A New Class of Private Chi-Square Tests

Daniel Kifer
Penn State University and the U.S. Census Bureau

Ryan Rogers
University of Pennsylvania

Abstract

In this paper, we develop new test statistics for private hypothesis testing. These statistics are designed specifically so that their asymptotic distributions, after accounting for noise added for privacy concerns, match the asymptotics of the classical (non-private) chi-square tests for testing if the multinomial data parameters lie in lower dimensional manifolds (examples include goodness of fit and independence testing). Empirically, these new test statistics outperform prior work, which focused on noisy versions of existing statistics.

1 Introduction

In 2008, Homer et al. [13] published a proof-of-concept attack showing that participation of individuals in scientific studies can be inferred from aggregate data typically published in genome-wide association studies (GWAS). Since then, there has been renewed interest in protecting confidentiality of participants in scientific data [14, 21, 25, 19] using privacy definitions such as differential privacy and its variations [7, 6, 3, 5].

An important tool in statistical inference is *hypothesis testing*, a general framework for determining whether a given model – called the null hypothesis H_0 – of a population should be rejected based on a sample from the population. One of the main benefits of hypothesis testing is that it gives a way to control the probability of false discovery or Type I error – falsely concluding that a model should be rejected when it is indeed true. Type II error is the probability of failing to reject H_0 when it is false. Typically, scientists want a test that guarantees a pre-specified Type I error (say 0.05) and has high *power* – complement of Type II error.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

The standard approach to hypothesis testing is to (1) estimate the model parameters from the data, (2) compute a *test statistic* T (a function of the data and the model parameters), (3) determine the (asymptotic) distribution of T under the assumption that the model generated the data, (4) compute the *p-value* (Type I error) as the probability of T being more extreme than the realized value from the data.[1]

Our main contribution is a general template for creating test statistics involving categorical data. Empirically, they improve on the power of previous work on differentially private hypothesis testing [12, 23], while maintaining at most some given Type I error. Our approach is to select certain properties of non-private hypothesis tests (e.g., their asymptotic distributions) and then build new test statistics that match these properties when Gaussian noise is added (e.g., to achieve *concentrated differential privacy* [5, 3] or *(approximate) differential privacy* [6]). Although the test statistics are designed with Gaussian noise in mind, other noise distributions can be applied, e.g. Laplace.[2]

We point out that implications of this work extend beyond simply alleviating privacy concerns. In *adaptive data analysis*, data may be reused for multiple analyses, each of which may depend on previous outcomes thus potentially overfitting. This problem was recently studied in the computer science literature by Dwork et al. [8], who show that differential privacy can help prevent overfitting despite reusing data. There have been several follow up works [9, 4, 1] that improve and extend the connection between differential privacy and generalization guarantees in adaptive data analysis. Specifically, [17] deals with *post-selection hypothesis testing* where they can ensure a bound on Type I error even for several adaptively chosen tests, as long as each test is differentially private.

[1] For one-sided tests, the p-value is the probability of seeing the computed statistic or anything larger under H_0.

[2] If we use Laplace noise instead, we cannot match properties like the asymptotic distribution of the non-private statistics, but the new test statistics still empirically improve the power of the tests. Due to space issues, these experiments appear in the supplementary materials.

We discuss related work in Section 2, provide background information about privacy in Section 3, present our extension of minimum chi-square theory in Section 4 and show how it can be applied to goodness of fit (Section 5) and independence testing (Section 6). Experiments appear in these latter two sections. We present conclusions in Section 7.

Please note that, due to space constraints, proofs can be found in the supplementary file.

2 Related Work

One of the first works to study the asymptotic distributions of statistics that use differentially private data came from Wasserman and Zhou [24]. Smith [20] then showed that for a large family of statistics, there is a corresponding differentially private statistic that shares the same asymptotic distribution as the original statistic. However, these results do not ensure that statistically valid conclusions are made for finite samples. It is then the goal of a recent line of work to develop statistical inference tools that give valid conclusions for even reasonably sized datasets.

Prior work on private statistical inference for categorical data can be roughly grouped into two main approaches. The first group adds appropriately scaled noise to the sampled data (or histogram of data) to ensure differential privacy and uses existing classical hypothesis tests, disregarding the additional noise distribution [14]. This approach is based on the argument that the impact of the noise becomes small as the sample size grows large. Along these lines, [22] studies how many more samples would be needed before the test with additional noise recovers the same level of power as the original test on the actual data. However, as pointed out in [11, 15, 16, 12], even for moderately sized datasets, the impact of privacy noise is non-negligible and therefore such an approach can lead to misleading and statistically invalid results, specifically with much higher Type I error than the prescribed amount.

The second group of work consists of tests that focus on adjusting step (3) in the standard approach to hypothesis testing given in the introduction. That is, these tests use the same statistic in the classical hypothesis tests (without noise) and after making the statistic differentially private, they determine the resulting modified asymptotic distribution of the private statistic [21, 25, 23, 12]. Unfortunately, the resulting asymptotic distribution cannot be written analytically, and so Monte Carlo (MC) simulations or numerical approximations are commonly used to determine at what point to reject the null hypothesis.

We focus on a different technique from these two different approaches, namely modifying step (2) in our outline of hypothesis testing. Thus, we consider transforming the test statistic itself so that the resulting distribution is close to the original asymptotic distribution when additional Gaussian noise is used. If the noise is non-Gaussian, then this is followed by another step that appropriately adjusts the asymptotic distribution. The idea of modifying the test statistic for *regression coefficients* to obtain a t-statistic in ordinary least squares has also been considered in [18].

3 Privacy Preliminaries

Formal privacy definitions can be used to protect scientific data with the careful injection of noise. Hypothesis testing must then properly account for this noise to avoid generating false conclusions. We briefly discuss examples of privacy definitions that can be used and then elaborate on how to add noise to satisfy those definitions.

Let $\mathcal{X}$ be an arbitrary domain for records. We define two datasets $\mathbf{x} = (x_1, \cdots, x_n), \mathbf{x}' = (x'_1, \cdots, x'_n) \in \mathcal{X}^n$ to be *neighboring* if they differ in at most one entry, i.e. there is some $i \in [n]$ where $x_i \neq x'_i$, but $x_j = x'_j$ for all $j \neq i$. We now define differential privacy (DP)[7, 6].

Definition 3.1 (Differential Privacy)**.** A randomized algoirthm $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ is (ϵ, δ)-DP if for all neighboring datasets $\mathbf{x}, \mathbf{x}'$ and each subset of outcomes $S \subseteq \mathcal{O}$,

$$\Pr[\mathcal{M}(\mathbf{x}) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(\mathbf{x}') \in S] + \delta.$$

If $\delta = 0$, we simply say $\mathcal{M}$ is ϵ-DP.

In this work, we focus on a recent variation of differential privacy, called *zero concentrated differential privacy* (zCDP) [3]; extensions of our work to ϵ-DP can be found in the supplementary material.

Definition 3.2 (zCDP)**.** A randomized algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ is ρ-zCDP if for all neighboring datasets $\mathbf{x}, \mathbf{x}'$ and all $\alpha \in (1, \infty)$, we have $D_\alpha(\mathcal{M}(\mathbf{x}) || \mathcal{M}(\mathbf{x}')) \leq \rho\alpha$, where, for distributions P and Q, the Renyi divergence $D_\alpha(P||Q)$ is $\frac{1}{\alpha - 1} \log \left(\int P(y)^\alpha Q(y)^{1-\alpha} \, dy \right)$.

zCDP lies between *pure*-DP where $\delta = 0$ and *approximate*-DP (where δ may be positive):

Theorem 3.3 ([3])**.** *If $\mathcal{M}$ is ϵ-DP, then $\mathcal{M}$ is $\frac{\epsilon^2}{2}$-zCDP. Further, if $\mathcal{M}$ is ρ-zCDP then $\mathcal{M}$ is $(\rho + 2\sqrt{\rho \ln(1/\delta)}, \delta)$-DP for every $\delta > 0$.*

The following property is useful because it ensures the privacy of the dataset no matter what an adversary does with the output of a zCDP algorithm.

Theorem 3.4 (Post Processing [3])**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ and $g : \mathcal{O} \to \mathcal{O}'$ be randomized algorithms. If $\mathcal{M}$ is ρ-zCDP then $g \circ \mathcal{M} : \mathcal{X}^n \to \mathcal{O}'$ is ρ-zCDP.*

We can privately release a function $f : \mathcal{X}^n \to \mathbb{R}^d$ of the data using the Gaussian Mechanism $\mathcal{M}_{\text{Gauss}}$ [6]. $\mathcal{M}_{\text{Gauss}}$ first computes the *global sensitivity* of f, which is defined as $\Delta_p(f) = \max_{\text{neighboring } \mathbf{x},\mathbf{x}' \in \mathcal{X}^n} \{||f(\mathbf{x}) - f(\mathbf{x}')||_p\}$. and then generates a noisy version of f as follows (here $\sigma = \Delta_2(f)/\sqrt{2\rho}$):

$$\mathcal{M}_{\text{Gauss}}(\mathbf{x}) \sim N(f(\mathbf{x}), \sigma^2 \cdot I_d). \tag{1}$$

Theorem 3.5 ([3]). *For a function $f : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian mechanism $\mathcal{M}_{Gauss}$ from (1) is ρ-zCDP.*

For this work we will be considering categorical data. That is, we assume the domain $\mathcal{X}$ has been partitioned into d buckets or outcomes and the function $f : \mathcal{X}^n \to \mathbb{R}^d$ returns a histogram counting how many records are in each bucket. Our test statistics will only depend on this histogram. Since neighboring datasets $\mathbf{x}, \mathbf{x}'$ of size n differ on only one entry, their corresponding histograms differ by ± 1 in exactly two buckets. Hence, we will say that two histograms are neighboring if they differ in at most two entries by at most 1. In this case, $\Delta_2(f) = \sqrt{2}$. To preserve privacy, we will add noise to the corresponding histogram $X = (X_1, \cdots, X_d)$ of our original dataset to get $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_d)$. We perform hypothesis testing on this noisy histogram $\tilde{X}$. By Theorem 3.4, we know that each of our hypothesis tests will be ρ-zCDP as long as we add Gaussian noise with variance $1/\rho$ to each count in X.

4 General Chi-Square Tests

In the non-private setting, a chi-square test involves a histogram X and a model H_0 that produces expected counts $\overline{X}$ over the d buckets. In general, H_0 will have $k < d$ parameters and will estimate the parameters from X. The chi-square test statistic is defined as $\mathrm{T}_{\text{chi}} = \sum_{i=1}^{d} (X_i - \overline{X}_i)^2 / \overline{X}_i$. If the data were generated from H_0 and if k parameters had to be estimated, then the asymptotic distribution of T_{chi} is χ^2_{d-k-1}, a chi-square random variable with $d - k - 1$ degrees of freedom. This is the property we want our statistics to have when they are computed from the noisy histogram $\tilde{X}$ instead of X. Note that in the classical chi-square tests (e.g. Pearson independence test), the statistic T_{chi} is computed and if it is larger than the $1 - \alpha$ percentile of χ^2_{d-k-1}, then the model is rejected.

The above facts are part of a more general *minimum chi-square asymptotic theory* [10], which we overview in Section 4.2. However, we first explain the differences between private and non-private asymptotics [23, 12].

4.1 Private Asymptotics

In non-private statistics, a function of n data records is considered a random variable, and non-private asymptotics considers this distribution as $n \to \infty$. In private asymptotics, there is another quantity σ_n^2, the variance of the added noise.

In the *classical private regime*, one studies what happens as $n/\sigma_n^2 \to \infty$; i.e., when the variance due to privacy is insignificant compared to sampling variance in the data (i.e. $O(n)$). In practice, asymptotic distributions derived under this regime result in unreliable hypothesis tests because privacy noise is significant [21].

In the *variance-aware private regime*, one studies what happens as $n/\sigma_n^2 \to$ constant as $n \to \infty$; that is, when the variance due to privacy is proportional to sampling variance. In practice, asymptotic distributions derived under this regime result in hypothesis tests with reliable Type I error (i.e. the p-values they generate are accurate) [12, 23]. From now on, we will be using the variance-aware privacy regime.[3]

4.2 Minimum Chi-Square Theory

In this section, we present important results about *minimum chi-square theory*. The discussion is based largely on [10] (Chapter 23). Our work relies on this theory to construct new private test statistics in Sections 5 and 6 whose asymptotic behavior matches the non-private asymptotic behavior of the classical chi-square test.

We consider a sequence of d-dimensional random vectors $V^{(n)}$ for $n \geq 1$ (e.g. the data histogram). The parameter space Θ is a non-empty open subset of $\mathbb{R}^k$, where $k \leq d$. The model A maps a k-dimensional parameter $\theta \in \Theta$ into a d-dimensional vector (e.g., the expected value of $V^{(n)}$), hence it maps Θ to a subset of a k-dimensional manifold in d-dimensional space.

In this abstract setting, the null hypothesis is that there exists a $\theta^0 \in \Theta$ such that:[4]

$$\sqrt{n}\left(V^{(n)} - A(\theta^0)\right) \xrightarrow{D} N(0, C(\theta^0)) \tag{2}$$

where $C(\theta) \in \mathbb{R}^{d \times d}$ is a covariance matrix. Intuitively, Equation 2 says that the Central Limit Theorem can be applied for θ^0.

We measure the distance between $V^{(n)}$ and $A(\theta)$ with a test statistic given by the following quadratic form:

$$D^{(n)}(\theta) = n\left(V^{(n)} - A(\theta)\right)^{\intercal} M(\theta) \left(V^{(n)} - A(\theta)\right) \tag{3}$$

[3] Note that taking n and σ_n^2 to infinity is just a mathematical tool for simplifying expressions while mathematically keeping privacy noise variance proportional to the data variance; it does not mean that the amount of actual noise added to the data depends on the data size.

[4] Here $\xrightarrow{D}$ means convergence in distribution, as in the Central Limit Theorem [10].

where $M(\theta) \in \mathbb{R}^{d\times d}$ is a symmetric positive-semidefinite matrix; different choices of M will result in different test statistics. We make the following standard assumptions about $A(\theta)$ and $M(\theta)$.

Assumption 4.1. *For all $\theta \in \Theta$, we have: 1) $A(\theta)$ is bicontinuous,[5] 2) $A(\theta)$ has continuous first partial derivatives, which we denote as $\dot{A}(\theta)$ with full rank k, 3) $M(\theta)$ is continuous in θ and there exists an $\eta > 0$ such that $M(\theta) - \eta I_d$ is positive definite in an open neighborhood of θ^0.*

If θ^0 is known, then we show in the supplementary file that setting $M(\theta) = C(\theta)^{-1}$ in (3) then $D^{(n)}(\theta^0)$ converges in distribution to χ^2_d. However, as we show in Section 5, this can be a sub-optimal choice of M.

When θ^0 is not known, we need to estimate a good parameter $\widehat{\theta}^{(n)}$ to plug into (3). One approach is to set $\widehat{\theta}^{(n)} = \arg\min_{\theta\in\Theta} D^{(n)}(\theta)$. However, this can be a difficult optimization. If there is a rough estimate of θ^0 based on the data, call it $\phi(V^{(n)})$, and if it converges in probability to θ^0 (i.e. $\phi(V^{(n)}) \xrightarrow{P} \theta^0$ as $n \to \infty$), then we can plug it into the middle matrix to get:

$$\widehat{D}^{(n)}(\theta) = n\left(V^{(n)} - A(\theta)\right)^\intercal M(\phi(V^{(n)}))\left(V^{(n)} - A(\theta)\right). \tag{4}$$

and then set our estimator $\widehat{\theta}^{(n)} = \arg\min_{\theta\in\Theta} \widehat{D}^{(n)}(\theta)$. The test statistic becomes $\widehat{D}^{(n)}(\widehat{\theta}^{(n)})$ and the following theorems describe its asymptotic properties under the null hypothesis. We use the shorthand $A = A(\theta^0)$, $M = M(\theta^0)$, and $C = C(\theta^0)$.

Theorem 4.2. *Let $\widehat{\theta}^{(n)} = \operatorname{argmin}_{\theta\in\Theta} \widehat{D}^{(n)}(\theta)$. Given Assumption 4.1 and (2), we have $\sqrt{n}(\widehat{\theta}^{(n)} - \theta^0) \xrightarrow{D} N(0, \Psi)$ where θ^0 is the true parameter and $\Psi = \left(\dot{A}^\intercal M \dot{A}\right)^{-1} \dot{A}^\intercal M C M \dot{A} \left(\dot{A}^\intercal M \dot{A}\right)^{-1}$.*

We then state the following result using a slight modification of Theorem 24 in [10].

Theorem 4.3. *Let ν be the rank of $C(\theta_0)$. If Assumption 4.1 and (2) hold, and, for all $\theta \in \Theta$, $C(\theta)M(\theta)C(\theta) = C(\theta)$ and $C(\theta)M(\theta)\dot{A}(\theta) = \dot{A}(\theta)$ then for $\widehat{\theta}^{(n)}$ given in Theorem 4.2 and $\widehat{D}^{(n)}(\theta)$ given in (4) we have: $\widehat{D}^{(n)}\left(\widehat{\theta}^{(n)}\right) \xrightarrow{D} \chi^2_{\nu-k}$*

5 Private Goodness of Fit Tests

As a warmup, we will first cover goodness of fit testing where the null hypothesis is simply testing whether the underlying unknown parameter is equal to a particular value. We consider categorical data $X^{(n)} = \left(X_1^{(n)}, \cdots, X_d^{(n)}\right) \sim \text{Multinomial}(n, \mathbf{p})$ where $\mathbf{p} = (p_1, \cdots, p_d)$ is some probability vector over the d outcomes. We want to test the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}^0$, where each component of $\mathbf{p}^0$ is positive, but we want to do so in a private way. We then have the following classical result [2].

Lemma 5.1. *Under the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}^0$, $X^{(n)}/n$ is asymptotically normal $\sqrt{n}\left(\frac{X^{(n)}}{n} - \mathbf{p}^0\right) \xrightarrow{D} N(0, \Sigma)$ where Σ has rank $d-1$ and can be written as*

$$\Sigma \overset{\text{defn}}{=} \texttt{Diag}(\mathbf{p}^0) - \mathbf{p}^0(\mathbf{p}^0)^\intercal. \tag{5}$$

5.1 Unprojected Private Test Statistic

To preserve ρ-zCDP, we will add appropriately scaled Gaussian noise to each component of the histogram $X^{(n)}$. We then define the zCDP statistic $U_\rho^{(n)} = \left(U_{\rho,1}^{(n)}, \cdots, U_{\rho,d}^{(n)}\right)$ where we write $Z \sim N(0, 1/\rho \cdot I_d)$ and

$$U_\rho^{(n)} \overset{\text{defn}}{=} \sqrt{n}\left(\frac{X^{(n)} + Z}{n} - \mathbf{p}^0\right). \tag{6}$$

We next derive (see proof in supplementary file) the asymptotic distribution of $U_\rho^{(n)}$ under both private asymptotic regimes in Section 4.1 (note that $\sigma^2 = 1/\rho$).

Lemma 5.2. *The random vector $U_{\rho_n}^{(n)}$ from (6) under the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}^0$ has the following asymptotic distribution. If $n\rho_n \to \infty$ then $U_{\rho_n}^{(n)} \xrightarrow{D} N(0, \Sigma)$. Further, if $n\rho_n \to \rho > 0$ then $U_{\rho_n}^{(n)} \xrightarrow{D} N(0, \Sigma_\rho)$ where Σ_ρ has full rank and*

$$\Sigma_\rho \overset{\text{defn}}{=} \Sigma + 1/\rho \cdot I_d. \tag{7}$$

Because Σ_ρ is invertible when the privacy parameter $\rho > 0$, we can create a new statistic based on $U_\rho^{(n)}$ that has a chi-square asymptotic distribution under variance-aware privacy asymptotics.

Theorem 5.3. *Let $U_{\rho_n}^{(n)}$ be given in (6) for $n\rho_n \to \rho > 0$. If the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}^0$ holds, then for $\Sigma_{n\rho_n}$ given in (7), we have*

$$Q_{\rho_n}^{(n)} \overset{\text{defn}}{=} \left(U_{\rho_n}^{(n)}\right)^\intercal \Sigma_{n\rho_n}^{-1} U_{\rho_n}^{(n)} \xrightarrow{D} \chi^2_d. \tag{8}$$

Note that Σ_ρ is ill-conditioned when ρ is large (data variance overwhelms privacy noise), since Σ is singular and $\Sigma\,\mathbf{1} = \mathbf{0}$. This makes the test statistic unstable. Further, the additional noise adds a degree of freedom to the asymptotic distribution of the original statistic. This additional degree of freedom results in increasing the point in which we reject the null hypothesis, i.e. the critical value. Thus, rejecting an incorrect model becomes harder as we increase the degrees of freedom, and hence decreases power.

[5] i.e. $\theta_j \to \theta \Leftrightarrow A(\theta_j) \to A(\theta)$.

5.2 Projected Private Test Statistic

Given that the test statistic in the previous section depends on a nearly singular matrix, we now derive a new test statistic for the private goodness of fit test. It has the remarkable property that its asymptotic distribution is χ^2_{d-1} under both private asymptotics.

We start with the following observation. In the classical chi-square test, the random variables $\left(\frac{(X_i^{(n)}-np_i^0)}{\sqrt{np_i^0}}\right)_{i=1}^d$ have covariance matrix $I_d - \sqrt{\mathbf{p}^0}\sqrt{\mathbf{p}^0}^\intercal$ under the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}^0$. The classical test essentially uncorrelates these random variables and projects them onto the subspace orthogonal to $\sqrt{\mathbf{p}^0}$. We will use a similar intuition for the privacy-preserving random vector $U_\rho^{(n)}$.

The matrix Σ_ρ in (7) has eigenvector $\mathbf{1}$ with eigenvalue $1/\rho$ – regardless of the true parameters of the data-generating distribution. Hence we think of this direction as pure noise. We therefore project $U_\rho^{(n)}$ onto the space orthogonal to $\mathbf{1}$ (i.e. enforce the constraint that the entries in $U_\rho^{(n)}$ add up to 0, as they would in the noiseless case). We then define the *projected statistic* $\boldsymbol{\mathcal{Q}}_\rho^{(n)}$ as the following where we write the projection matrix $\boldsymbol{P} \stackrel{\text{defn}}{=} I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal$

$$\boldsymbol{\mathcal{Q}}_\rho^{(n)} \stackrel{\text{defn}}{=} \left(U_\rho^{(n)}\right)^\intercal \boldsymbol{P}\Sigma_{n\rho}^{-1}\boldsymbol{P}U_\rho^{(n)}. \tag{9}$$

It will be useful to write out the middle matrix in $\boldsymbol{\mathcal{Q}}_{\rho_n}^{(n)}$ for analyzing its asymptotic distribution.

Lemma 5.4. *For the covariance matrix $\Sigma_{n\rho_n}$ given in (7), we have the following identity when $n\rho_n \to \rho > 0$ $\boldsymbol{P}\Sigma_{n\rho_n}^{-1}\boldsymbol{P} \to \Sigma_\rho^{-1} - \frac{\rho}{d}\cdot\mathbf{1}\mathbf{1}^\intercal$ Further, when $n\rho_n \to \infty$, we have the following $\boldsymbol{P}\Sigma_{n\rho_n}^{-1}\boldsymbol{P} \to \boldsymbol{P}\texttt{Diag}\left(\mathbf{p}^0\right)^{-1}\boldsymbol{P}$*

The projected statistic is asymptotically chi-square distributed in both private asymptotic regimes.

Theorem 5.5. *Let $U_\rho^{(n)}$ be given in (6). The projected statistic $\boldsymbol{\mathcal{Q}}_\rho^{(n)}$ has the following asymptotic distribution for $n\rho_n \to \rho > 0$ and $n\rho_n \to \infty$ (as $n \to \infty$) if the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}^0$ holds: $\boldsymbol{\mathcal{Q}}_{\rho_n}^{(n)} \xrightarrow{D} \chi^2_{d-1}$.*

5.3 Comparison of Statistics

We now want to compare the two private chi-square statistics in (8) and (9) to see which may lead to a larger *power* (i.e. smaller Type II error). The following theorem shows that we can write the unprojected statistic (8) as a combination of both the projected statistic (9) and squared independent Gaussian noise.

Theorem 5.6. *Consider histogram data $X^{(n)}$ that has Gaussian noise $Z \sim N(0, 1/\rho \cdot I_d)$ added to it. For the statistics $Q_\rho^{(n)}$ and $\boldsymbol{\mathcal{Q}}_\rho^{(n)}$ based on the noisy counts given in (8) and (9) respectively, we have $Q_\rho^{(n)} = \boldsymbol{\mathcal{Q}}_\rho^{(n)} + \frac{\rho}{d}\left(\sum_{i=1}^d Z_i\right)^2$. Further, for any fixed data $X^{(n)}$, $\boldsymbol{\mathcal{Q}}_\rho^{(n)}$ is independent of $\left(\sum_{i=1}^d Z_i\right)^2$.*

Algorithm 1 (`zCDP-GOF`) shows how to perform goodness of fit testing with either of these two test statistics, i.e. unprojected (8) or projected (9). We note that our test is zCDP for neighboring histogram datasets due to it being an application of the Gaussian mechanism and Theorem 3.4. Hence:

Theorem 5.7. *`zCDP-GOF`$(\cdot; \rho, \alpha, \mathbf{p}^0)$ is ρ-zCDP.*

Algorithm 1 zCDP Chi-Square Goodness of Fit Test

procedure `zCDP-GOF`$(X^{(n)};\ \rho,\ \alpha,\ H_0 : \mathbf{p} = \mathbf{p}^0)$
Set $\tilde{X}^{(n)} \leftarrow X^{(n)} + Z$ where $Z \sim N(0, 1/\rho \cdot I_d)$.
  For the *unprojected statistic*:
  $\mathrm{T} \leftarrow \frac{1}{n}\left(\tilde{X}^{(n)} - n\mathbf{p}^0\right)^\intercal \Sigma_{n\rho}^{-1}\left(\tilde{X}^{(n)} - n\mathbf{p}^0\right)$
  $t \leftarrow (1-\alpha)$ quantile of χ^2_d
  For the *projected statistic*:
  $\mathrm{T} \leftarrow \frac{1}{n}\left(\tilde{X}^{(n)} - n\mathbf{p}^0\right)^\intercal \boldsymbol{P}\Sigma_{n\rho}^{-1}\boldsymbol{P}\left(\tilde{X}^{(n)} - n\mathbf{p}^0\right)$
  $t \leftarrow (1-\alpha)$ quantile of χ^2_{d-1}
if $\mathrm{T} > t$ **then** Reject

When the null hypothesis is false (i.e., $\mathbf{p} \neq \mathbf{p}^0$), both statistics converge to a non-central chi-square distribution (the analysis can be found in the supplementary file). We then turn to empirical results.

5.4 Experiments for Goodness of Fit Testing

Throughout all of our experiments, we will fix $\alpha = 0.05$ and privacy parameter $\rho = 0.001$. All of our tests are designed to achieve Type I error at most α.[6]

We then empirically check the power of our new tests in `zCDP-GOF` for both the projected and unprojected statistic. Subject to the constraint that our tests achieve Type I error at most α, we seek to maximize *power*, or the probability of rejecting the null hypothesis when a distribution $\mathbf{p}^1 \neq \mathbf{p}^0$, called the *alternate hypothesis*, is true. We expect to see the projected statistic achieve higher power than the unprojected statistic due to Theorem 5.6. Further, the critical value we use for the projected statistic is smaller than the critical value for the unprojected statistic, which might improve the power of the projected statistic.

Here we present a typical experimental scenario. We set the null hypothesis $\mathbf{p}^0 = (1/2, 1/6, 1/6, 1/6)$

[6] Due to space limitations we give the empirical Type I error for various $\mathbf{p}^0$ and n in the supplementary file.

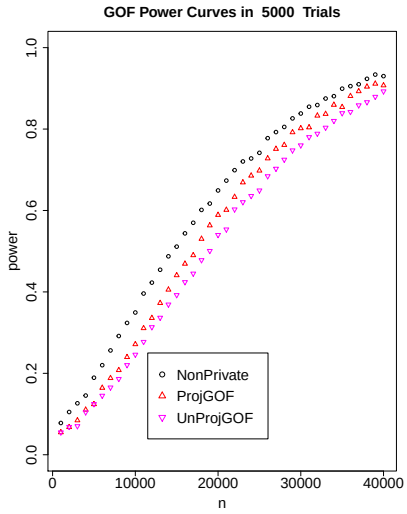

Figure 1: Comparing power between the projected and unprojected statistics in `zCDP-GOF` with the classical non-private test with 5000 trials each, $\rho = 0.001$ and $\alpha = 0.05$.

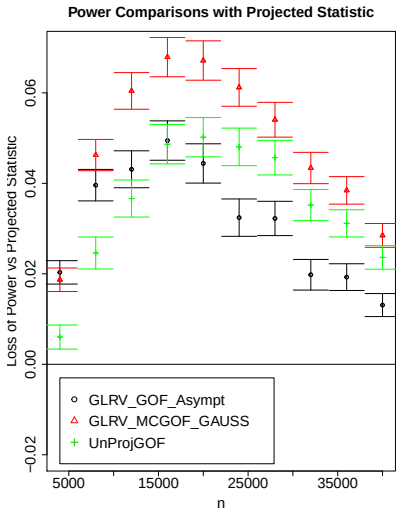

Figure 2: The empirical power loss from using other private goodness of fit tests instead of the projected statistic in `zCDP-GOF` for 100,000 trials, $\rho = 0.001$ and $\alpha = 0.05$.

and alternate hypothesis $\mathbf{p}^1 = \mathbf{p}^0 + 0.01 \cdot (1, -1/3, -1/3, -1/3)$ for various sample sizes (we empirically found this to be a tough alternative hypothesis for our statistics). For each sample size n, we sample 5,000 independent datasets from the alternate hypothesis and test $H_0 : \mathbf{p} = \mathbf{p}^0$ in `zCDP-GOF`. The resulting power plots are in Figure 1 for `zCDP-GOF` from Algorithm 1. We label "NonPrivate" as the classical chi-square goodness of fit test used on the actual data (and thus not private). Further, we write "ProjGOF" as the test from `zCDP-GOF` with the projected statistic whereas "UnProjGOF" uses the unprojected statistic. Clearly, the projected outperforms the unprojected statistic.

We then compare the projected and unprojected statistic in `zCDP-GOF` to prior work in Figure 2. Since the projected statistic outperforms the other tests, we plot the difference in power between the projected statistic and the other tests. We label "GLRV_MCGOF_GAUSS" as the Monte-Carlo (MC) test with Gaussian noise from [12],[7] and "GLRV_GOF_Asympt" as the asymptotics-based test with Gaussian noise from [12, 23]. The error bars show 1.96 times the standard error in the difference of proportions from 100,000 trials, giving a 95% confidence interval.

6 General Chi-Square Private Tests

We now consider the case where the null hypothesis contains many distributions, so that the best fitting distribution must be estimated and used in the test statistics. The data is multinomial $X^{(n)} \sim \text{Multinomial}(n, \mathbf{p}(\theta^0))$ and $\mathbf{p}$ is a function that converts parameters into a d-dimensional multinomial probability vector. The null hypothesis is $H_0 : \theta^0 \in \Theta$; i.e. $\mathbf{p}(\theta^0)$ belongs to a subset of a lower-dimensional manifold. We again use Gaussian noise $Z \sim N(0, 1/\rho \cdot I_d)$ to ensure ρ-zCDP, and we define

$$U_\rho^{(n)}(\theta) \overset{\text{defn}}{=} \sqrt{n}\left(\frac{X^{(n)} + Z}{n} - \mathbf{p}\left(\theta\right)\right). \tag{10}$$

With θ^0 being the unknown true parameter, we are now ready to define our two test statistics in terms of some function $\phi : \mathbb{R}^d \to \mathbb{R}$, such that $\phi(X^{(n)}+Z) \xrightarrow{P} \theta^0$ (recall from Section 4.2 that ϕ is a simple but possibly a suboptimal estimate of the true parameter θ^0 based on the noisy data) and the covariance matrix $\Sigma_\rho(\theta) \overset{\text{defn}}{=} \texttt{Diag}\left(\mathbf{p}(\theta)\right) - \mathbf{p}(\theta)\mathbf{p}(\theta)^\intercal + 1/\rho \cdot I_d$.

We define the *unprojected* statistic $R_\rho^{(n)}(\theta)$ as follows:

$$\widehat{M} \overset{\text{defn}}{=} \left(\Sigma_{n\rho}\left(\phi(X^{(n)} + Z)\right)\right)^{-1}$$
$$R_\rho^{(n)}(\theta) \overset{\text{defn}}{=} U_\rho^{(n)}(\theta)^\intercal \widehat{M} U_\rho^{(n)}(\theta). \tag{11}$$

This is a specialization of (4) in Section 4.2 with the following substitutions: $V^{(n)} = \left(\frac{X^{(n)}+Z}{n}\right)$, $A(\theta) = \mathbf{p}(\theta)$, and $M(\theta) = \left(\Sigma_{n\rho}(\theta)\right)^{-1}$.

For the *projected* statistic $\boldsymbol{\mathcal{R}}_\rho^{(n)}(\theta)$, the corresponding substitutions are $\boldsymbol{P} = I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal$, $V^{(n)} = \boldsymbol{P} \cdot \left(\frac{X^{(n)}+Z}{n}\right)$, $A(\theta) = \boldsymbol{P} \cdot \mathbf{p}(\theta)$, and again $M(\theta) = \left(\Sigma_{n\rho}(\theta)\right)^{-1}$ giving:

$$\boldsymbol{\mathcal{R}}_\rho^{(n)}(\theta) \overset{\text{defn}}{=} U_\rho^{(n)}(\theta)^\intercal \cdot \boldsymbol{P}\widehat{M}\boldsymbol{P} \cdot U_\rho^{(n)}(\theta). \tag{12}$$

[7] We set the the number of MC trials $m = 59$ in these experiments, which guarantees at most 5% Type I error.

We then assume that for both the projected and unprojected statistic Assumption 4.1 holds using their relative vectors $V^{(n)}$, $A(\theta)$, and matrix $M(\theta)$. We now present the asymptotic distribution of both statistics, which is proved using the result in Theorem 4.3 (the full proof is in the supplementary file).

Theorem 6.1. *Under $H_0 : \theta^0 \in \Theta$, the following are true as $n \to \infty$. Setting $\widehat{\theta}^{(n)} = \arg\min_{\theta\in\Theta} R^{(n)}_{\rho_n}(\theta)$ we have $R^{(n)}_{\rho_n}(\widehat{\theta}^{(n)}) \xrightarrow{D} \chi^2_{d-k}$ if $n\rho_n \to \rho > 0$. Furthermore, setting $\widehat{\theta}^{(n)} = \arg\min_{\theta\in\Theta} \boldsymbol{\mathcal{R}}^{(n)}_{\rho_n}(\theta)$ we have $\boldsymbol{\mathcal{R}}^{(n)}_{\rho_n}(\widehat{\theta}^{(n)}) \xrightarrow{D} \chi^2_{d-k-1}$ if $n\rho_n \to \rho$ or $n\rho_n \to \infty$.*

Again, the projected statistic has the same distribution under both private asymptotic regimes and matches the non-private chi-square test asymptotics. We present our more general test zCDP-Min-χ^2 in Algorithm 2. The quick-and-dirty estimator $\phi(\cdot)$ is application-specific (Section 6.1 gives independence testing as an example).[8] Further, for neighboring histogram data, we have the following privacy guarantee.

Theorem 6.2. *zCDP-Min-$\chi^2(\cdot; \rho, \alpha, \phi, \Theta)$ is ρ-zCDP.*

Algorithm 2 zCDP General Chi-Square Test

procedure zCDP-Min-$\chi^2(X^{(n)}$; ρ, α, ϕ, $H_0 : \theta^0 \in \Theta)$

Set $\tilde{X}^{(n)} \leftarrow X^{(n)} + Z$ where $Z \sim N(0, 1/\rho \cdot I_d)$.

Set $\widehat{M} = \Sigma_{n\rho}\left(\phi(\tilde{X}^{(n)})\right)^{-1}$

For the *unprojected statistic*:

$$\mathrm{T}(\theta) = \frac{1}{n}\left(\tilde{X}^{(n)} - n\mathbf{p}(\theta)\right)^{\intercal} \widehat{M}\left(\tilde{X}^{(n)} - n\mathbf{p}(\theta)\right)$$

Set $\widehat{\theta}^{(n)} = \mathrm{argmin}_{\theta\in\Theta}\,\mathrm{T}(\theta)$

$t \leftarrow (1-\alpha)$ quantile of χ^2_{d-k}

For the *projected statistic*:

$$\mathrm{T}(\theta) = \frac{1}{n}\left(\tilde{X}^{(n)} - n\mathbf{p}(\theta)\right)^{\intercal} \boldsymbol{P}\widehat{M}\boldsymbol{P}\left(\tilde{X}^{(n)} - n\mathbf{p}(\theta)\right)$$

Set $\widehat{\theta}^{(n)} = \mathrm{argmin}_{\theta\in\Theta}\,\mathrm{T}(\theta)$

$t \leftarrow (1-\alpha)$ quantile of χ^2_{d-k-1}

if $\mathrm{T}(\widehat{\theta}^{(n)}) > t$ **then** Reject

6.1 Application - Independence Test

We showcase our general chi-square test zCDP-Min-χ^2 by giving results for independence testing. Conceptually, it is convenient to think of the data histogram as an $r \times c$ table, with $\mathbf{p}_{i,j}$ being the probability a person is in the bucket in row i and column j. We then consider two multinomial random variables $Y \sim \text{Multinomial}(1, \pi^{(1)})$ for $\pi^{(1)} \in \mathbb{R}^r$ (the marginal row probability vector) and $Y' \sim \text{Multinomial}(1, \pi^{(2)})$ for $\pi^{(2)} \in \mathbb{R}^c$ (the marginal column probability vector). Under the null hypothesis of independence between Y and Y', $\mathbf{p}_{i,j} = \pi^{(1)}_i \pi^{(2)}_j$. Generally, we write the probabilities as $\mathbf{p}(\pi^{(1)}, \pi^{(2)}) = \pi^{(1)}\left(\pi^{(2)}\right)^{\intercal}$ so that $X^{(n)} \sim \text{Multinomial}\left(n, \mathbf{p}(\pi^{(1)}, \pi^{(2)})\right)$. Thus we have the underlying parameter vector $\theta^0 = \left(\pi^{(1)}_1, \cdots, \pi^{(1)}_{r-1}, \pi^{(2)}_1, \cdots, \pi^{(2)}_{c-1}\right)$ - we do not need the last component of $\pi^{(1)}$ or $\pi^{(2)}$ because we know that each must sum to 1. Also, we have $d = rc$ and $k = (r-1) + (c-1)$ in this case. We want to test whether Y is independent of Y'. For our data, we are given a collection of n independent trials of Y and Y'. We then count the number of joint outcomes in a contingency table given in Table 1. Each cell in the contingency table contains element $X^{(n)}_{i,j}$ that gives the number of occurrences of $Y_i = 1$ and $Y'_j = 1$. Since our test statistics notationally treat the data as a vector, when needed, we convert $X^{(n)}$ to a vector that goes from left to right along each row of the table.

Table 1: Contingency Table.

$Y \;\backslash\; Y'$	1	2	$\cdots$	c	Marginals
1	$X^{(n)}_{1,1}$	$X^{(n)}_{1,2}$	$\cdots$	$X^{(n)}_{1,c}$	$X^{(n)}_{1,\cdot}$
2	$X^{(n)}_{2,1}$	$X^{(n)}_{2,2}$	$\cdots$	$X^{(n)}_{2,c}$	$X^{(n)}_{2,\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$X^{(n)}_{r,1}$	$X^{(n)}_{r,2}$	$\cdots$	$X^{(n)}_{r,c}$	$X^{(n)}_{r,\cdot}$
Marginals	$X^{(n)}_{\cdot,1}$	$X^{(n)}_{\cdot,2}$	$\cdots$	$X^{(n)}_{\cdot,c}$	n

In order to compute the statistic $R^{(n)}_{\rho}(\widehat{\theta}^{(n)})$ or $\boldsymbol{\mathcal{R}}^{(n)}_{\rho}(\widehat{\theta}^{(n)})$ in zCDP-Min-χ^2, we need to find a quick-and-dirty estimator $\phi(X^{(n)} + Z)$ that converges in probability to $\mathbf{p}\left(\pi^{(1)}, \pi^{(2)}\right)$ as $n \to \infty$. We will use the estimator for the unknown probability vector based on the marginals of the table with noisy counts, so that for naïve estimates $\tilde{\pi}^{(1)}_i = \frac{X^{(n)}_{i,\cdot} + Z_{i,\cdot}}{\tilde{n}}$, $\tilde{\pi}^{(2)}_j = \frac{X^{(n)}_{\cdot,j} + Z_{\cdot,j}}{\tilde{n}}$ where $\tilde{n} = n + \sum_{i,j} Z_{i,j}$ we have[9] $\phi\left(X^{(n)} + Z\right) = \left(\tilde{\pi}^{(1)}_1, \cdots, \tilde{\pi}^{(1)}_{r-1}, \tilde{\pi}^{(2)}_1, \cdots, \tilde{\pi}^{(2)}_{c-1}\right)$. Note that $Z \sim N(0, 1/\rho_n \cdot I_{rc})$ so it is easy to see that under both private asymptotic regimes ($n\rho_n \to \rho$ and $n\rho_n \to \infty$) we have $\tilde{\pi}^{(1)}_i \xrightarrow{P} \pi^{(1)}_i$ and $\tilde{\pi}^{(2)}_j \xrightarrow{P} \pi^{(j)}_j$ for all $i \in [r]$ and $j \in [c]$ as $n \to \infty$.

We then use this statistic $\phi(X^{(n)} + Z)$ in our unprojected and projected statistic in zCDP-Min-χ^2 to have

[8] For goodness-of-fit testing, ϕ always returns $\mathbf{p}^0$ and $k = 0$ so zCDP-Min-χ^2 is a generalization of zCDP-GOF.

[9] We note that in the case of small sample sizes, we follow a common rule of thumb where if any of the expected cell counts are less than 5, i.e. if $n\tilde{\pi}^{(1)}_i \tilde{\pi}^{(2)}_j \leq 5$ for any $(i,j) \in [r] \times [c]$, then we do not make any conclusion.

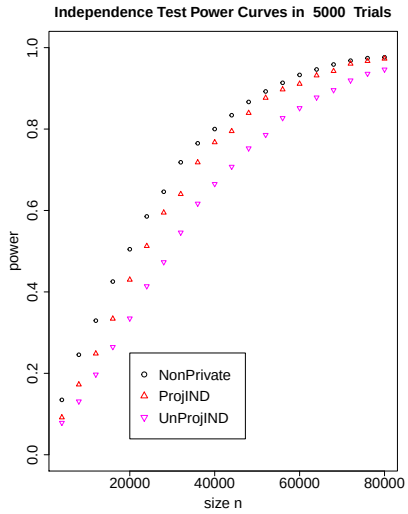

Figure 3: Comparing power between the projected and unprojected statistics in zCDP-Min-χ^2 for independence testing with the classical Pearson chi-square test with 5000 trials each, $\rho = 0.001$ and $\alpha = 0.05$.

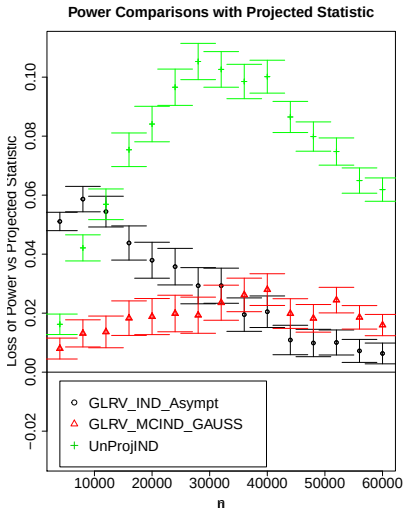

Figure 4: The empirical power loss from using other private independence tests instead of the projected statistic in zCDP-Min-χ^2 for 50,000 trials, $\rho = 0.001$ and $\alpha = 0.05$.

a ρ-zCDP hypothesis test for independence between two categorical variables. Note that in this setting, the projected statistic has a $\chi^2_{(r-1)(c-1)}$ distribution, which is exactly the same asymptotic distribution used in the classical Pearson chi-square independence test.

For our results we will again fix $\alpha = 0.05$ and $\rho = 0.001$. We verify experimentally in the supplementary file that our tests achieve at most α Type I error.

We then compare the power zCDP-Min-χ^2 achieves for either of our test statistics. As a sample of our experiments, we set $r = c = 2$ and $\pi^{(1)} = (2/3, 1/3), \pi^{(2)} = (1/2, 1/2)$. We then sample our contingency table $X^{(n)}$ from Multinomial$(n, \mathbf{p}(\pi^{(1)}, \pi^{(2)}) + \boldsymbol{\Delta})$ where $\boldsymbol{\Delta} = 0.01 \cdot (1, 0, -1, 0)$, so that the null hypothesis is indeed false and should be rejected. We give the empirical power of zCDP-Min-χ^2 in Figure 3 using both the unprojected $R_\rho^{(n)}(\widehat{\theta}^{(n)})$ from (11) and projected statistic $\boldsymbol{\mathcal{R}}_\rho^{(n)}(\widehat{\theta}^{(n)})$ from (12) for 5,000 independent trials and various sample sizes n. Note that again we pick $\widehat{\theta}^{(n)}$ from Theorem 4.2 relative to the statistic we use. We label "NonPrivate" as the classical Pearson chi-square test used on the actual data and "ProjIND" as the test from zCDP-Min-χ^2 with the projected statistic whereas "UnProjIND" uses the unprojected statistic.

The projected statistic again outperforms prior work, so in Figure 4, we plot the difference in power between the projected statistic in zCDP-Min-χ^2 and the competitors (the unprojected statistic and independence tests from [12]) in 50,000 trials. Note that we label "GLRV_MCIND_GAUSS" (resp., "GLRV_IND_Asympt") as the Monte Carlo (resp., asymptotics-based) test with Gaussian noise from [12].

Additional experiments can be found in the supplementary material.

7 Conclusions

We have demonstrated a new broad class of private hypothesis tests zCDP-Min-χ^2 for categorical data based on the minimum chi-square theory. We gave two statistics (*unprojected* and *projected*) that converge to a chi-square distribution when we use Gaussian noise and thus lead to zCDP hypothesis tests. Unlike prior work, these statistics have the same asymptotic distributions in the private asymptotic regime as the classical chi-square tests have in the classical asymptotic regime.

Our simulations show that with both statistics our tests achieve at most α Type I error (see supplementary file). Empirically, the test using the projected statistic significantly improves the Type II error when compared to the unprojected statistic and prior work [12]. Further, our new tests give comparable power to the classical (nonprivate) chi-square tests. The supplementary file contains further applications to GWAS data and other privacy-preserving noise distributions (e.g. Laplace).

8 Acknowledgments

This work was partially supported by NSF grant 1228669.

References

[1] R. Bassily, K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21*, pages 1046–1059, 2016.

[2] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. Discrete multivariate analysis: Theory and practice, 1975.

[3] M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *ArXiv e-prints*, May 2016.

[4] R. Cummings, K. Ligett, K. Nissim, A. Roth, and Z. S. Wu. Adaptive learning with robust generalization guarantees. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 772–814, 2016.

[5] C. Dwork and G. N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016.

[6] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on The Theory and Applications of Cryptographic Techniques*, EUROCRYPT'06, pages 486–503, Berlin, Heidelberg, 2006. Springer-Verlag.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06*, pages 265–284, 2006.

[8] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. In *STOC*, 2015.

[9] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, 2015.

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

[11] S. E. Fienberg, A. Rinaldo, and X. Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Proceedings of the 2010 International Conference on Privacy in Statistical Databases*, PSD'10, pages 187–199, Berlin, Heidelberg, 2010. Springer-Verlag.

[12] M. Gaboardi, H. Lim, R. M. Rogers, and S. P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2111–2120, 2016.

[13] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8), 08 2008.

[14] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1079–1087, New York, NY, USA, 2013. ACM.

[15] V. Karwa and A. Slavković. Differentially private graphical degree sequences and synthetic graphs. In J. Domingo-Ferrer and I. Tinnirello, editors, *Privacy in Statistical Databases*, volume 7556 of *Lecture Notes in Computer Science*, pages 273–285. Springer Berlin Heidelberg, 2012.

[16] V. Karwa and A. Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Ann. Statist.*, 44(P1):87–112, 02 2016.

[17] R. Rogers, A. Roth, A. Smith, and O. Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science, New Brunswick, NJ, USA, October 9 - 11*, pages 487–494, 2016.

[18] O. Sheffet. Differentially private least squares: Estimation, confidence and rejecting the null hypothesis. *arXiv preprint arXiv:1507.02482*, 2015.

[19] S. Simmons, C. Sahinalp, and B. Berger. Enabling privacy-preserving {GWASs} in heterogeneous human populations. *Cell Systems*, 3(1):54 – 61, 2016.

[20] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

[21] C. Uhler, A. Slavkovic, and S. E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013.

[22] D. Vu and A. Slavković. Differential privacy for clinical trial data: Preliminary evaluations. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 138–143, Washington, DC, USA, 2009. IEEE Computer Society.

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

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

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