Privacy Aware Experimentation over Sensitive Groups: A General Chi Square Approach

Rina Friedberg RFRIEDBERG@LINKEDIN.COM and **Ryan Rogers** RROGERS@LINKEDIN.COM

Abstract

As companies work to provide the best possible experience for members, users, and customers, it is crucial to understand how different people – particularly individuals from sensitive groups - have different experiences. For example, do women visit our platform less frequently than members of other genders? Or perhaps, are people with disabilities disproportionately affected by a change to our user interface? However, to run these statistical tests or form estimates to answer these questions, we need to know sensitive attributes. When dealing with personal data, privacy techniques should be considered, especially when we are dealing with sensitive groups, e.g. race/ethnicity or gender. We study a new privacy model where users belong to certain sensitive groups, and we show how to conduct statistical inference on whether there are significant differences in outcomes between the various groups. We introduce a general chi-squared test that accounts for differential privacy in group membership, and show how this covers a broad set of hypothesis tests, improving statistical power over tests that ignore the noise due to privacy.

Keywords: Hypothesis testing, differential privacy, experimentation

1. Introduction

When measuring the impact that new products or enhancements have on users, we rely on A/B testing to help us determine if the new feature significantly improves the user experience. To determine whether certain user groups are negatively impacted, although overall metrics might improve, we would like to measure the outcomes across these groups. However, to run these statistical tests, we need to know sensitive group information, which should be kept private. We start by covering the various existing privacy models one might consider, and why they fall short in providing privacy of the group that each sample belongs to or drastically impact utility for the task at hand.

One approach would be to keep a dataset of sensitive attributes of users in a secure environment that cannot be directly accessed. However, if multiple experiments are conducted on the secure dataset and only the aggregated results are revealed, it may still be possible to reconstruct the sensitive attributes Dinur and Nissim (2003). Privacy mitigations such as differential privacy Dwork et al. (2006) can be used to ensure each outcome for each experiment has a sufficient amount of noise to protect the privacy of each individual. However, with potentially hundreds of A/B tests ran every day, the overall privacy loss as defined in differential privacy (DP) becomes massive relatively quickly. We could add so much noise that each result is worthless, but we want to ensure both usefulness of A/B testing and privacy of these features over a very long or even infinite time horizon.

Another approach is to create a synthetic dataset subject to (central) DP, with some generative process (Near and Darais, 2021). Note that we would no longer need to worry

about the privacy loss accumulating with each experiment, since each computation would be post-processing on the synthetic data, and privacy loss cannot be increased due to post-processing (Dwork et al., 2006). Unfortunately, for A/B testing, we want to be able to perform a join on the data by some user ID so that we know which members were in the control or treatment groups and what their outcomes were, however this is not possible, as the synthetic data via generative process would no longer have individual IDs.

We then turn to local differential privacy as a solution to creating a synthetic dataset while keeping member IDs - this ensures that (a) we have a privatized dataset that can be used an unlimited number of times without compromising privacy of which group each user belongs to, and (b) allows us to join with another dataset by a unique identifier for each user. For our setting, we are not concerned with the privacy of the outcomes themselves, as these outcomes are traditionally used internally to determine the impact of different experiments. Recent work from Juarez and Korolova (2022) has also considered privatizing the group membership with local DP, but also privatize the outcome, as the outcome might be "correlated with group membership." This is precisely what we want to test. Considering the privacy of only the features can also be thought of as a complement to some recent work on label-differential privacy where the labels themselves are the private information while the features are not treated as sensitive.

The privacy model we consider here can be viewed as a variant to a traditional local DP set up, as local DP would consider privatizing the features, e.g. the group membership, and the outcome together. In this privacy model, which we call the *local group DP* (LGDP) model, we will only privatize the group membership. We will consider several scenarios with different privacy mechanisms.

First, we consider binary outcomes, which can easily be extended to categorical outcomes; we consider testing a difference in proportions between two groups, as well as χ^2- independence testing for a difference between proportions over multiple sensitive groups simultaneously. We then consider outcomes that are real valued. We consider both confidence intervals and hypothesis tests for the difference in means, as well as differences in means across multiple sensitive groups simultaneously via one-way Analysis of Variation (ANOVA) tests. Lastly, we will consider the application to A/B tests where users are split across two treatment variants randomly and which variant each user belongs to is known to the analyst and does not need privatization. In this case, we test whether there is a significant change from the control to treatment group in the difference of outcomes between two sensitive groups. In particular, this can be used to determine whether a feature creates a smaller difference between means of sensitive groups. We now summarize our contributions below:

- We present a new privacy model over sensitive groups called local group DP that is less restrictive than the local DP setting, and would privatize the group for each sample.
- We present a unified framework based on general χ^2 tests that will allow us to easily derive new asymptotically valid statistical tests for various privacy mechanisms and hypothesis tests, including Z-tests, independence tests, t-tests, and ANOVA.
- We demonstrate that confidence intervals with traditional methods that do not factor in the privacy mechanism lead to false conclusions, while our approach is able to compute empirically valid confidence intervals that converges to the original test as $\epsilon \to \infty$.
- We show that when testing for a significant difference in means across $g > 2$ groups, we achieve higher statistical power for the same level of privacy than traditional approaches.

- We apply our general framework to A/B tests for testing whether the difference across sensitive groups has changed between the control and treatment groups.

1.1. Related Work

Airbnb recently described an approach that could be used to determine disparate impacts across sensitive groups in Project Lighthouse (Basu et al., 2022). Project Lighthouse incorporates k-anonymity and ℓ-diversity to protect the privacy of a dataset, but would require repeating the procedure for a new experiment that might use additional features (see Section A.1 for more discussion).

We now cover several additional related works, although we point out that these do not consider the privacy model we do here. There have been several works that have considered statistical hypothesis tests under local DP, including simple tests like distribution testing or goodness of fit and composite tests like independence testing (Gaboardi and Rogers, 2018; Sheffet, 2018; Acharya et al., 2019a) as well as A/B testing (Ding et al., 2018; Waudby-Smith et al., 2022). Further, in the local setting, there has been work in mean estimation and confidence intervals (Gaboardi et al., 2019; Joseph et al., 2019; Waudby-Smith et al., 2022) as well as minimax optimal schemes (Duchi et al., 2018; Bhowmick et al., 2019). We also mention several works on hypothesis testing in the global model of DP, including distribution testing or goodness of fit (Cai et al., 2017; Canonne et al., 2019, 2020; Awan and Slavkovic, 2020) and more composite tests, like independence testing (Vu and Slavkovic, 2009; Uhler et al., 2013; Yu et al., 2014; Wang et al., 2015; Gaboardi et al., 2016; Kifer and Rogers, 2017; Kakizaki et al., 2017), ANOVA Campbell et al. (2018); Swanberg et al. (2019), and linear regression (Sheffet, 2017). Also in the global model there has been work on mean estimation and confidence intervals (Karwa and Vadhan, 2017; Wang et al., 2019; Biswas et al., 2020; Covington et al., 2021).

2. Privacy Preliminaries

We start with the definition of local DP (Warner, 1965; Evfimievski et al., 2003; Kasiviswanathan et al., 2011), which will treat the group and the outcome of each group to be sensitive information, and contrast that with the privacy model we consider here.

Definition 1 *An algorithm $M : \mathcal{X} \to \mathcal{Y}$ is ϵ-locally differentially private if for all possible inputs $x, x' \in \mathcal{X}$ and for all outcomes $S \subseteq \mathcal{Y}$ we have* $\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(x') \in S]$.

We will instead focus on the case where a user's outcome is already known to the data analyst, but the group membership is not.

Definition 2 *An algorithm $M : \{1, \ldots, g\} \times \mathcal{O} \to \mathcal{Y}$ is ϵ-local group DP (LGDP) if $M'(\cdot, o) : \{1, \ldots, g\} \to \mathcal{Y}$ is ϵ-local DP for all outcomes $o \in \mathcal{O}$, i.e. ϵ-local DP in its first argument.*

This less restrictive definition of privacy ensures deniability for the group a user belongs to. In particular, for attributes, like race and ethnicity, the information is incredibly sensitive and does not change over time, while outcomes, such as salary or conversion on current ad campaigns, can change over time. Another way to interpret LGDP is that it is local DP on the group membership of each user, disregarding the outcomes that are joined with the user.

We now present three fundamental local DP mechanisms for privatizing categorial inputs, where one will actually be a special case of another. We assume that the sensitive data is a group $j \in [g]$. There have been several great works on improving run time and communication costs of these mechanisms (Acharya et al., 2019b; Feldman et al., 2022); we are primarily interested in the tradeoffs between statistical power and privacy.

Definition 3 (g-Randomized Response[1] from Warner (1965)) *The g-randomized response mechanism $M : \{1, \dots, g\} \to \{1, \dots, g\}$ returns its input with probability $\frac{e^\epsilon}{e^\epsilon + g - 1}$ and otherwise uniformly selects a different outcome with equal probability, i.e.*

$$\Pr[M(j) = j] = \frac{e^\epsilon}{e^\epsilon + g - 1}, \qquad \text{and } \Pr[M(j) = \ell] = \frac{1}{e^\epsilon + g - 1} \ \forall \ell \neq j.$$

Definition 4 (Bit Flipping[2] from Erlingsson et al. (2014)) *The bit flipping mechanism $M : \{1, \dots, g\} \to \{0, 1\}^g$ with input j creates a vector of length g with all zeros except a one in position j, iterates through each coordinate, and flips the bit with probability $\frac{1}{e^{\epsilon/2}+1}$ or otherwise keeps it the same.*

Definition 5 (Subset Mechanism from Ye and Barg (2017)) *The subset mechanism with parameter $k < g$ is $M : \{1, \dots, g\} \to \mathcal{Y}_k$ where $\mathcal{Y}_k = \{y \in \{0, 1\}^g : \sum_{i=1}^d y_i = k\}$. For $M(j) = (M(j)[\ell] : \ell \in \{1, \dots, g\})$, the probability the j-th coordinate $M(j)[j] = 1$ is $\frac{ke^\varepsilon}{ke^\varepsilon + g - k}$, and otherwise $M(j)[j] = 0$. If $M(j)[j] = 1$, then we uniformly select $k-1$ distinct coordinates from the set $\{1, \dots, g\} \setminus \{j\}$ to be 1. If $M(j)[j] = 0$, then we uniformly select k distinct coordinates from the set $\{1, \dots, g\} \setminus \{j\}$ to be 1.*

The subset mechanism can be thought of as a way to unify the g-randomized response and the bit flipping mechanism. In fact when $k = 1$, we recover the g-randomized response mechanism. We will pick $k = \lceil \frac{g}{e^\epsilon + 1} \rceil$, as Proposition III.2 in Ye and Barg (2017) claims this to be the optimal choice for k.

3. General Chi Squared Theory

Our analysis approach is the general chi-squared test, which extends a chi-squared test for independence to much broader hypothesis tests (Ferguson, 1996) (Chapter 23). Previous work from Kifer and Rogers (2017) and Gaboardi and Rogers (2018) used the general χ^2 tests to design valid hypothesis tests for categorical data in goodness of fit and independence testing subject to DP. Here, we summarize the approach adapted for LGDP. Consider random vectors $\{Y_i \in \mathbb{R}^d : i \in \{1, \dots, n\}\}$ where each Y_i is selected i.i.d. from some distribution with parameters $\theta^\star \in \Theta$ where Θ is a non-empty open subset of $\mathbb{R}^\nu$, with $\nu < d$. Let $\bar{Y}$ denote the sample average, and suppose we also have a function A mapping $\mathbb{R}^\nu$ to $\mathbb{R}^d$.

We write the null hypothesis H_0 as $\theta^\star \in \Theta$ such that with $\bar{Y} := \frac{1}{n} \sum_{i=1}^n Y_i$ we have,

$$\sqrt{n} \left(\bar{Y} - A(\theta^\star) \right) \xrightarrow{D} N(0, C(\theta^\star)) \tag{1}$$

where $C(\theta) \in \mathbb{R}^{d \times d}$ is a covariance matrix and we write $\xrightarrow{D}$ to denote convergence in distribution as sample size n increases. We measure the distance between $\bar{Y}$ and $A(\theta)$ with a test statistic given by the following, where $M \in \mathbb{R}^d \times \mathbb{R}^d$ is symmetric positive-definite.

$$D^{(n)}(\theta) = n \left(\bar{Y} - A(\theta) \right)^\intercal M(\theta) \left(\bar{Y} - A(\theta) \right). \tag{2}$$

Note that we make standard regularity assumptions about $A(\theta)$, $M(\theta)$, and Θ, detailed as Assumption 4.1 in Kifer and Rogers (2017). When $\theta^\star$ is not known, we need to estimate a good parameter $\hat{\theta}^{(n)}$ to plug into (2). Kifer and Rogers (2017) show that we need only use a rough estimate of $\theta^\star$, denoted $\phi(\bar{Y})$, which converges in probability to $\theta^\star$. So, we set our parameter estimate to be $\hat{\theta}^{(n)} = \arg\min_{\theta\in\Theta} \hat{D}^{(n)}(\theta)$, and the test statistic becomes $\hat{D}^{(n)}(\hat{\theta}^{(n)})$. Statistical theory from Kifer and Rogers (2017) establishes that under the null,

$$\min_{\theta\in\Theta\subseteq\mathbb{R}^\nu} \hat{D}^{(n)}(\theta) \xrightarrow{D} \chi^2_{d'-\nu}.$$

In Algorithm 1, we outline our general approach to designing new hypothesis tests with $1-\alpha$ significance, which, given our assumptions hold, have asymptotically a χ^2 distribution.

Algorithm 1 General χ^2 approach

- Based on the hypothesis test, compute a random vector $Y_i \in \mathbb{R}^d$ that are sampled i.i.d. from the unknown population distribution with parameter space $\Theta \subset \mathbb{R}^\nu$ where $\nu < d$.
- Compute the covariance matrix $C(\theta)$ of Y_i under the null hypothesis for general $\theta \in \Theta$.
- Compute estimates $\hat{\theta}_n$ that converge in probability to $\theta^\star$ and write $\hat{C} := C(\hat{\theta}_n)$.
- Calculate the (general) inverse of $C(\hat{\theta}_n)^\dagger$, where $C(\theta^*)$ has rank at most $d' \leq d$.
- Compute the χ^2 test statistic $\hat{D}$, and if $\hat{D} > \chi^2_{d'-\nu,1-\alpha}$, then reject the null hypothesis.

$$\hat{D} = \min_{\theta\in\Theta} \left\{ n \left(\bar{Y} - \mathbb{E}[Y_i;\theta]\right)^\intercal C(\hat{\theta}_n)^\dagger \left(\bar{Y} - \mathbb{E}[Y_i;\theta]\right) \right\} \tag{3}$$

4. Difference in Proportions with Binary Outcomes

Suppose we have a binary outcome, and we want to test for a difference between the success probabilities across two sensitive groups. We will have data $\{(G_i, X_i)\}_{i=1}^n$, where we first sample the group X_i belongs to as $G_i \sim \text{Bern}(\pi) + 1$, for an unknown $\pi \in [0,1]$; then we have $X_i|G_i \sim \text{Bern}(p_{G_i})$, where $p_g \in [0,1]$ is the probability of success for group $g \in \{1,2\}$. We want to test the null hypothesis $\mathtt{H}_0 : p_1 = p_2 + \Delta$. For ε-LGDP, we can use randomized response $M : \{1,2\} \to \{1,2\}$ where $\Pr[M(g) = g] = \frac{e^\varepsilon}{1+e^\varepsilon}$.

We show how to build a general χ^2 test for testing $\mathtt{H}_0$, adopting the general theory outlined in Section 3, which was also used to derive tests in the local DP setting by Gaboardi and Rogers (2018). There have been several works on privatizing χ^2 tests, even in the more restrictive local DP setting (Gaboardi and Rogers, 2018; Sheffet, 2018; Acharya et al., 2019a).

In Table 1 we present a contingency table of outcomes, where Group 1 has probability of success p_1 and Group 2 has probability of success p_2. The LGDP setting means that individuals can move across columns (groups), but not rows (outcomes).

To privatize group membership, we will use randomized response and privacy loss parameter $\varepsilon > 0$. For $i \in [n]$ and $g \in \{1,2\}$, let $Z_i^\varepsilon[g] \sim \text{Bern}\left(\frac{e^\varepsilon}{1+e^\varepsilon}\right)$, where data point i will swap groups when $Z_i = 0$. When we sample n combined outcomes over groups 1 and 2, we will consider a single multinomial random variable with the outcome probabilities flattened

Table 1: Contingency Table giving outcome probabilities.

Outcome Probs	Group 1	Group 2
Success	$Y[1,1]$	$Y[1,2]$
Failure	$Y[2,1]$	$Y[2,2]$

to an array; we can write this as the following, where $\mathbf{0}$ denotes the 2×2 matrix of zeros.

$$\begin{aligned} Y_i^\varepsilon &= \begin{pmatrix} Z_i^\varepsilon[1] \cdot W_i \cdot X_i[1] + (1 - Z_i^\varepsilon[2]) \cdot (1 - W_i) \cdot X_i[2] \\ (1 - Z_i^\varepsilon[1]) \cdot W_i \cdot X_i[1] + Z_{i,1}^\varepsilon \cdot (1 - W_i) \cdot X_i[2] \\ Z_i^\varepsilon[1] \cdot W_i \cdot (1 - X_i[1]) + (1 - Z_{i,1}^\varepsilon) \cdot (1 - W_i) \cdot (1 - X_i[2]) \\ (1 - Z_i^\varepsilon[1]) \cdot W_i \cdot (1 - X_i[1]) + Z_i^\varepsilon[2] \cdot (1 - W_i) \cdot (1 - X_i[2]) \end{pmatrix} \\ &= \begin{bmatrix} Z_i^\varepsilon & \mathbf{0} \\ \mathbf{0} & Z_i^\varepsilon \end{bmatrix} Y_i, \quad \text{where } Z_i^\varepsilon = \begin{bmatrix} Z_i^\varepsilon[1] & 1 - Z_i^\varepsilon[2] \\ 1 - Z_i^\varepsilon[1] & Z_i^\varepsilon[2] \end{bmatrix}, \end{aligned}$$

simplifying to a block matrix. Note that the first column of Z_i^ε is exactly randomized response applied to group 1, while the second column is randomized response applied to group 2.

So, we can consider privatized data Y^ε to be generated from $\text{Multinom}(n, \boldsymbol{\theta}^\varepsilon(\pi, p_1, p_2))$,

$$\boldsymbol{\theta}^\varepsilon(\pi, p_1, p_2) = \begin{pmatrix} \frac{e^\varepsilon}{e^\varepsilon+1}(\pi p_1) + \frac{1}{e^\varepsilon+1}((1-\pi)p_2) \\ \frac{e^\varepsilon}{e^\varepsilon+1}((1-\pi)p_2) + \frac{1}{e^\varepsilon+1}(\pi p_1) \\ \frac{e^\varepsilon}{e^\varepsilon+1}(\pi(1-p_1)) + \frac{1}{e^\varepsilon+1}((1-\pi)(1-p_2)) \\ \frac{e^\varepsilon}{e^\varepsilon+1}((1-\pi)(1-p_2)) + \frac{1}{e^\varepsilon+1}(\pi(1-p_1)) \end{pmatrix}$$

We then form estimates $\hat{\pi}$ and $\hat{p}_1, \hat{p}_2$ for π and p_1, p_2 from under $\mathtt{H}_0$, $p_1 - p_2 = \Delta$.

$$\hat{p}_2 = \frac{Y^\varepsilon[1,1] + Y^\varepsilon[2,1]}{n} - \hat{\pi}\Delta, \quad \hat{p}_1 = \hat{p}_2 + \Delta, \quad \hat{\pi} = \left(\frac{e^\varepsilon + 1}{e^\varepsilon - 1}\right)\left(\frac{Y[1,1]^\varepsilon + Y[1,2]^\varepsilon}{n} - \frac{1}{e^\varepsilon+1}\right)$$

The private χ^2-statistic $\hat{D}^\varepsilon$ then becomes the following,

$$\hat{D}^\varepsilon(\Delta) = n \cdot \min_{\substack{p_1 - p_2 = \Delta \\ \pi \in (0,1)}} \left\{ (Y^\varepsilon/n - \boldsymbol{\theta}^\varepsilon(\pi, p_1, p_2))^\intercal \, \text{Diag}\left(\boldsymbol{\theta}^\varepsilon(\hat{\pi}, \hat{p}_1, \hat{p}_2)\right)^{-1} (Y^\varepsilon/n - \boldsymbol{\theta}^\varepsilon(\pi, p_1, p_2)) \right\}.$$

One way to achieve valid confidence intervals for the difference $p_1 - p_2$ is to test for multiple values of Δ to see which intervals should be rejected under $\mathtt{H}_0$. That is, we search over the space $\Delta \in [-1, 1]$, with a tolerance level τ, and check whether $\hat{D}(Y; \Delta) \leq \chi^2_{1,1-\alpha}$. As we move from $\Delta = -1$, we will cross a point $\Delta = \Delta^L$ where $\hat{D}(Y; \Delta^L) > \chi^2_{1,1-\alpha}$, yet $\hat{D}(Y; \Delta^L + \tau) \leq \chi^2_{1,1-\alpha}$. This is the left-end point of our confidence interval; we find Δ^R analogously. See the appendix for more results and discussion.

5. Testing Difference in Two Means: t-tests

Typically, one would use the classical t-test to test the difference between two means between samples $\{X_i[j]\}_{i=1}^{n_j} \overset{i.i.d.}{\sim} \mathtt{N}\left(\mu_j, \sigma_j^2\right)$ for $j \in \{1, 2\}$. When introducing privacy, we want to know

whether the t-test can still be used or whether a different test should be used. To use the general χ^2 statistic, we will need to handle real outcomes. First, we form a contingency table for continuous outcomes without discretizing, by considering the moments of the samples (Table 2). We use $W_i \sim \text{Bern}(\pi)$ to determine the group of sample i.

Table 2: Contingency Table for continuous outcomes.

Sample Orders	Group 0	Group 1
0-th	$\sum_{i=1}^n W_i$	$\sum_{i=1}^n (1 - W_i)$
1-st	$\sum_{i=1}^n W_i X_i[1]$	$\sum_{i=1}^n (1 - W_i) X_i[2]$

We write $Z_i^\varepsilon[j] \sim \text{Bern}(\frac{e^\varepsilon}{e^\varepsilon + 1})$ for $i \in [n]$ and $j \in \{1, 2\}$, and form the privatized Y_i^ε,

$$Y_i^\varepsilon = \begin{pmatrix} Z_i^\varepsilon[1] \cdot W_i + (1 - Z_i^\varepsilon[2]) \cdot (1 - W_i) \\ Z_i^\varepsilon[1] \cdot W_i \cdot X_i[1] + (1 - Z_i^\varepsilon[2]) \cdot (1 - W_i) \cdot X_i[2] \\ (1 - Z_i^\varepsilon[1]) \cdot W_i \cdot X_i[1] + Z_i^\varepsilon[2] \cdot (1 - W_i) \cdot X_i[2] \end{pmatrix}.$$

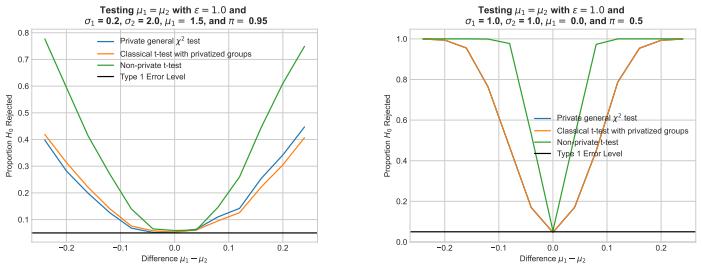

Figure 1: Power of t-test without modification on privatized data and the χ^2 test (13) with $\epsilon = 1.0$ and $n = 10000$ samples. The left plot has parameters $(\pi, \mu_1, \sigma_1, \sigma_2) = (0.95, 1.5, 2, 0.2)$, while the right plot has $(\pi, \mu_1, \sigma_1, \sigma_2) = (0.5, 0, 1, 1)$.

From here, we can derive the following test statistic, which we compare to a χ_1^2.

$$D^\varepsilon = \min_{\substack{\pi \in (0,1), \\ \mu_1, \mu_2 : \mu_1 = \mu_2 + \Delta}} \left\{ \left(Y^\varepsilon - \boldsymbol{\theta}^\varepsilon(\pi, \mu_1, \mu_2)\right)^\intercal C(\hat{\pi}, \hat{\mu_1}, \hat{\mu_2}, \hat{\sigma_1}, \hat{\sigma_2})^{-1} \left(Y^\varepsilon - \boldsymbol{\theta}^\varepsilon(\pi, \mu_1, \mu_2)\right) \right\} \quad (4)$$

Figure 5 gives results for testing the difference in means, comparing our test with the naive approach of using the t-test as if no privacy has been introduced, as well as the baseline t-test on non-privatized groups. We chose a fairly extreme setting of parameters in the data distribution to show that the χ^2 approach can achieve similar power to the classical t-test approach, while in more symmetric settings, i.e. $\sigma_1 \approx \sigma_2$ and $\pi \approx 0.5$, they perform similarly. We apply the general rule of thumb[3] that if at any point our estimate group size $\hat{\pi} \cdot n$ in either group is less than 5, we simply return a zero statistic and hence fail to reject.

3. There are similar rules of thumb in traditional statistical tests, where tests are inconclusive if a group size is too small.

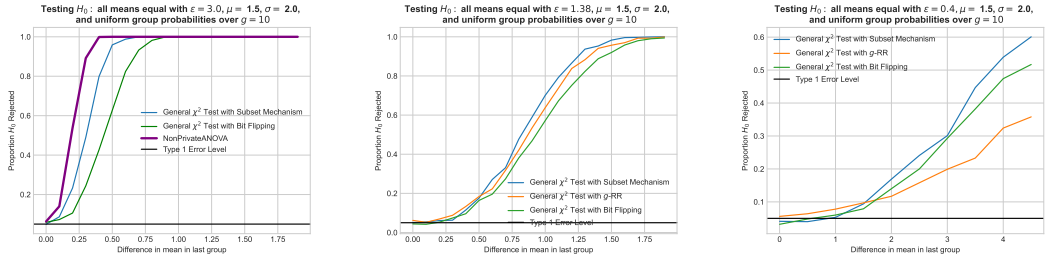

Figure 2: Comparing power curves for ANOVA with $n = 10000$ samples and various ϵ.

6. Testing Differences in Means across Several Groups

We now consider testing whether there is a difference in means across $g > 2$ groups. That is, let $X_i[j]$ be i.i.d. $\mathtt{N}\left(\mu_j, \sigma_j^2\right)$ for $i \in [n]$ with $j \in [g]$; we test $H_0 : \mu_1 = \mu_2 = \cdots \mu_g$. In this case, we would perform a one-way ANOVA test, which makes the assumption that all σ_j are equal, i.e. $\sigma_j = \sigma$ for all $j \in [g]$. The one-way ANOVA test compares H_0 to the alternative hypothesis H_1: not all means are equal. Observe that the one-way ANOVA does not allow us to conclude specifically *which* mean or means may be different across the groups, and would usually be followed up with either a set of t-tests or a method like Tukey's method, or a reasonable conclusion from clear deviations observed (Rice, 2006).

It is straightforward to fit this hypothesis test to our generalized χ^2 test framework by considering the variable W_i, which will determine the group that sample i is in. That is, $W_i \sim \text{Multinomial}(n, \pi)$ for $i \in [n]$ and $\pi \in [0,1]^g$ is a probability vector. We omit the details here (available in the appendix). We present results in Figure 2 with $\mu = \mu_j$ for all $j \in \{1, \ldots, g\}$ with $g = 10$ but we vary the mean in the last coordinate μ_{10} to see the fraction of times we reject the null. We compare the tests for g-randomized response, bit flipping, and the subset mechanism, as well as using the one-way ANOVA test as is on data that has been privatized with g-randomized response. We see that the subset mechanism outperforms the various tests at the different privacy levels.

7. Application to A/B Testing

To close, we apply our general approach to the A/B testing setting, specifically to test whether the difference in means between two groups has remained the same or changed in an A/B test. Going back to our motivation, this extension has huge potential as a practical solution for differential privacy in analyzing A/B tests over sensitive groups. With this method, companies can add privacy to data describing demographic information, and then proceed to test hypotheses about different groups and outcomes under many different A/B tests, without incurring privacy loss. As companies move to ask questions about different groups' experiences on their platforms, testing without statistical bias and with high power is crucial for making the best decisions.

We give confidence intervals for the difference between means across treatment and control in Figure 3 using the general χ^2 test statistic and the unmodified t-test statistic. Observe

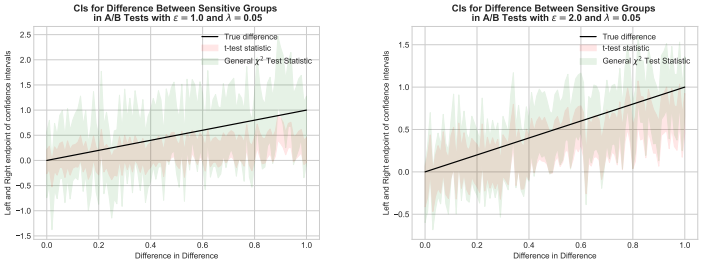

Figure 3: Confidence intervals for the difference in means across groups between treatment and control; $H_0 : \mu_{1,t} - \mu_{2,t} = \mu_{1,c} - \mu_{2,c} + \Delta$. We compare t-tests and general χ^2 tests on privatized groups with treatment probability $\lambda = 0.05$ and $\varepsilon \in \{1, 2\}$.

that the t-test does not cover the true difference in difference in the confidence intervals, as compared to the general χ^2 test, which does not suffer from the same downward bias.

8. Conclusion

We introduced the local group DP definition as a less restrictive privacy model than the local DP model, while ensuring the group that each member belongs to remains private no matter the number of tests that are conducted. We considered the binary outcome setting, before extending the general χ^2 framework to test differences in means, one-way ANOVA, and an application to A/B testing. Our results show that these general χ^2 tests can also be used to compute valid confidence intervals for the true difference in proportions and means even when group membership is privatized. Furthermore, the power of the tests over multiple groups can significantly benefit by using the subset mechanism and the corresponding test, rather than the traditional tests that do not account for privacy.

A limitation of this work is that if a new group is to be added to a set of existing groups, increasing the total number of groups from g to g', then it is not clear how to take samples that are privatized over the smaller set of g groups. The question then becomes, how do we design statistical tests that combines datasets where some samples are privatized over g groups while others are privatized over $g' > g$ groups? Answering these questions will help researchers to design tests that solve practical problems in hypothesis testing under DP.

Acknowledgments

The authors would like to thank Parvez Ahammad, YinYin Yu, and Rahul Tandra for their support, and Adrian Cardoso and Kenneth Tay for thoughtful feedback and reviews. We also thank participants in the Fields Institute Workshop on Differential Privacy and Statistical Data Analysis for their feedback and discussions.

References

Jayadev Acharya, Clement Canonne, Cody Freitag, and Himanshu Tyagi. Test without trust: Optimal locally private distribution testing. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2067–2076. PMLR, 16–18 Apr 2019a. URL https://proceedings.mlr.press/v89/acharya19b.html.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1120–1129. PMLR, 16–18 Apr 2019b. URL https://proceedings.mlr.press/v89/acharya19a.html.

Jordan Alexander Awan and Aleksandra Slavkovic. Differentially private inference for binomial data. *Journal of Privacy and Confidentiality*, 10(1), Jan. 2020. doi: 10.29012/jpc.725. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/725.

Siddhartha Basu, Ruthie Berman, Adam Bloomston, John Campbell, Anne Diaz, Nanako Era, Benjamin Evans, Sukhada Palkar, and Skyler Wharton. Measuring discrepancies in airbnb guest acceptance rates using anonymized demographic data, 2022. URL https://arxiv.org/abs/2204.12001.

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning, 2019. URL https://arxiv.org/pdf/1812.00984.

Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14475–14485. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a684eceee76fc522773286a895bc8436-Paper.pdf.

Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'IT: Private and sample efficient identity testing. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 635–644. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/cai17a.html.

Zachary Campbell, Andrew Bray, Anna M. Ritz, and Adam Groce. Differentially private anova testing. *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 281–285, 2018.

Clément L. Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 310–321, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059. doi: 10.1145/3313276.3316336. URL https://doi.org/10.1145/3313276.3316336.

Clément L Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, and Lydia Zakynthinou. Private identity testing for high-dimensional distributions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10099–10111. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/72b32a1f754ba1c09b3695e0cb6cde7f-Paper.pdf.

Christian Covington, Xi He, James Honaker, and Gautam Kamath. Unbiased statistical estimation and valid confidence intervals under differential privacy, 2021. URL https://arxiv.org/abs/2110.14465.

Bolin Ding, Harsha Nori, Paul Li, and Joshua Allen. Comparing population means under local differential privacy: With significance and power. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, pages 202–210, 2003.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation (with discussion). *Journal of the American Statistical Association*, 113(521):182–215, 2018.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, page 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540327312. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2957-6. doi: 10.1145/2660267.2660348. URL http://doi.acm.org/10.1145/2660267.2660348.

Alexandre V. Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, pages 211–222, 2003.

Vitaly Feldman, Jelani Nelson, Huy Lê Nguyen, and Kunal Talwar. Private frequency estimation via projective geometry, 2022. URL https://arxiv.org/abs/2203.00194.

T.S. Ferguson. *A Course in Large Sample Theory*. Chapman & Hall Texts in Statistical Science Series. Taylor & Francis, 1996. ISBN 9780412043710.

Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1626–1635. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gaboardi18a.html.

Marco Gaboardi, Hyun Lim, Ryan Rogers, and Salil Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2111–2120, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/rogers16.html.

Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private mean estimation: z-test and tight confidence intervals. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2545–2554. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/gaboardi19a.html.

Matthew Joseph, Janardhan Kulkarni, Jieming Mao, and Steven Z. Wu. Locally private gaussian estimation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlche Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/a588a6199feff5ba48402883d9b72700-Paper.pdf.

Marc Juarez and Aleksandra Korolova. "you can't fix what you can't measure": Privately measuring demographic performance disparities in federated learning, 2022. URL https://arxiv.org/abs/2206.12183.

Kazuya Kakizaki, Kazuto Fukuchi, and Jun Sakuma. Differentially private chi-squared test by unit circle mechanism. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1761–1770. JMLR.org, 2017.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals, 2017. URL https://arxiv.org/abs/1711.03908.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Daniel Kifer and Ryan Rogers. A New Class of Private Chi-Square Hypothesis Tests. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 991–1000. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/rogers17a.html.

Joseph Near and David Darais. Differentially private synthetic data, 2021. URL https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data.

John A. Rice. *Mathematical Statistics and Data Analysis.* Belmont, CA: Duxbury Press., third edition, 2006.

Or Sheffet. Differentially private ordinary least squares. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3105–3114. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sheffet17a.html.

Or Sheffet. Locally private hypothesis testing. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4605–4614. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/sheffet18a.html.

Marika Swanberg, Ira Globus-Harris, Iris Griffith, Anna Ritz, Adam Groce, and Andrew Bray. Improved differentially private analysis of variance. *Proceedings on Privacy Enhancing Technologies*, 2019(3):310–330, 2019. doi: doi:10.2478/popets-2019-0049. URL https://doi.org/10.2478/popets-2019-0049.

Caroline Uhler, Aleksandra B. Slavkovic, and Stephen E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), Aug. 2013. doi: 10.29012/jpc.v5i1.629. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/629.

Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143, 2009. doi: 10.1109/ICDMW.2009.52.

Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data, 2015. URL https://arxiv.org/abs/1511.03376.

Yue Wang, Daniel Kifer, and Jaewoo Lee. Differentially private confidence intervals for empirical risk minimization. *Journal of Privacy and Confidentiality*, 9(1), Mar. 2019. doi: 10.29012/jpc.660. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/660.

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

Ian Waudby-Smith, Zhiwei Steven Wu, and Aaditya Ramdas. Locally private nonparametric confidence intervals and sequences, 2022. URL https://arxiv.org/abs/2202.08728.

Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under local differential privacy. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 759–763, 2017. doi: 10.1109/ISIT.2017.8006630.

Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2014.01.008. URL https://www.sciencedirect.com/science/article/pii/S1532046414000100. Special Issue on Informatics Methods in Medical Privacy.