it is shown that certain tests, e.g., χ^2-tests, perform poorly when used with the estimated statistic (Gaboardi et al. 2016), leading to much higher type-I error than the specified amount. The second type of approaches (Uhlerop, Slavković, and Fienberg 2013; Yu et al. 2014; Wang, Lee, and Kifer 2015; Gaboardi et al. 2016) try to derive the asymptotic distribution of the estimated test statistic. Since this asymptotic distribution cannot be written analytically, Monte Carlo simulations or numerical approximations are used to calculate the p-value in step 2). More recently, for χ^2-tests, unit circle mechanism (Kakizaki, Fukuchi, and Sakuma 2017) utilizes the geometrical property of the test statistics and achieves a sharp reduction on the type-II errors; and independently, new test statistics (Rogers and Kifer 2017) are proposed, so that their asymptotic distributions with DP noise injected are close to the asymptotics of the classical (non-private) tests.

To our best knowledge, our work is the first on statistical hypothesis tests to compare population means under LDP. One of our primary applications is A/B testing in software companies, so LDP is a proper privacy guarantee for each user without the need of trusting the data collector. LDP χ^2-testing is studied in (Rogers 2017), to test goodness of fit and independence for multinomial distributions. LDP hypothesis tests to distinguish between two specific distributions are studied in (Kairouz, Oh, and Viswanath 2014).

Another relevant line of works are about parameter estimations under LDP, including, e.g., mean/density estimations (Duchi, Jordan, and Wainwright 2013; Duchi, Wainwright, and Jordan 2016; Ding, Kulkarni, and Yekhanin 2017), and histogram estimations (Duchi, Wainwright, and Jordan 2013; Kairouz, Bonawitz, and Ramage 2016; Wang et al. 2016; Wang, Wu, and Hu 2016). Communication and computation-efficient mechanisms are developed for histogram estimations over large domains to find heavy hitters (Bassily and Smith 2015; Wang et al. 2017; Bassily et al. 2017). Industrial deployments of LDP techniques on this line enhance privacy using memorization (Erlingsson, Pihur, and Korolova 2014; Fanti, Pihur, and Erlingsson 2016). How to find heavy hitters is also studied in a hybrid-privacy model with both LDP and DP users (Avent et al. 2017).

Preliminaries

Each *user* has a private **real-valued** *counter* $x \in \Sigma = [0, m]$ (to measure, e.g., app usage). Our approaches can be easily extended for general domains like $[-m, m]$, but we focus on $[0, m]$ for the simplicity of presentation. Let $[n] = \{1, 2, \ldots, n\}$ and $X = \{x_i\}_{i\in[n]}$ be a (sample) set of counters from n users. We use $\mu_X = \sum_i x_i/n$ and $s_X^2 = \sum_i (x_i - \mu_X)^2/(n-1)$ to denote the *sample mean* and *sample variance* of X, respectively. We use $\mathbf{X}$ (or $\mathbf{A}$, $\mathbf{B}$) to denote both a population and the distribution of this population, from which a sample X (or A, B) is drawn.

Hypothesis testing to compare population means. Let $\mathbf{A}$ and $\mathbf{B}$ be the distributions of counters in the *control* and the *treatment* populations, respectively. Let $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$ be the *population means* (*expectations*). The *null hypothesis* H_0 is $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} = d_0$, and the *alternative hypothesis* H_1 is, e.g., $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} \neq d_0$. We randomly pick n_A and n_B users from the populations $\mathbf{A}$ and $\mathbf{B}$, respectively, and let $A = \{a_i\}_{i\in[n_A]}$ and $B = \{b_i\}_{i\in[n_B]}$ be the corresponding samples.

A test is an algorithm $\mathfrak{T}$ that takes the two samples A and B and decides whether to reject or accept H_0 at a pre-specified *significance level* α. We require $\mathfrak{T}$ to have *type-I error* at most α, i.e., $\mathbf{Pr}[\mathfrak{T}(A, B; \alpha, H_0) = \text{reject} \mid H_0] \leq \alpha$, and *type-II error* $\mathbf{Pr}[\mathfrak{T}(A, B; \alpha, H_0) = \text{accept} \mid H_1] = \beta$ as small as possible. $1 - \beta$ is called the *statistical power* of $\mathfrak{T}$.

The probability is taken over the randomness from the data generation (of A and B) and the possible randomness in $\mathfrak{T}$.

A key step in a test is to compute the observed value of a test statistic from samples. To compare population means, it is usually a function of six parameters: sample means μ_A and μ_B, sample variances s_A^2 and s_B^2, and sample sizes n_A and n_B. In t-tests, we also need to obtain the *degrees of freedom*. For example, in Welch's t-test, they are, respectively

$$t = \frac{(\mu_A - \mu_B) - d_0}{\sqrt{s_A^2/n_A + s_B^2/n_B}} \text{ and } df = \frac{(s_A^2/n_A + s_B^2/n_B)^2}{\frac{(s_A^2/n_A)^2}{n_A - 1} + \frac{(s_B^2/n_B)^2}{n_B - 1}}. \quad (1)$$

Local model of differential privacy (LDP). In the *local model of differential privacy* (LDP) (Duchi, Wainwright, and Jordan 2013; Bassily and Smith 2015), also called randomized response model (Warner 1965), γ-amplification (Evfimievski, Gehrke, and Srikant 2003), or FRAPP (Agrawal and Haritsa 2005), private data from each user is randomized by an algorithm $\mathfrak{A}$ before being sent to data collector.

Definition 1 (Local model of differential privacy) *A randomized algorithm $\mathfrak{A} : \Sigma \to \mathcal{Z}$ is ϵ-locally differentially private (ϵ-LDP) if for any pair of values $x, y \in \Sigma$ and any subset of output $S \subseteq \mathcal{Z}$, we have that*

$$\mathbf{Pr}[\mathfrak{A}(x) \in S] \leq e^{\epsilon} \cdot \mathbf{Pr}[\mathfrak{A}(y) \in S].$$

One interpretation of LDP is that no matter what output is released from $\mathfrak{A}$, it is approximately equally as likely to have come from one value $x \in \Sigma$ as any other. Unlike the DP model used in (Gaboardi et al. 2016; Rogers and Kifer 2017), users do not need to trust the data collector in LDP.

Problem statement: LDP mean-comparison test. Each user in the control and the treatment has a counter. Two random samples of counters A and B are drawn from the control and the treatment distributions $\mathbf{A}$ and $\mathbf{B}$, respectively. With the null hypothesis H_0: $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} = d_0$, we want to design a test, such that: i) each counter in A and B is collected from the user in an ε-LDP way, ii) its type-I error $\leq$ significance level α, and iii) type-II error is as small as possible.

Building block: 1-bit LDP data collection. We will utilize the following ε-LDP mechanism $\mathfrak{M}_{\varepsilon,m}$ from (Ding, Kulkarni, and Yekhanin 2017) to privatize each counter in samples A and B. For each user with a counter x, it generates a noisy bit (0 or 1), independently, and sends to the data collector

$$\mathfrak{M}_{\varepsilon,m}(x) = \begin{cases} 1, \text{ with probability } \frac{1}{e^{\epsilon}+1} + \frac{x}{m} \cdot \frac{e^{\epsilon}-1}{e^{\epsilon}+1}; \\ 0, \text{ otherwise.} \end{cases} \quad (2)$$

Proposition 1 *The mechanism $\mathfrak{M}_{\varepsilon,m}$ is ε-LDP.*

This mechanism can be interpreted as firstly rounding x to a bit 1 with probability x/m or 0 otherwise, and flipping the bit with probability $\frac{1}{e^{\epsilon}+1}$. It is communication-efficient (only one bit is sent) and can be seen as a simplification of the multidimensional mean-estimation mechanism in (Duchi, Jordan, and Wainwright 2013).

$\mathfrak{M}_{\varepsilon,m}$ can be used for mean estimation. Suppose there are n users: $X = \{x_i\}_{i\in[n]}$. We collect $x_i' = \mathfrak{M}_{\varepsilon,m}(x_i)$ from each user i. The mean $\mu_X = \sum_i x_i/n$ can be estimated from the noisy bits $X' = \mathfrak{M}_{\varepsilon,m}(X) = \{x_i'\}_{i\in[n]}$ as:

$$\hat{\mu}_{\varepsilon,m}(X') = \frac{m}{n} \sum_{i=1}^{n} \frac{x_i' \cdot (e^{\varepsilon}+1) - 1}{e^{\varepsilon} - 1}. \quad (3)$$

It is shown in (Ding, Kulkarni, and Yekhanin 2017) that:

Proposition 2 *The estimator $\hat{\mu}_{\varepsilon,m}(X')$ is unbiased: i) $\mathbf{E}[\hat{\mu}_{\varepsilon,m}(X')] = \mu_X$, and ii) $\mathbf{Var}[\hat{\mu}_{\varepsilon,m}(X')] = \mathrm{O}\Big(\frac{m^2}{n\varepsilon^2}\Big)$.*

Estimation-based LDP Test

Given two samples $A = \{a_i\}_{i\in[n_A]}$ and $B = \{b_i\}_{i\in[n_B]}$, one straightforward starting point is to use $\mathfrak{M}_{\varepsilon,m}$ to collect each counter, to preserve ε-LDP for each user, and estimate parameters (sample means μ_A and μ_B and sample variances s_A^2 and s_B^2) in, e.g., (1), based on $\mathfrak{M}_{\varepsilon,m}(A) = \{\mathfrak{M}_{\varepsilon,m}(a_i)\}$ and $\mathfrak{M}_{\varepsilon,m}(B) = \{\mathfrak{M}_{\varepsilon,m}(b_i)\}$ to conduct a t-test.

We can obtain estimators to μ_A and μ_B using $\hat{\mu}_{\varepsilon,m}$ (3). However, it is difficult to estimate the sample variances from the LDP data collection $\mathfrak{M}_{\varepsilon,m}(A)$ and $\mathfrak{M}_{\varepsilon,m}(B)$. The intuition is as follows. Consider two distributions: a counter from distribution $\mathbf{X}$ is always a constant $m/2$; a counter from $\mathbf{Y}$ is 0 with probability $1/2$, and m otherwise. After applying $\mathfrak{M}_{\varepsilon,m}$ on two samples X and Y from $\mathbf{X}$ and $\mathbf{Y}$, respectively, the LDP samples $\mathfrak{M}_{\varepsilon,m}(X)$ and $\mathfrak{M}_{\varepsilon,m}(Y)$ follow the same distribution (cannot be distinguished from each other), but the gap between s_X^2 and s_Y^2 is $\Omega\big(m^2\big)$ with high probability. In general, we have a hardness result:

Proposition 3 *If each counter in $X = \{x_i\}$ is collected using $\mathfrak{M}_{\varepsilon,m}$, any estimator $\hat{s}^2$ to s_X^2 based on $\mathfrak{M}_{\varepsilon,m}(X) = \{\mathfrak{M}_{\varepsilon,m}(x_i)\}$ has worst-case error $|\hat{s}^2 - s_X^2|$ at least $\Omega\big(m^2\big)$.*

Proof: In the full version (Ding et al. 2017). □

LDP Data Collection for Variance Estimation. The above proposition essentially says that estimating the sample variance of a sample X based on $\mathfrak{M}_{\varepsilon,m}(X)$ leads to unbounded error, as a sample variance itself is bounded by $\mathrm{O}\big(m^2\big)$. In order to obtain a reasonable estimation for sample variances, we need to collect two LDP bits from each user.

Indeed, the sample variance s_X^2 can be rewritten as

$$s_X^2 = \frac{1}{n-1} \sum_i (x_i - \mu_X)^2 = \frac{n}{n-1} \left(\mu_{X^2} - \mu_X^2\right), \quad (4)$$

where X^2 is defined to be $\{x_i^2\}_{i\in[n]}$ and $\mu_{X^2} = \frac{1}{n}\sum_i x_i^2$.

The sequential composability (McSherry 2009) of DP also holds for LDP (considering a dataset with one user). We split the privacy budget into $\varepsilon = \varepsilon_1 + \varepsilon_2$. For each user with a counter x_i, the first bit to be collected is $x_i' = \mathfrak{M}_{\varepsilon_1,m}(x_i)$, which can be also used to estimate mean μ_X. The second bit is $x_i'' = \mathfrak{M}_{\varepsilon_2,m^2}(x_i^2)$, which will be used to estimate μ_{X^2} (the range of x_i^2 is $[0, m^2]$). Note that the two bits are collected with independent randomnesses. From the sequential composability, we preserve $(\varepsilon_1 + \varepsilon_2)$-LDP for each user.

After collecting $X' = \{x_i'\}_{i\in[n]}$ and $X'' = \{x_i''\}_{i\in[n]}$ from users in X, the sample variance can be estimated as

$$\hat{s}^2_{\varepsilon_1,\varepsilon_2,m}(X', X'') = \frac{n\left(\hat{\mu}_{\varepsilon_2,m^2}(X'') - \hat{\mu}^2_{\varepsilon_1,m}(X')\right)}{n-1}, \quad (5)$$

where $\hat{\mu}_{\varepsilon_1,m}$ and $\hat{\mu}_{\varepsilon_2,m^2}$ are defined as in (3). We have the following result about the accuracy of $\hat{s}^2_{\varepsilon_1,\varepsilon_2,m}$.

Proposition 4 *$s^2_{\varepsilon_1,\varepsilon_2,m}(X', X'')$ is an estimator to s^2_X:*
i) $s^2_X - \mathrm{O}\Big(\frac{m^2}{n\varepsilon_1^2}\Big) \le \mathbf{E}\big[\hat{s}^2_{\varepsilon_1,\varepsilon_2,m}(X', X'')\big] \le s^2_X$, *and*
ii) $\mathbf{Var}\big[\hat{s}^2_{\varepsilon_1,\varepsilon_2,m}(X', X'')\big] = \mathrm{O}\Big(\frac{m^4}{n^2\varepsilon_1^4} + \frac{m^4}{n\varepsilon_2^2}\Big)$.

Proof: In the full version (Ding et al. 2017). □

Using Mean/Variance Estimation in Tests. The test $\mathfrak{T}^{\mathsf{est}}_{\varepsilon_1,\varepsilon_2}$ uses the above mechanism to collect data and estimate μ_A, μ_B, s^2_A, and s^2_B under LDP (consider $X = A, B$), and put the estimates back into (1) to calculate t and df in order to conduct t-tests. Refer to the full version (Ding et al. 2017) for detailed description. There is no theoretical guarantee on the testing errors, but from the above discussion, we have:

Theorem 1 *$\mathfrak{T}^{\mathsf{est}}_{\varepsilon_1,\varepsilon_2}$ satisfies $(\varepsilon_1 + \varepsilon_2)$-LDP.*

Transformation-based LDP Test

In this section, we give a LDP testing algorithm with guaranteed significance and power. The main idea is that, if a counter x follows some (unknown) population distribution with a (unknown) mean μ, the LDP bit $\mathfrak{M}_{\varepsilon,m}(x)$ follows a Bernoulli distribution with the mean determined by μ and ε. So in order to compare population means, we can conduct tests directly on the LDP bits and compare Bernoulli means.

The following proposition firstly gives the relationship between the original population distribution and the resulting Bernoulli distribution on the outputs of $\mathfrak{M}_{\varepsilon,m}$.

Proposition 5 *If a counter $x \in [0, m]$ follows a distribution $\mathbf{X}$ with mean $\mu_{\mathbf{X}}$, the LDP bit $x^{\mathsf{bin}} = \mathfrak{M}_{\varepsilon,m}(x)$ (as in (2)) follows a Bernoulli distribution with the mean*

$$p_{\mathbf{X}} = \mathbf{Pr}\big[x^{\mathsf{bin}} = 1\big] = \frac{\mu_{\mathbf{X}}}{m} \cdot \frac{e^\varepsilon - 1}{e^\varepsilon + 1} + \frac{1}{e^\varepsilon + 1}. \quad (6)$$

Proof: Let f be the PDF of $\mathbf{X}$. We have

$$\begin{aligned}
&\mathbf{Pr}\big[x^{\mathsf{bin}} = 1\big] = \int_0^m \mathbf{Pr}\big[x^{\mathsf{bin}} = 1 \mid x\big]\, f(x)dx \\
=&\int_0^m \left(\frac{1}{e^\epsilon + 1} + \frac{x}{m} \cdot \frac{e^\epsilon - 1}{e^\epsilon + 1}\right) f(x)dx \\
=&\frac{1}{e^\varepsilon + 1}\int_0^m f(x)dx + \frac{1}{m} \cdot \frac{e^\epsilon - 1}{e^\epsilon + 1}\int_0^m x f(x)dx \\
=&\frac{1}{e^\varepsilon + 1} \cdot 1 + \frac{1}{m} \cdot \frac{e^\epsilon - 1}{e^\epsilon + 1} \cdot \mathbf{E}[x] \\
=&\frac{\mu_{\mathbf{X}}}{m} \cdot \frac{e^\varepsilon - 1}{e^\varepsilon + 1} + \frac{1}{e^\varepsilon + 1}.
\end{aligned}$$

Therefore, x^{bin} follows a Bernoulli distribution. □

The test process $\mathfrak{T}^{\mathsf{bin}}_\varepsilon$ is described in Algorithm 1. Two samples, $A = \{a_i\}_{i\in[n_A]}$ and $B = \{b_i\}_{i\in[n_B]}$, are drawn from two distributions $\mathbf{A}$ and $\mathbf{B}$ with means $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$. After applying the mechanism $\mathfrak{M}_{\varepsilon,m}$, the LDP bits $A^{\mathsf{bin}} = \mathfrak{M}_{\varepsilon,m}(A) = \{\mathfrak{M}_{\varepsilon,m}(a_i)\}_{i\in[n_A]}$ and $B^{\mathsf{bin}} = \mathfrak{M}_{\varepsilon,m}(B)$ collected (lines 1-5) are two samples from Bernoulli distributions with means $p_{\mathbf{A}}$ and $p_{\mathbf{B}}$. With $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} = d_0$, from (6), we have $p_{\mathbf{A}} - p_{\mathbf{B}} = d_0^{\mathsf{bin}} = (d_0/m)\cdot((e^\varepsilon - 1)/(e^\varepsilon + 1))$. An important observation here is that the relative order between $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$ is the same as the one between $p_{\mathbf{A}}$ and $p_{\mathbf{B}}$, i.e., $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} \ge d_0 \Leftrightarrow p_{\mathbf{A}} - p_{\mathbf{B}} \ge d_0^{\mathsf{bin}}$. Therefore, we can conduct a test on A^{bin} and B^{bin} to compare $p_{\mathbf{A}}$ and $p_{\mathbf{B}}$ with a null hypothesis $H_0^{\mathsf{bin}} : p_{\mathbf{A}} - p_{\mathbf{B}} = d_0^{\mathsf{bin}}$ (line 6), in order to compare $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$ and reject or accept H_0 (line 7).

Input: Two samples $A = \{a_i\}_{i\in[n_A]}$ and $B = \{b_i\}_{i\in[n_B]}$.
Null hypothesis H_0**:** $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} = d_0$.
Parameters: privacy budget ε and significance level α.

1: For users $i = 1$ to n_A do
2: Encode $a_i^{\mathsf{bin}} = \mathfrak{M}_{\varepsilon,m}(a_i)$ and send a_i^{bin} to the server.
3: For users $i = 1$ to n_B do
4: Encode $b_i^{\mathsf{bin}} = \mathfrak{M}_{\varepsilon,m}(b_i)$ and send b_i^{bin} to the server.
5: Server receives $A^{\mathsf{bin}} = \{a_i^{\mathsf{bin}}\}$ and $B^{\mathsf{bin}} = \{b_i^{\mathsf{bin}}\}$.
6: Let the transformed null hypothesis be

$$H_0^{\mathsf{bin}} : p_{\mathbf{A}} - p_{\mathbf{B}} = \frac{d_0}{m} \cdot \frac{e^\varepsilon - 1}{e^\varepsilon + 1},$$

where $p_{\mathbf{A}}$ ($p_{\mathbf{B}}$) is the distribution mean of A^{bin} (B^{bin}).
7: Conduct a t-test with null hypothesis H_0^{bin} on A^{bin} and B^{bin} at significance level α: accept (or reject) H_0 if and only if H_0^{bin} is accepted (or rejected).

Algorithm 1: $\mathfrak{T}^{\mathsf{bin}}_\varepsilon$: Transformation-based LDP Test

Indeed, $\mathfrak{T}^{\mathsf{bin}}_\varepsilon$ preserves ε-LDP for each user from Proposition 1 and the way how A^{bin} and B^{bin} are collected.

Theorem 2 *$\mathfrak{T}^{\mathsf{bin}}_\varepsilon$ satisfies ε-LDP.*

We do need a larger sample size in $\mathfrak{T}^{\mathsf{bin}}_\varepsilon$ to get a satisfactory statistical power than the size needed in a non-private t-test on the real values in A and B. Intuitively, for a fixed gap between population means $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$, the gap between $p_{\mathbf{A}}$ and $p_{\mathbf{B}}$ is smaller if the domains size m is larger or ε is smaller; note that the smaller the gap between $p_{\mathbf{A}}$ and $p_{\mathbf{B}}$ is, the harder for the test $\mathfrak{T}^{\mathsf{bin}}_\varepsilon$ to draw a significant conclusion. Lower bounds of the power of $\mathfrak{T}^{\mathsf{bin}}_\varepsilon$ (or sample sizes needed) are derived in Theorem 3 at a significance level α.

Theorem 3 *$\mathfrak{T}^{\mathsf{bin}}_\varepsilon$ (Algorithm 1) has a significance level α, i.e., type-I error $\le \alpha$. Suppose the alternative H_1: $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} > d_0$ is true with $(\mu_{\mathbf{A}} - \mu_{\mathbf{B}}) - d_0 = \theta$. The statistical power ($1-$ type-II error) of $\mathfrak{T}^{\mathsf{bin}}_\varepsilon$, denoted by $P(\theta)$, is*

$$P(\theta) \triangleq \mathbf{Pr}\big[\mathfrak{T}^{\mathsf{bin}}_\varepsilon(A, B) = \text{reject} \mid (\mu_{\mathbf{A}} - \mu_{\mathbf{B}}) - d_0 = \theta\big] \ge$$

$$1 - \exp\left(-\left(\frac{\theta(e^\varepsilon - 1)}{m(e^\varepsilon + 1)}\sqrt{\frac{2n_An_B}{n_A + n_B}} - \sqrt{\ln\frac{1}{\alpha}}\right)^2\right), \quad (7)$$

if the samples are large enough: $n_A + n_B = \Omega\Big(\frac{m^2}{\theta^2\varepsilon^2}\ln\frac{1}{\alpha}\Big)$.

If we use Normal distributions to approximate Binomial distributions and Student's t-distributions (under the condition that n_A and n_B are large enough), we have:

$$P(\theta) \ge 1 - F\left(F^{-1}(1-\alpha) - \frac{p_\theta}{\hat{\sigma}_{\mathbf{A}+\mathbf{B}}}\right) \ge \quad (8)$$

$$1 - F\left(F^{-1}(1-\alpha) - p_\theta \cdot \sqrt{\frac{4(n_A - 1)(n_B - 1)}{n_A + n_B - 2}}\right) \quad (9)$$

where $p_\theta = \frac{\theta}{m} \cdot \frac{e^\varepsilon - 1}{e^\varepsilon + 1}$, $F(\cdot)$ *is the CDF of the Normal distribution* $\mathfrak{N}(0,1)$*, and the sample variance* $\hat{\sigma}_{\mathbf{A}+\mathbf{B}} =$

$$\sqrt{\frac{\mathbf{1}_{A^{\mathsf{bin}}}/n_A - \mathbf{1}^2_{A^{\mathsf{bin}}}/n_A^2}{n_A - 1} + \frac{\mathbf{1}_{B^{\mathsf{bin}}}/n_B - \mathbf{1}^2_{B^{\mathsf{bin}}}/n_B^2}{n_B - 1}},$$ *where*

$\mathbf{1}_X = |\{x \in X \mid x = 1\}|$ *is the number of* 1*'s in* X. (10)

In particular, if $n_A = n_B = n$ *and we require that the statistic power* $P(\theta) \geq 1 - \beta$*, it suffices to have*

$$n = \left(F^{-1}(1-\alpha) - F^{-1}(\beta)\right)^2 \cdot \frac{1}{2p_\theta^2} + 1. \tag{11}$$

Proof: Let's focus on the case $d_0 = 0$. The proof can be easily generalized for $d_0 > 0$ by adding a constant.

Consider a test with a null hypothesis H_0^{bin} (in line 7 of Algorithm 1) using the following test statistic:

$$z(A^{\mathsf{bin}}, B^{\mathsf{bin}}) = \frac{1}{n_A}\mathbf{1}_{A^{\mathsf{bin}}} - \frac{1}{n_B}\mathbf{1}_{B^{\mathsf{bin}}}$$

($\mathbf{1}_X$ is defined in (10)). Using the linearity of expectation, we have $\mathbf{E}\left[z(A^{\mathsf{bin}}, B^{\mathsf{bin}})\right] = p_{\mathbf{A}} - p_{\mathbf{B}}$ ($p_{\mathbf{X}}$ is defined in (6)).

The proof of (7) is in the full version (Ding et al. 2017), using a weaker test and McDiarmid's inequality.

We now focus on (8)-(9). Let's first state the Normal approximation: *a Binomial distribution* $\mathfrak{B}(n,p)$ *with* n *trials and success probability* p *can be approximated by a normal distribution* $\mathfrak{N}(np, np(1-p))$*, if* n *is large enough.*

From Proposition 5, we have $\mathbf{1}_{A^{\mathsf{bin}}} \sim \mathfrak{B}(n_A, p_{\mathbf{A}})$ and $\mathbf{1}_{B^{\mathsf{bin}}} \sim \mathfrak{B}(n_B, p_{\mathbf{B}})$. Using Normal approximations to $\mathbf{1}_{A^{\mathsf{bin}}}$ and $\mathbf{1}_{B^{\mathsf{bin}}}$, under H_0^{bit} (when $d_0 = 0$), we have

$$\frac{z(A^{\mathsf{bin}}, B^{\mathsf{bin}})}{\sigma_{\mathbf{A}+\mathbf{B}}} \sim \mathfrak{N}(0,1),$$

where $\sigma_{\mathbf{A}+\mathbf{B}} = \sqrt{p_{\mathbf{A}}(1-p_{\mathbf{A}})/n_A + p_{\mathbf{B}}(1-p_{\mathbf{B}})/n_B}$.

In the test (line 7), we use $\hat{\sigma}_{\mathbf{A}+\mathbf{B}}$ in (10) to approximate $\sigma_{\mathbf{A}+\mathbf{B}}$. When n_A and n_B are large enough,

$$\frac{z(A^{\mathsf{bin}}, B^{\mathsf{bin}})}{\hat{\sigma}_{\mathbf{A}+\mathbf{B}}} \sim \mathfrak{N}(0,1),$$

from the Normal approximation to Student's t-distribution.

At a significance level of α, we need to find an rejection threshold z_0 of $z(A^{\mathsf{bin}}, B^{\mathsf{bin}})$, s.t., under H_0^{bin}, $\mathbf{Pr}[\text{reject}] = \mathbf{Pr}\left[z(A^{\mathsf{bin}}, B^{\mathsf{bin}}) \geq z_0\right] \leq \alpha$. Therefore, based on the CDF of the Normal distribution $\mathfrak{N}(0,1)$, we reject $H_0^{\mathbf{bit}}$ iff

$$z(A^{\mathsf{bin}}, B^{\mathsf{bin}}) \geq z_0 = F^{-1}(1-\alpha) \cdot \hat{\sigma}_{\mathbf{A}+\mathbf{B}}.$$

Now let's estimate the statistical power under the alternative hypothesis with $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} = \theta$, or equivalently,

$$p_{\mathbf{A}} - p_{\mathbf{B}} = \frac{\theta}{m} \cdot \frac{e^\varepsilon - 1}{e^\varepsilon + 1} \triangleq p_\theta.$$

Under the above condition, we have

$$\frac{z(A^{\mathsf{bin}}, B^{\mathsf{bin}}) - p_\theta}{\hat{\sigma}_{\mathbf{A}+\mathbf{B}}} \sim \mathfrak{N}(0,1).$$

And thus the statistical power $P(\theta) =$

$$\begin{aligned}
&= \mathbf{Pr}\left[\frac{z(A^{\mathsf{bin}}, B^{\mathsf{bin}})}{\hat{\sigma}_{\mathbf{A}+\mathbf{B}}} \geq F^{-1}(1-\alpha) \;\middle|\; p_{\mathbf{A}} - p_{\mathbf{B}} = p_\theta\right] \\
&= \mathbf{Pr}\left[\frac{z(A^{\mathsf{bin}}, B^{\mathsf{bin}}) - p_\theta}{\hat{\sigma}_{\mathbf{A}+\mathbf{B}}} \geq F^{-1}(1-\alpha) - \frac{p_\theta}{\hat{\sigma}_{\mathbf{A}+\mathbf{B}}}\right] \\
&= 1 - F\left(F^{-1}(1-\alpha) - \frac{p_\theta}{\hat{\sigma}_{\mathbf{A}+\mathbf{B}}}\right) \\
&\geq 1 - F\left(F^{-1}(1-\alpha) - p_\theta \cdot \sqrt{\frac{4(n_A-1)(n_B-1)}{n_A+n_B-2}}\right).
\end{aligned}$$

The sample size lower bound in (11) is directly from (9). □

How to use Theorem 3. All three of (7)-(9) can be used to estimate the lower bound of statistical power, and the largest one can be picked. (8) is likely to be the tightest one, but it needs the sample variance $\hat{\sigma}_{\mathbf{A}+\mathbf{B}}$ (10) in the LDP samples A^{bin} and B^{bin}, while the other two only need the sample sizes. Before we draw samples from populations, (11) can be used to estimate the sample sizes needed. As will be verified in the experiments, the estimated sample sizes are sufficient to guarantee the required statistical power. It is interesting to note that the type-II error of $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$ has a dominated term $\mathrm{O}(-\exp(\sqrt{n}))$ similar to the one of unit circle mechanism (Kakizaki, Fukuchi, and Sakuma 2017) for a different test (χ^2-test) under DP. The additional term $\frac{1}{m}$ is from LDP.

About Laplace mechanism. $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$ can be adapted if each user's counter is collected using the Laplace-perturbation mechanism (Duchi, Jordan, and Wainwright 2013). However, sending Laplace-perturbed counters is costly and we cannot expect better statistical power from it.

Extensions for Hybrid Privacy Requirements

Hybrid privacy model. The transformation-based LDP test $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$ can be extended for population with *hybrid privacy requirements*: more formally, *in a random sample* $A = \{a_i\}$ *(or* $B = \{b_i\}$*) drawn from the distribution* $\mathbf{A}$ *(or* $\mathbf{B}$*), some users require* ε*-LDP, while the others do not.*

Rescaling LDP bits. Indeed, for users who do not require ε-LDP, we can simply send their exact counter a_j (or b_j) to the server. The question is, for those who require ε-LDP, e.g., a_i, how to combine their LDP bits $\mathfrak{M}_{\varepsilon,m}(a_i) \in \{0,1\}$ with exact counters $a_j \in [0,m]$ to conduct hypothesis tests.

The proposed test $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ in Algorithm 2 "re-scales" LDP bits $\mathfrak{M}_{\varepsilon,m}(a_i)$ to form a mixed sample A^{mix} together with the exact counters. For a user a_i who requires ε-LDP, if $\mathfrak{M}_{\varepsilon,m}(a_i) = 0$, $a_i^{\mathsf{mix}} = -m/(e^\varepsilon - 1)$ is sent (line 3), and if $\mathfrak{M}_{\varepsilon,m}(a_i) = 1$, $a_i^{\mathsf{mix}} = me^\varepsilon/(e^\varepsilon - 1)$ is sent (line 4). If a user a_i does not requires ε-LDP, simply send $a_i^{\mathsf{mix}} = a_i$ (line 5). The same process is applied for sample B (lines 7-12). This process can be easily extended if different users require different values of the privacy parameter ε.

We can show that $A^{\mathsf{mix}} = \{a_i^{\mathsf{mix}}\}$ and $B^{\mathsf{mix}} = \{b_i^{\mathsf{mix}}\}$ received by the server (line 13) can be considered as being drawn from distributions $\mathbf{A}^{\mathsf{mix}}$ and $\mathbf{B}^{\mathsf{mix}}$, with means $\mu_{\mathbf{A}}^{\mathsf{mix}} = \mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}^{\mathsf{mix}} = \mu_{\mathbf{B}}$, respectively, but higher variances.

Input: Two samples $A = \{a_i\}_{i\in[n_A]}$ and $B = \{b_i\}_{i\in[n_B]}$.
Null hypothesis H_0**:** $\mu_{\mathbf{A}} - \mu_{\mathbf{B}} = d_0$.
Parameters: privacy budget ε and significance level α.

1: For users $i = 1$ to n_A do
2: If a_i requires ε-LDP then:
3: If $\mathfrak{M}_{\varepsilon,m}(a_i) = 0$ then: $a_i^{\mathsf{mix}} = -m/(e^\varepsilon - 1)$;
4: Else: $a_i^{\mathsf{mix}} = me^\varepsilon/(e^\varepsilon - 1)$.
5: Else: $a_i^{\mathsf{mix}} = a_i$.
6: Send a_i^{mix} to the server.
7: For users $i = 1$ to n_B do
8: If b_i requires ε-LDP then:
9: If $\mathfrak{M}_{\varepsilon,m}(b_i) = 0$ then: $b_i^{\mathsf{mix}} = -m/(e^\varepsilon - 1)$;
10: Else: $b_i^{\mathsf{mix}} = me^\varepsilon/(e^\varepsilon - 1)$.
11: Else: $b_i^{\mathsf{mix}} = b_i$.
12: Send b_i^{mix} to the server.
13: Server receives $A^{\mathsf{mix}} = \{a_i^{\mathsf{mix}}\}$ and $B^{\mathsf{mix}} = \{b_i^{\mathsf{mix}}\}$.
14: Conduct a t-test with the null hypothesis $H_0^{\mathsf{mix}} : \mu_{\mathbf{A}}^{\mathsf{mix}} - \mu_{\mathbf{B}}^{\mathsf{mix}} = d_0$ on A^{mix} and B^{mix} at significance level α: accept (or reject) H_0 iff H_0^{mix} is accepted (or rejected).

Algorithm 2: $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$: For Hybrid Privacy Requirements

Proposition 6 *If a counter $x \in [0, m]$ follows a distribution $\mathbf{X}$ with mean $\mu_{\mathbf{X}}$, x^{mix} (derived as a_i^{mix} or b_i^{mix} in Algorithm 2) follows a distribution $\mathbf{X}^{\mathsf{mix}}$ with mean $\mu_{\mathbf{X}}^{\mathsf{mix}} = \mu_{\mathbf{X}}$.*

Let $\sigma_{\mathbf{X}}^2$ be the variance of $\mathbf{X}$. If each user in $\mathbf{X}$ requires ε-LDP with probability r, the variance of $\mathbf{X}^{\mathsf{mix}}$ is

$$\sigma_{\mathbf{X}^{\mathsf{mix}}}^2 = \sigma_{\mathbf{X}}^2 \cdot (1 - r) + \mathrm{O}\big(m^2/\varepsilon^2\big) \cdot r. \tag{12}$$

Proof: $\mu_{\mathbf{X}}^{\mathsf{mix}} = \mu_{\mathbf{X}}$ is from (6) and the fact that if x requires LDP, $x^{\mathsf{mix}} = \mathfrak{M}_{\varepsilon,m}(x) \cdot m \cdot (e^\varepsilon + 1)/(e^\varepsilon - 1) - m/(e^\varepsilon - 1)$.

The variance of $\mathbf{X}^{\mathsf{mix}}$ can be then calculated as follows:

$$\begin{aligned}
&\sigma_{\mathbf{X}^{\mathsf{mix}}}^2 = \mathbf{E}\big[(x^{\mathsf{mix}} - \mathbf{E}\big[x^{\mathsf{mix}}\big])^2\big] \\
=&\mathbf{E}\big[(x^{\mathsf{mix}} - \mathbf{E}\big[x^{\mathsf{mix}}\big])^2 \mid x \text{ does not require LDP}\big] \cdot (1 - r) \\
+&\mathbf{E}\big[(x^{\mathsf{mix}} - \mathbf{E}\big[x^{\mathsf{mix}}\big])^2 \mid x \text{ requires LDP}\big] \cdot r \\
=&\sigma_{\mathbf{X}}^2 \cdot (1 - r) + \frac{m^2(e^\varepsilon + 1)^2}{(e^\varepsilon - 1)^2} \cdot \mathbf{Var}[\mathfrak{M}_{\varepsilon,m}(x)] \cdot r,
\end{aligned}$$

where $\mathbf{Var}[\mathfrak{M}_{\varepsilon,m}(x)] = \mathrm{O}(1)$ follows from (2). □

Since the distributions $\mathbf{A}$ and $\mathbf{A}^{\mathsf{mix}}$ (or, $\mathbf{B}$ and $\mathbf{B}^{\mathsf{mix}}$) have the same mean, we can conduct a t-test with the null hypothesis $H_0^{\mathsf{mix}} : \mu_{\mathbf{A}}^{\mathsf{mix}} - \mu_{\mathbf{B}}^{\mathsf{mix}} = d_0$ on A^{mix} and B^{mix} in order to accept or reject H_0 on A and B (line 14) – this is because H_0 is a necessary and sufficient condition of H_0^{mix}. For the same reason, the significance and the power of $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ are guaranteed to be the same as those of the t-test on line 14.

Theorem 4 *$\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ (Algorithm 2) satisfied the hybrid privacy model. It has a significance level α, i.e., type-I error $\le \alpha$. Its power is the same as the power of the t-test on line 14.*

Proof: The significance/power guarantee is from the above discussion. The privacy guarantee for each user follows from Proposition 1 and how A^{mix} and B^{mix} are collected. □

We do not give a closed-form lower bound of $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$'s statistical power. Since it is equal to the power of the test conducted on line 14 of Algorithm 2, it depends on the variances of $\mathbf{A}^{\mathsf{mix}}$ and $\mathbf{B}^{\mathsf{mix}}$, which are determined by $\sigma_{\mathbf{A}}^2$, $\sigma_{\mathbf{B}}^2$, r (the fraction of users requiring LDP), and m as in (12). And since $\sigma_{\mathbf{A}}^2$ and $\sigma_{\mathbf{B}}^2$ are usually much smaller than m^2, r has a significant impact on the power: the smaller r is, the smaller the variances $\sigma_{\mathbf{A}^{\mathsf{mix}}}^2$ and $\sigma_{\mathbf{B}^{\mathsf{mix}}}^2$ are, and the larger the power is. This property will be verified in our experiments.

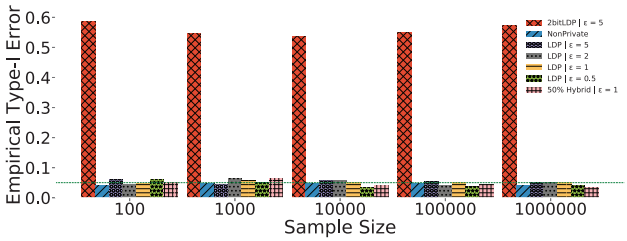


Figure 1: Empirical type-I error of different approaches

More general hybrid privacy models. $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ can be easily extended for the model where each user i requires a different privacy budget ε_i (just replacing ε with ε_i in lines 3-4 and 9-10 of Algorithm 2). In a different model considered for the heavy-hitter problem (Avent et al. 2017), some users require LDP while others only require DP on the testing output: how to conduct effective tests in this model remains open.

Experimental Evaluation

Dataset and parameters. There are 20 million users in this real-world dataset. Each user has a counter with the value in $[0, 15000]$, i.e., $m = 15000$. There are categorical attributes, e.g., country, associated with each user's counter.

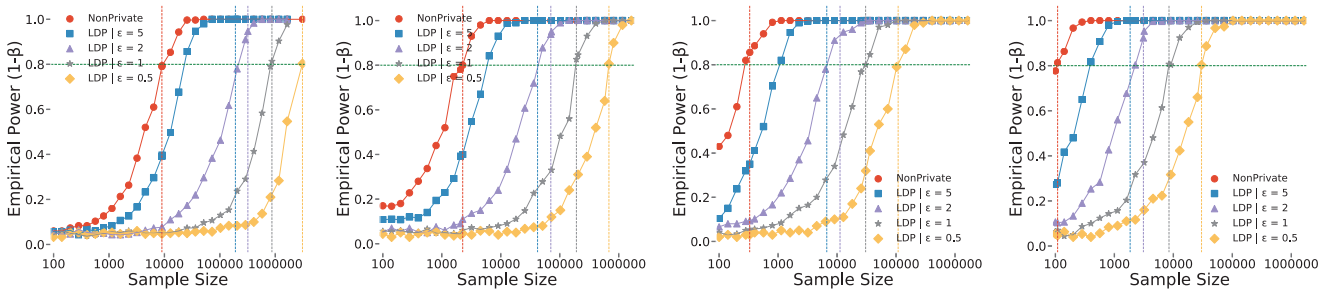
We vary the privacy parameter ε from 0.5 to 5. The prespecified significance level $\alpha = 0.05$, and the null hypothesis is $H_0 : \mu_{\mathbf{A}} - \mu_{\mathbf{B}} = 0$. We draw samples from control/treatment with equal sizes $n_A = n_B$. For each value of sample sizes, we repeat drawing samples and conducting (LDP) tests 1000 times, and report the average (empirical) type-I error and statistical power (1 − type-II error).

Evaluating Significance

In the first set of experiments, we draw samples from distributions with the same means, and thus H_0 holds. We report the average type-I error of different approaches in Figure 1, i.e., the empirical probability that H_0 is rejected.

We conduct Welch's t-test on the non-privatized samples (NonPrivate in Figure 1), estimation-based LDP test $\mathfrak{T}_{\varepsilon_1,\varepsilon_2}^{\mathsf{est}}$ (2bitLDP), LDP test $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$ in Algorithm 1 (LDP), and $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ in Algorithm 2 when 50% users require LDP (50% Hybrid).

Since it is required that type-I error ≤ significance level α, ideally, the empirical type-I error should be close to α or less. Figure 1 verifies Theorems 3-4: for different sample sizes and ε, both $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$ and $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ always have empirical type-I errors close to 0.05. However, $\mathfrak{T}_\varepsilon^{\mathsf{est}}$ does not perform well, even with large sample sizes and large $\varepsilon = \varepsilon_1 + \varepsilon_2 = 2.5 + 2.5$; the reason is that although the sample variance estimator (5) used in $\mathfrak{T}_\varepsilon^{\mathsf{est}}$ has bounded bias (Proposition 4), its variance is too high as m is large, and thus its empirical



(a) Gap $\theta = 60$ (b) Gap $\theta = 120$ (c) Gap $\theta = 300$ (d) Gap $\theta = 600$

Figure 2: Empirical power ($1 -$ type-II error) of different approaches in different scenarios for varying θ and ε

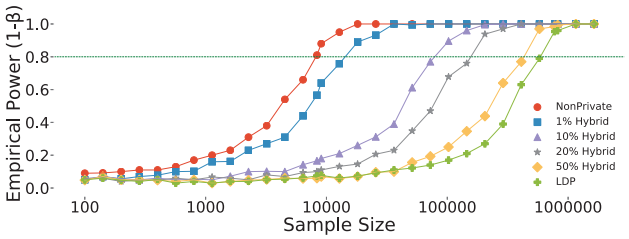


Figure 3: Empirical power ($1 -$ type-II error) of $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ for hybrid privacy requirements with varying LDP fraction

error has a non-negligible impact on the test statistic, leading to much higher type-I error than the pre-specified α.

Evaluating Power

We will evaluate the power of $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$ (LDP in Figure 2) and $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ (x% Hybrid in Figure 3) in the rest experiments, as they have been shown to satisfy the significance level empirically.

We draw samples from two populations with means differing by some constant θ, and thus H_0 should be rejected. We report the average power in Figures 2-3, i.e., the empirical probability that H_0 is (correctly) rejected.

Varying gap θ between control and treatment. We vary θ between two sub-populations from 60 to 600 in Figures 2(a)-2(d). In particular, $\theta = 60$ is a real scenario that we are comparing populations between two countries A and B. We inject shifts to create synthetic cases with $\theta =$ 120-600.

As θ grows, both $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$ (for different ε) and non-privatized Welch's t-test need smaller samples to achieve an empirical power 0.8 (a threshold commonly used in practice). Intuitively, the larger the mean gap θ is, the easier it is for a test to distinguish the two populations. This trend is also consistent with theoretical bounds of the power derived in Theorem 3.

For $\theta = 60$, to achieve an empirical power 0.8, $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$ (with $\varepsilon = 5$) needs roughly three times as many samples as the non-privatized Welch's t-test does. It is totally acceptable, because, with the strong LDP privacy guaranteed for each user without the need of trusting the data collector, more users would be willing to share their data (in an LDP way).

Theorem 3 gives a way (11) to estimate the sample size needed to achieve certain power in $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$. We can verify this analytical result here: in Figures 2(a)-2(d), for each ε, we plot a dashed line in the same color as the color of the corresponding power curve – this dashed line, calculated by (11), denotes the sample size needed to achieve a power of 0.8. It turns out to be a "safe" estimation: samples with this size always give the required or better power in Figure 2.

Hybrid privacy requirements. We evaluate our test $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ for hybrid privacy requirements in the scenario with $\theta = 60$ and $\varepsilon = 1$ in Figure 3: "$x\%$ Hybrid" represents $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ on a population with a random portion of $x\%$ users requiring 1-LDP, and "LDP" represents $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$. Intuitively, the less users require LDP, the easier it is for us to conduct tests. This intuition is consistent with the empirical performance of $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$, which calibrates noise for each LDP user and mixes them with exact samples from those who do not require LDP.

Even within one sample, LDP users may follow a different distribution from non-LDP ones. So $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ still needs larger samples than the non-private test does. But the sample size needed to achieve a high power (e.g., 0.8) decreases quickly as the LDP ratio goes down: when 50% users requires LDP, the sample size needed in $\mathfrak{T}_\varepsilon^{\mathsf{mix}}$ is around half of the one in $\mathfrak{T}_\varepsilon^{\mathsf{bin}}$; and when 1% users requires LDP, the sample size needed is close to the one in the non-private t-test.

Conclusion

We study how to conduct hypothesis testing for comparing population means under LDP. We propose two approaches. Both inject noise into each user's data in the samples before sending it to the data collector to ensure LDP. The first one, called estimation-based LDP test, decodes LDP samples aggregatively at the data collector to recover the observed test statistics. The second one, called transformation-based LDP test, studies the relationship between the population distributions and the distributions of LDP samples. It conducts transformed tests directly on LDP samples and converts conclusions for the original tests. The second one has provable significance and lower bounds of power, and it can be extended for population with hybrid privacy requirements.

References

Agrawal, S., and Haritsa, J. R. 2005. A framework for high-accuracy privacy-preserving mining. In *Proceedings of the 21st International Conference on Data Engineering (ICDE)*, 193–204.

Apple. 2017. https://developer.apple.com/library/content/releasenotes/General/WhatsNewIniOS/Articles/iOS10.html.

Avent, B.; Korolova, A.; Zeber, D.; Hovden, T.; and Livshits, B. 2017. BLENDER: enabling local search with a hybrid differential privacy model. In *26th USENIX Security Symposium (USENIX Security)*, 747–764.

Bassily, R., and Smith, A. D. 2015. Local, private, efficient protocols for succinct histograms. In *Proceedings of the 47th ACM Symposium on Theory of Computing (STOC)*, 127–135.

Bassily, R.; Nissim, K.; Stemmer, U.; and Thakurta, A. 2017. Practical locally private heavy hitters. *CoRR* abs/1707.04982.

Ding, B.; Nori, H.; Li, P.; and Allen, J. 2017. Comparing population means under local differential privacy: with significance and power. *CoRR*.

Ding, B.; Kulkarni, J.; and Yekhanin, S. 2017. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

Duchi, J. C.; Jordan, M. I.; and Wainwright, M. J. 2013. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 429–438.

Duchi, J. C.; Wainwright, M. J.; and Jordan, M. I. 2013. Local privacy and minimax bounds: Sharp rates for probability estimation. In *Advances in Neural Information Processing Systems 26 (NIPS)*, 1529–1537.

Duchi, J. C.; Wainwright, M. J.; and Jordan, M. I. 2016. Minimax optimal procedures for locally private estimation. *CoRR* abs/1604.02390.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *3rd Theory of Cryptography Conference (TCC)*, 265–284.

Erlingsson, Ú.; 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)*, 1054–1067.

Evfimievski, A.; Gehrke, J.; and Srikant, R. 2003. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems (PODS)*, 211–222.

Fanti, G. C.; Pihur, V.; and Erlingsson, Ú. 2016. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. *PoPETs* 2016(3):41–61.

Fienberg, S. E.; Rinaldo, A.; and Yang, X. 2010. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Proceedings of Privacy in Statistical Databases (PSD) 2010*, 187–199.

Gaboardi, M.; Lim, H.; Rogers, R. M.; and Vadhan, S. P. 2016. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, 2111–2120.

Greenberg, A. 2017. https://www.wired.com/story/uber-privacy-elastic-sensitivity/.

Hackett, R. 2015. http://fortune.com/2015/08/26/ashley-madison-hack/.

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

Kairouz, P.; Bonawitz, K.; and Ramage, D. 2016. Discrete distribution estimation under local privacy. 2436–2444.

Kairouz, P.; Oh, S.; and Viswanath, P. 2014. Extremal mechanisms for local differential privacy. In *Advances in Neural Information Processing Systems 27 (NIPS)*, 2879–2887.

Kakizaki, K.; Fukuchi, K.; and Sakuma, J. 2017. Differentially private chi-squared test by unit circle mechanism. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1761–1770.

Karwa, V., and Slavković, A. B. 2012. Differentially private graphical degree sequences and synthetic graphs. In *Proceedings of Privacy in Statistical Databases (PSD) 2012*, 273–285.

Karwa, V., and Slavković, A. 2016. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Ann. Statist.* 44(1):87–112.

Kohavi, R., and Round, M. 2004. Front line internet analytics at amazon.com. *Emetrics Summit*.

Kohavi, R.; Deng, A.; Frasca, B.; Longbotham, R.; Walker, T.; and Xu, Y. 2012. Trustworthy online controlled experiments: five puzzling outcomes explained. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 786–794.

McSherry, F. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 19–30.

Panger, G. 2016. Reassessing the facebook experiment: critical thinking about the validity of big data research. *Information, Communication & Society* 19(8):1108–1126.

Rogers, R., and Kifer, D. 2017. A new class of private chi-square hypothesis tests. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 991–1000.

Rogers, R. 2017. Leveraging privacy in data analysis. *PhD Dissertation*.

Tang, D.; Agarwal, A.; O'Brien, D.; and Meyer, M. 2010. Overlapping experiment infrastructure: more, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 17–26.

Uhlerop, C.; Slavković, A.; and Fienberg, S. E. 2013. Privacy-preserving data sharing for genome-wide association studies. *The Journal of Privacy and Confidentiality* 5(1):137–166.

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

Wang, S.; Huang, L.; Wang, P.; Nie, Y.; Xu, H.; Yang, W.; Li, X.; and Qiao, C. 2016. Mutual information optimally local private discrete distribution estimation. *CoRR* abs/1607.08025.

Wang, T.; Blocki, J.; Li, N.; and Jha, S. 2017. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security)*, 729–745.

Wang, Y.; Lee, J.; and Kifer, D. 2015. Differentially private hypothesis testing, revisited. *CoRR* abs/1511.03376.

Wang, Y.; Wu, X.; and Hu, D. 2016. Using randomized response for differential privacy preserving data collection. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference (EDBT/ICDT Workshops)*.

Warner, S. L. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60(309):63–69.

Yu, F.; Fienberg, S. E.; Slavkovic, A. B.; and Uhler, C. 2014. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics* 50:133–141.