The Test of Tests: A Framework for Differentially Private Hypothesis Testing

Zeki Kazan [1] **Kaiyan Shi** [2] **Adam Groce** [3] **Andrew Bray** [4]

Abstract

We present a generic framework for creating differentially private versions of *any* hypothesis test in a black-box way. We analyze the resulting tests analytically and experimentally. Most crucially, we show good practical performance for small data sets, showing that at $\epsilon = 1$ we only need 5-6 times as much data as in the fully public setting. We compare our work to the one existing framework of this type, as well as to several individually-designed private hypothesis tests. Our framework is higher power than other generic solutions and at least competitive with (and often better than) individually-designed tests.

1. Introduction

Hypothesis tests are one of the most basic and common statistical analyses that analysts perform on data. The goal of a hypothesis test is to see whether some "effect" in the data (e.g., men are taller than women) is plausibly the result of random variation in the sample, rather than a true fact about the population. Hypothesis tests are the bedrock of statistical analysis in the social sciences, medicine, and other fields, and a variety of hypothesis tests are used, depending on the type of data and the sort of effect one is considering.

However, data in these fields often consists of private information about individuals. Researchers are under moral and legal obligations to protect the privacy of that data and can often only access that data if they can guarantee their analysis will not violate the privacy of those individuals. Differential privacy has emerged as the most convincing formal definition of privacy protection in this setting.

Differentially private versions of many popular hypothesis tests have been created, including private analogues of χ^2 tests (Fienberg et al., 2011; Gaboardi et al., 2016; Johnson & Shmatikov, 2013; Rogers & Kifer, 2017; Uhlerop et al., 2013; Vu & Slavkovic, 2009; Wang et al., 2015), ANOVA tests (Campbell et al., 2018; Swanberg et al., 2019), and many others (Alabi & Vadhan, 2022; Barrientos et al., 2019; Canonne et al., 2020; Couch et al., 2019; Ding et al., 2018; D'Orazio et al., 2015; Narayanan, 2022; Nguyên & Hui, 2017; Solea, 2014; Sheffet, 2017). However, this is work that privacy researchers must carefully repeat for each possible hypothesis test. While this is feasible for the most frequently used tests like χ^2 and ANOVA, it is not plausible to expect this work to be repeated for the wide range of hypothesis tests that exist, many of which are highly specific to particular situations. For example, economists (e.g. (Gatignon & Xuereb, 1997; Jaworski & Kohli, 1993)) use the Chow test to test for a structural break in a regression line; researchers studying ordinal data (e.g. (Kramer, 1996; Uddin & Huynh, 2018)) use ordered logistic regression. Conducting these sorts of analyses privately currently requires collaboration with privacy experts and prohibitive time and effort spent on the study of private statistics before the applied question can even be considered.

In this paper we present a general framework that can automatically create a private version of *any* existing hypothesis test and demonstrate its practicality. For example, at $\epsilon = 1$ our method generally requires no more than 5 or 6 times as much data to detect a given effect as would be required in the non-private setting. This makes our off-the-shelf tool competitive with (and occasionally superior to) some individually tailored private hypothesis tests.

1.1. Our Contributions

The framework we present can be viewed as an instantiation of a method mentioned in the literature. Its origins are not clear, and we think it is best viewed as folklore. It is mentioned in (Canonne et al., 2019) and a similar technique is used in (Cai et al., 2017). Perhaps the clearest description of the method is in (Canonne et al., 2020):

> "There exists a black-box method for obtaining a differentially private tester from any non-private tester A using the sample-and-aggregate frame-

[1] Department of Statistical Science, Duke University, Durham, NC, USA [2] Department of Computer Science, University of Maryland, College Park, MD, USA [3] Department of Computer Science, Reed College, Portland, OR, USA [4] Department of Statistics, UC Berkeley, Berkeley, CA, USA. Correspondence to: Adam Groce <agroce@reed.edu>, Andrew Bray <andrew-bray@berkeley.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

> work (Nissim et al., 2007). Specifically, given any tester A with sample complexity n, we can obtain an ε-differentially private tester with sample complexity $O(n/\varepsilon)$."

Unfortunately, these two sentences are the full extent to which this method is considered in (Canonne et al., 2020). The authors use it only as a point of comparison to show that their method has better asymptotic performance. We are interested not in the asymptotic case, but in concrete performance that will allow the practical use of private statistics on real data.

The authors quoted above also do not fully specify the method to which they are referring. Experienced privacy researchers can fill in the details on their own, but they can be filled in different ways. Our goal here is to fill in the details completely, giving pseudocode and publicly available implementations, but also to fill in these details in the best possible way and to give concrete analysis of the power of the resulting tests. In particular, we do the following:

- We give a framework for creating a private version of any known (non-private) hypothesis test. This uses the subsample-and-aggregate method, with the aggregation done by the uniformly most powerful binomial test given by (Awan & Slavković, 2018).
- We give precise analytic expressions for the power of our test in terms of the power of the underlying non-private test. These finite sample (rather than asymptotic) calculations mean that given a specific public test, one can easily tune parameters in our framework to optimize its power. These calculations are where we derive the claim that we can get $\epsilon = 1$ privacy with 5-6 times the data needed for the public test, but we stress that this is an upper bound *without* test-specific parameter tuning, and in practice our statistical power is often significantly higher.
- We implement our framework and use it to privatize several specific tests[1]. In particular we consider the context where Cannone et al. dismissed this method as less powerful than their proposal. We find that despite their superior asymptotic performance, our framework outperforms their test in a range of practical settings. For example, with a large effect size we obtain 80% power at $n = 65$, while their test is invalid for $n < 359$ and does not reach 80% power until $n = 6500$.

In concurrent work, Peña and Barrientos (Peña & Barrientos, 2022) also provide a generic framework that implements the idea quoted from Canonne et al. above. We delayed the publication of this work to add a full comparison to their framework, which can be found in Section 4. Compared to their framework, ours has meaningfully higher power, and (unlike theirs) can be run for all database sizes, ϵ values, and choices of public test.

Below, we provide an overview of differentially private hypothesis testing. In section 3 we outline our test procedure, providing pseudo-code and an analytic expression for the test's power. In Section 4 we compare our framework to the only existing alternative, that of Peña and Barrientos. Finally, in Section 5 we compare our general framework to some specific existing private tests.

2. Background

In this section, we first discuss hypothesis testing in general. We then introduce differential privacy and the results we will use. Finally, we describe prior work on differentially private hypothesis testing.

2.1. Hypothesis Testing

Consider a researcher who wants to determine if a new miracle weight-loss drug works as advertised. They measure the weight-loss of individuals in two groups, giving one the drug and one a placebo. They wish to know if the drug had a significant effect. Their first step is to formulate a null hypothesis (H_0) - a theory of how the data is distributed. Here H_0 may be that the differences in the groups are due to random variation; the drug has no advantage over the placebo.

To test whether or not the data $\mathbf{x}$ is consistent with H_0, the researcher will compute a *test statistic* $\tau(\mathbf{x})$. The choice of a function τ to compute the test statistic largely determines which hypothesis test being used. For a random database $\mathbf{X}$ drawn according to H_0, the distribution of the statistic $T = \tau(\mathbf{X})$ can be determined either analytically or through simulation. The researcher then computes a *p-value*, the probability that the observed test statistic or a more extreme value would occur under H_0.

Definition 2.1. For an observed test statistic $t = \tau(\mathbf{x})$ and null hypothesis H_0, the one-sided *p-value*, p, is defined as

$$p = \Pr[T \geq t \mid T = \tau(\mathbf{X}) \text{ and } \mathbf{X} \leftarrow H_0].$$

If the function τ is well-chosen, then the more the underlying distribution of $\mathbf{X}$ differs from the distribution under H_0, the more likely a low p-value will be. Typically a significance threshold α is chosen, and H_0 is rejected as a plausible explanation of the data when $p < \alpha$. The choice of α determines the *type I error rate*, the probability of incorrectly rejecting a true null hypothesis.

[1] Code that implements our methods is available at https://github.com/diff-priv-ht/test-of-tests.

We define the *critical value* t^* to be the value of the test statistic t when $p = \alpha$. We use this to define the *statistical power*, a measure of how likely a hypothesis test is to pick up an effect (i.e. to reject a false null hypothesis). The power is a function of how much the underlying distribution of $\mathbf{X}$ differs from the distribution under H_0 as well as the size of the database.

Definition 2.2. For a given alternate data distribution H_A, the *statistical power*, θ, of a hypothesis test is

$$\theta = \Pr[T \geq t^* \mid T = \tau(\mathbf{X}) \text{ and } \mathbf{X} \leftarrow H_A].$$

The goal of hypothesis test design is to maximize statistical power, ideally finding a single test that has good performance for a range of effects.

2.2. Differential Privacy

To convince the public to allow their confidential data to be used for statistical analyses, researchers need to guarantee that sensitive information will not be compromised. Previous methods adopted to protect individual privacy, such as anonymization, have been shown to fail in numerous cases (e.g. (Sweeney, 2002; Narayanan & Shmatikov, 2008; Homer et al., 2008)).

Differential privacy, proposed in 2006 by Dwork et al. (Dwork et al., 2006), is a formal definition of privacy. It protects an individual's privacy by requiring that any output occurs with roughly equal probability regardless of value of that individual's information. Databases that differ only in the data of one individual are called *neighboring* databases.

Definition 2.3 (Differential Privacy)**.** A randomized algorithm $\tilde{f}$ on databases is (ε, δ) differentially private if for all $\mathcal{S} \subseteq \text{Range}(\tilde{f})$ and for databases $\mathbf{x}, \mathbf{x}'$ that only differ only in the values of one row:

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

It is possible that $\delta = 0$. Under this condition, the randomized algorithm $\tilde{f}$ is said to be ε-differentially private. In general, ε indicates the privacy level (a smaller ε indicates a higher privacy guarantee) and δ determines the likelihood of privacy failure. An (ε, δ)-differential privacy guarantees that, with $1 - \delta$ probability, the privacy loss is bounded by e^{ε}.

Differential privacy is resistant to post processing — if an algorithm is differentially private, any further analysis or computation on the output (without dependence on the database) will also result in private output.

Theorem 2.4 (Post Processing)**.** *Let $\tilde{f}$ be an (ε, δ)-differentially private randomized algorithm. Let g be an arbitrary randomized algorithm. Then $g \circ \tilde{f}$ is (ε, δ)- differentially private.*

Any differentially private algorithm must be randomized. The most popular (and simple) method is the *Laplace mechanism*, introduced by Dwork et al. (Dwork et al., 2006), which adds noise drawn from the Laplace distribution to the output of the query one seeks to privatize.

Definition 2.5 (Laplace Distribution)**.** The Laplace Distribution centered at 0 with scale b has probability density function

$$\mathsf{Lap}(x|b) = \frac{1}{2b}\exp\Big(-\frac{|x|}{b}\Big).$$

We write $\mathsf{Lap}(b)$ to denote the Laplace distribution with scale b.

The magnitude through which the alteration of a single row in the database can change the output of a query is called the global sensitivity.

Definition 2.6 (Global sensitivity)**.** The global sensitivity of a function f is:

$$GS_f = \max_{\mathbf{x}, \mathbf{x}'} |f(\mathbf{x}) - f(\mathbf{x}')|,$$

where $\mathbf{x}$ and $\mathbf{x}'$ are neighboring databases.

The standard deviation of the Laplace Distribution used to introduce noise depends on both ε and GS_f.

Definition 2.7 (Laplace Mechanism)**.** Given any function f, the Laplace mechanism is defined as

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + Y,$$

where Y is drawn from $\mathsf{Lap}(GS_f/\varepsilon)$, and GS_f is the global sensitivity of f.

Theorem 2.8 (Laplace Mechanism)**.** *The Laplace mechanism $(\varepsilon, 0)$-differentially private.*

Although the Laplace mechanism ensures that an output will not violate privacy, sometimes the global sensitivity is so large that the Laplace noise overwhelms the signal. The *subsample and aggregate* technique (Nissim et al., 2007) is designed to mitigate this problem. Subsample and aggregate works exactly as it sounds. The database $\mathbf{x}$ with n rows is first partitioned into m groups of approximately equal size. Then a non-private function f is computed in each group independently. Finally, these intermediate results are aggregated through some differentially private mechanism.

2.3. Related Works

There is an extensive (and rapidly expanding) literature examining the problem of converting public hypothesis tests to the private setting. One line of work (Smith, 2008; 2011; Wasserman & Zhou, 2010) studies how fast the distributions of private test statistics converge to the public. These results, however, are often asymptotic and offer little in the way of implementable tests. (Wang et al., 2018) studies the problem of generating a reference distribution more thoroughly, providing a general recipe for approximating the sampling distributions of private test statistics.

Another line of work examines the problem of privatizing the test statistic for the χ^2 test of independence. This includes works in the context of genome-wide association study (GWAS) data (Fienberg et al., 2011; Johnson & Shmatikov, 2013; Uhlerop et al., 2013), although they tend to use asymptotic arguments for the uniformity of p-values. Other work (Gaboardi et al., 2016; Wang et al., 2015) has shown that Monte Carlo methods can produce better reference distributions. (Vu & Slavkovic, 2009) provides concrete methods for producing a p-value by adjusting for Laplace noise, while (Rogers & Kifer, 2017) proposes alternate test statistics that have reference distributions with preferable properties.

Recent works in differentially private hypothesis testing have begun to include in-depth power analyses. (Awan & Slavković, 2018) constructed the universally most powerful test for binomial data (see Section 3.1 for further discussion). (Brenner & Nissim, 2010) shows that a universally most powerful test cannot exist for data with a domain containing more than two elements. Nguyên and Hui propose methods for differentially private survival analysis (Nguyên & Hui, 2017). Two works have addressed the problem of studying the difference in means of normal distributions (Ding et al., 2018; D'Orazio et al., 2015), and several consider the problem of hypothesis testing for linear regression coefficients (Alabi & Vadhan, 2022; Barrientos et al., 2019; Sheffet, 2017). Of these, (Barrientos et al., 2019) is notable for sharing some conceptual ideas with the framework we propose here. A few works propose tests for the mean of a normal distribution in the univariate (Solea, 2014) and mulivariate (Canonne et al., 2020; Narayanan, 2022) settings. Two works study the one-way ANOVA (Campbell et al., 2018; Swanberg et al., 2019), although these are outperformed by work on nonparametric alternatives (Couch et al., 2019). (Avella-Medina, 2021) proposes a hypothesis test based on M-estimators that is applicable to general parametric models, including many of the above.

3. Framework

Here we introduce our test of tests (ToT) framework and analyze its power. We also discuss how to optimize the framework's parameters for a given situation.

3.1. Private Binomial Test

Awan and Slavković (Awan & Slavković, 2018) develop a uniformly most powerful test for binomial data. They define the Truncated-Uniform-Laplace (Tulap) Distribution, the sum of the discrete Laplace and uniform distributions. The distribution is parameterized by a location parameter, m, and a scale parameter, $b \in (0, 1)$. Its CDF has a closed form; see Definition 4.1 in (Awan & Slavković, 2018). [2]

Let $A \sim \text{Binomial}(n, p)$. Awan and Slavković show that the private test statistic $Z|A \sim \text{Tulap}(A, e^{-\varepsilon})$ is an ε-differentially private estimate of A. They also provide an algorithm for producing a p-value to test the hypothesis

$$H_0 : p \leq p_0 \quad \text{and} \quad H_A : p > p_0$$

and show that the p-value produced is the smallest ε-DP p-value for this test. See Theorem 7.2 and Algorithm 2 in (Awan & Slavković, 2018) for further details.

3.2. Our Algorithm

We now describe our general algorithm, which we call *test of tests* (ToT), which can privatize all hypothesis tests. The formalization is presented in Algorithm 1 and a graphical representation in Figure 1. We are given a database $\mathbf{x}$ of size n, and our goal is to run an ε-private version of hypothesis test τ [3] on that database with significance threshold α. We first partition the input database into m equal sized subsets $\mathbf{x}_1, \ldots, \mathbf{x}_m$. In practice, if $m \nmid n$, then the subsets should be of sizes $\lfloor \frac{n}{m} \rfloor$ and $\lceil \frac{n}{m} \rceil$ as appropriate. The following results will assume that $m \mid n$ for simplicity. In each subset, we conduct the public test τ, computing the p-value and accepting/rejecting according to a sub-test significance threshold of α_0. If the number of data points in a subsample is insufficient to run the public test, the p-value is drawn from $\text{Unif}(0, 1)$. Let a be the number of rejects. Under the null distribution, each instance of τ rejects with probability α_0, so a follows a binomial distribution.

We then conduct Awan and Slavković's private binomial test on a to see if it is consistent with a binomial distribution with parameter α_0. To privatize a, we define $z = \text{Tulap}(a, e^{-\varepsilon})$. Let $B \sim \text{Binomial}(m, \alpha_0)$ and $N \sim \text{Tulap}(0, e^{-\varepsilon})$. Then, the reference distribution is $B + N$ and so the p-value is $P(B + N \geq z)$.

Note that of the inputs listed, $\mathbf{x}$, τ, ϵ, and α are true inputs from the user, while m and α_0 are parameters that can be optimized. We discuss this optimization in Section 3.4.

The privacy and validity of Algorithm 1 follow immediately from its design. These proofs (and most others in this paper) are placed in Appendix A due to space constraints.

Theorem 3.1. *Algorithm 1 is ε-differentially private.*

Theorem 3.2. *Algorithm 1 is valid. That is, when the data is drawn from H_0 the probability of rejection is at most α.*

[2] The Tulap distribution has a third parameter, q, but we always set $q = 0$ because our aim is to have $\delta = 0$. Allowing $\delta > 0$ could be done by changing q, and would increase the power of our test.

[3] For concision, we use τ to represent a test that utilizes $\tau(\mathbf{x})$.

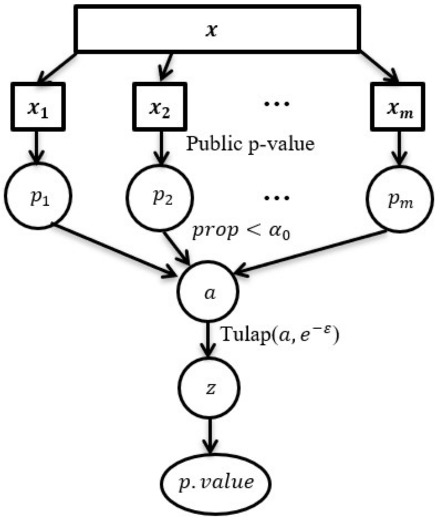

Figure 1. A graphical representation of Algorithm 1.

Algorithm 1 Test of Tests

```
Input: x, $\tau$, $\varepsilon$, $\alpha$, $m$, $\alpha_0$
Partition x into subsets $\mathbf{x}_1, \ldots, \mathbf{x}_m$
for $j = 1$ to $m$ do
    if $\tau$ can be run on $\mathbf{x}_j$ then
        $p_j \longleftarrow \tau(\mathbf{x}_j)$
    else
        $p_j \sim$ Unif$(0, 1)$
    end if
end for
$a \longleftarrow |\{p_j : p_j < \alpha_0\}|$
$z \longleftarrow$ Tulap$(a, e^{-\varepsilon})$
$p.value \longleftarrow P(B + N \geq z)$
Output: $z$, $p.value$
```

3.3. Theoretical Power

We can now analyze the statistical power of the test of tests framework. We begin by noting its asymptotic sample complexity as a function of ε. This was stated without proof by (Canonne et al., 2020), and we provide a proof in Appendix A.

Theorem 3.3. *The number of samples required for our test to achieve ρ power is $n = \mathcal{O}(c/\varepsilon)$, where c is the number of samples needed by the non-private test, τ.*

The focus of this work is not asymptotic performance, but practical performance on small n, and for that analysis we need an exact computation of the power of any ToT instantiation.

Theorem 3.4. *Let θ be the power of the public test τ in each of the m subsamples with significance level α_0. Let $A \sim$ Binomial(m, θ), $Z|A \sim$ Tulap$(A, e^{-\varepsilon})$, $B \sim$ Binomial(m, α_0), and $N \sim$ Tulap$(0, e^{-\varepsilon})$. Then the power of our test is*

$$\mathcal{P}(\varepsilon, \alpha, m, \alpha_0, \theta) = (1 - F_Z(F^{-1}_{B+N}(1 - \alpha))).$$

Note that F^{-1}_{B+N} does not have a known analytic form, so when computing the power via Theorem 3.4, the quantiles of the distribution must be determined numerically.

If one is interested in a particular public hypothesis test with known characteristics, the above result can be used to determine a bound on the sample size required for the privatized test to achieve ρ power. (Simple proof in Appendix A.)

Corollary 3.5. *Suppose that a public hypothesis test τ requires at most n data points to achieve θ power at a significance level α_0 for any choice of the data. Then, in order for the private test with privacy parameter ε to achieve ρ power at a significance level α, the necessary number of data points is bounded above by $n\tilde{m}$, where $\tilde{m}$ is the smallest m such that $\rho \leq \mathcal{P}(\varepsilon, \alpha, m, \alpha_0, \theta)$.*

Since the power $\mathcal{P}$ is strictly increasing with respect to m, it is straightforward to determine $\tilde{m}$ numerically. This allows general statements about how much more data a private test will need compared to the equivalent public test. Some examples are shown in Table 1. For example, the first row shows that *any* public test that achieves 80% power at $\alpha = 0.05$ can be privatized at $\varepsilon = 1$ (by using exactly that public test as the subtest) to get the same power and significance with $\tilde{m} = 5$, meaning that the private test needs 5 times the data of the public test. For 95% power 6 times the data of the public test is needed. (For $\varepsilon = 0.01$ those multiples are 44x and 52x respectively.)

We stress that these general statements, while they are very strong, are only upper bounds. That is because without specifying a test, one cannot say what would happen when the α_0 for the subtests is different than the α one is attempting to achieve in the overall test. Given any particular test, one can vary α_0 and find better settings. For example, a z-test with $\alpha = 0.05$ run on data with an effect size of 0.65 standard deviations will reach 80% power at $n = 20$, meaning that the statement above would guarantee no more than $n = 100$ needed to get the same power in the $\varepsilon = 1$ private setting. But allowing α_0 to take values other than 0.05, we find that one can actually do this with $n = 70$, meaning a 3.5× cost of privacy, rather than 5×. At $\varepsilon = 0.01$, it requires $n = 420$, for a 21× cost of privacy, instead of the 44 given by the upper bound in the table.

As another example, take an ANOVA test with three groups run on data with equal within-group and between-group variance. The upper bounds in the table for 95% power require 6× data at $\varepsilon = 1$ and 52× data at $\varepsilon = 0.01$, but the optimized test requires 3.6× and 41× data instead.

θ	α_0	ρ	α	ε	$\tilde{m}$
0.80	0.05	0.80	0.05	1	5
0.80	0.05	0.80	0.05	0.1	44
0.95	0.05	0.95	0.05	1	6
0.95	0.05	0.95	0.05	0.1	52

Table 1. For a public test τ that requires achieves θ power at significance level α_0, in order for our ε-private test to achieve ρ power at a significance level of α, we require at most a factor of $\tilde{m}$ more data.

3.4. Optimization

The variables m (the number of subsamples) and α_0 (the sub-test significance threshold) must be optimized. Fortunately, we find that doing an extremely thorough optimization for these parameters is not necessary. The optimal m, α_0 combination for one effect size generally does an adequate job across a large range of effect sizes, with a decrease in power generally in the range of 1 to 2%.

When an approximate expected effect size is known, we can easily compute θ, the power of the public test τ, for any sample size. For fixed m, standard techniques can be used to find the α_0 that maximizes $\mathcal{P}(\varepsilon, \alpha, m, \alpha_0, \theta)$. This can then be repeated for all m in a reasonable set to find the otpimal m, α_0 pair. For our simulations, we use the set $\{1, 2, 3, \ldots, \lfloor\sqrt{n}\rfloor, \ldots \lfloor\frac{n}{3}\rfloor, \lfloor\frac{n}{2}\rfloor, n\}$ and find that this process takes less than 20 seconds for a t-test at $n = 100$.

In practice, however, an approximate expected effect size is often not known a priori. In this setting, we suggest fixing a desired power, ρ, and optimizing for the m, α_0 pair that minimizes the effect size detectable with ρ power. This can be achieved by beginning with a grid of effect sizes and performing a binary search, using the above process for known effect size at each step, to find the minimum effect size in the grid detectable with ρ power. Then use the m, α_0 from that combination to run the test of tests. In our simulations, we use a length-16 grid and find this process takes less than a minute and a half for a t-test at $n = 100$.

We note, interestingly, that the optimization tends to favor high values of m, with very small subsamples and high significance thresholds on the subtests. It turns out that aggregating a large number of minimally-informative tests is preferable to a small number of more reliable tests.

4. Comparison to Peña-Barrientos Framework

Peña and Barrientos (Peña & Barrientos, 2022), simultaneously to this work, proposed their own framework (henceforth referred to as PB) for privatizing arbitrary public hypothesis tests. Like our framework, theirs follows the folklore subsample-and-aggregate idea mentioned by Canonne et al. (Canonne et al., 2019). Both methods begin by running the public test on subsamples of the data set, but the methods of aggregation are different. PB develop what is essentially a custom-built binomial test based on a randomized response-type method. We instead use the Awan and Slavković binomial test, which is provably optimal. As a result, our framework is the highest-power framework possible within this general type of design (see Appendix A).

Theorem 4.1. *For any choice of public test τ and privacy parameter ϵ, the statistical power of the private test resulting from the ToT framework will be higher than that resulting from the PB framework.*

The PB framework has drawbacks beyond the simple lowering of power. The details of the test mean that it cannot get valid results at all parameter settings. In particular, there is a minimum m value at which the test can be run. Since the public test itself often requires a certain amount of data, this means that a meaningful amount of data is sometimes required before the PB test can be used at all. (For example, with $\varepsilon = 1$ and $\alpha = 0.05$, PB requires $m \geq 7$, increasing to $m \geq 67$ when $\varepsilon = 0.1$.)

Furthermore, given values of the other parameters, α_0 *must* be set to a specific value so that the resulting α of the larger test is accurate. This removes a degree of freedom in optimization, further worsening power. PB give two methods of setting parameters. The first involves no real optimizing at all, suggesting that m be set as low as possible. This is meant for the "low-power" setting, where the goal is to achieve significant power at the lowest possible n. (In contrast, we find that very high m often results in better power.) We use this method in our comparison calculations, though we note that they also suggest one could calculate power curves at a variety of parameter values and choose the best parameters through visual inspection. This method is necessary to reach high power, because using the lowest possible m results in an upper bound on the power of the PB test, meaning that the power does not approach 100% as n grows. (This bound can be as low as 80% in realistic scenarios.)

In Appendix A we provide an analogue of our Theorem 3.4 for the PB test so that we can directly compare power instead of relying upon approximate simulations. Figure 2 shows the exact power of PB and of our ToT framework for two examples. We use a t-test as the public test. In the top panel, with a moderate effect size and $\varepsilon = 1$, the PB framework requires 40% more data to achieve 80% power ($n = 200$ for ToT, $n = 280$ for PB). In the bottom panel, with a larger effect size and $\varepsilon = 0.1$, the difference is much greater. Here ToT only requires $n = 125$ to get 80% power, while PB cannot be run at all until $n \geq 134$ and doesn't get 80% power until $n = 348$ (a 178% increase). Additional comparisons privatizing a z-test and an ANOVA can be found in Appendix B.

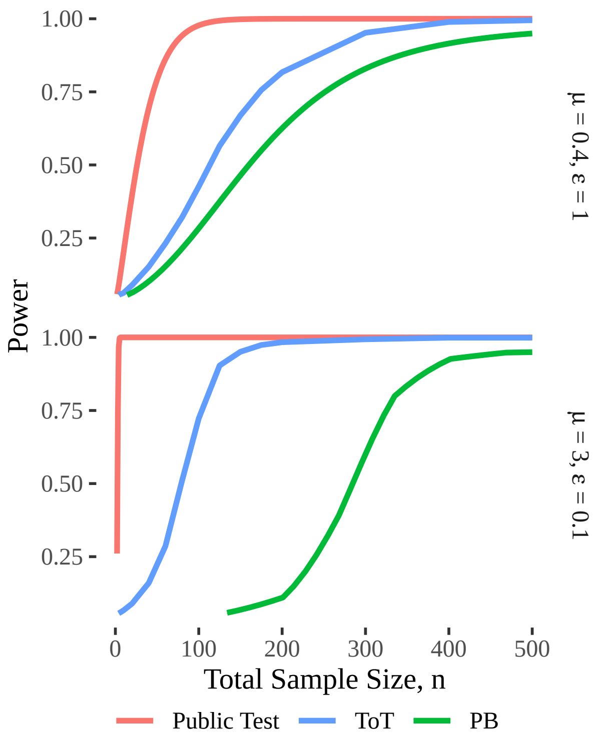

Figure 2. Power comparison between (Peña & Barrientos, 2022) and the test of tests for various sample sizes n with $\alpha = 0.05$ and a t-test with $\sigma = 1$. The top panel has an effect size of $\mu = 0.4$ and $\varepsilon = 1$; the bottom has $\mu = 3$ and $\varepsilon = 0.1$. We optimize m, α_0 for the test of tests as discussed in Section 3.4 with target $\rho = 0.9$.

5. Comparisons to Tailored Tests

In this section, we demonstrate the use of the ToT framework on a selection of hypothesis tests, namely a test for the mean of multivariate normal data and a one-way ANOVA. These tests have both been the subject of prior work, so we can compare our general-purpose technique to tests experts carefully developed for specific situations.

5.1. Mean of Multivariate Normal Data

Since the general method we are using was described (and dismissed) in (Canonne et al., 2020) (henceforth referred to as CKMUZ), we begin by using our framework to develop a test for the same situation. Here the analyst observes data drawn from a multivariate normal distribution, $\mathbf{x} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\mathbf{X}_i \sim \mathcal{N}_d(\boldsymbol{\mu}, \mathbb{I}_d)$. The null hypothesis is that $\boldsymbol{\mu} = \mathbf{0}$, while the alternate has $\boldsymbol{\mu} \neq \mathbf{0}$. A public hypothesis test for this setting uses the test statistic $Z = n \sum_{j=1}^{d} \bar{X}_j^2$, which is known to follow the distribution $\chi^2(df = d)$. We will use this test for the p-value computation step in Algorithm 1. We also compute the power of this test.

Theorem 5.1. *Let F_0 and F_A be the CDFs of $\chi^2(df = d)$ and $\chi^2\left(df = d, \lambda = n\sum_{j=1}^{d} \mu_j^2\right)$, respectively, where n is the sample size and μ_j is the j^{th} entry of $\boldsymbol{\mu}$. Then the power of the public test with significance level α is* $1 - F_A(F_0^{-1}(1-\alpha))$.

This analytic expression for the power is not needed to perform the test, but having it allows optimization to be done more efficiently and means our figures show the exact power of our test, rather than a Monte Carlo approximation.

CKMUZ proposes a computationally efficient private test for this setting[4] and proves that its asymptotic dependency on ε and effect size is superior to the strategy we use here. (Narayanan, 2022) gives another test with yet better asymptotic performance. Unfortunately, this algorithm is described only in general asymptotic terms, without the concrete details necessary for implementation. As a result, we compare to the test given by CKMUZ.

The CKMUZ test does not have an adjustable α value. Instead, the analyst is inputs a parameter, γ,[5] that is a lower bound on the total variation distance between the null distribution $\mathcal{N}_d(\mathbf{0}, \mathbb{I}_d)$ and the alternate distribution $\mathcal{N}_d(\boldsymbol{\mu}, \mathbb{I}_d)$. The test is then guaranteed to distinguish the two distributions with probability 2/3 with a required sample complexity of $\tilde{\mathcal{O}}\left(d^{1/2}/\gamma^2 + d^{1/2}/(\gamma\varepsilon)\right)$. This means that Type 1 error will approach 1/3 for sufficiently high n, but it can be much higher at low n. In fact, there is a threshold of $\max\left\{25\log\frac{d}{\delta}, \frac{5}{\varepsilon}\log\frac{1}{\delta}\right\}$ below which the test always rejects (100% Type 1 error). Just above this threshold it has Type 1 error of roughly 50%, where it remains for the entire range of n values we are considering. (This Type 1 error is proven analytically in Appendix A and confirmed experimentally in Appendix B.) As a result, we set the α value in our test to 0.5 for a fair comparison, but we note that our test has the advantage that α can be set arbitrarily.

For our comparison, we set $d = 100$ and $\varepsilon = 1$. ToT uses pure differential privacy, with $\delta = 0$, but CKMUZ requires a nonzero δ. We set $\delta = 10^{-3}$, which we believe to be very favorable, much higher than is generally considered acceptable in practice. We set $\gamma = 0.1$. We do not present the power curve for the CKMUZ test until it stops summarily rejecting all inputs, which for these parameters happens at $n = 359$. We consider two possible effects, one where the true mean differs by 0.1 standard deviations in all coordinates, and one where it differs by 0.5 in only a single coordinate. The results can be seen in Figure 3.

[4] (Narayanan, 2022) points out an error in this work, but it is in a second algorithm that is irrelevant to this comparison.

[5] We call this parameter γ, rather than the α from the paper to avoid conflict in notation.

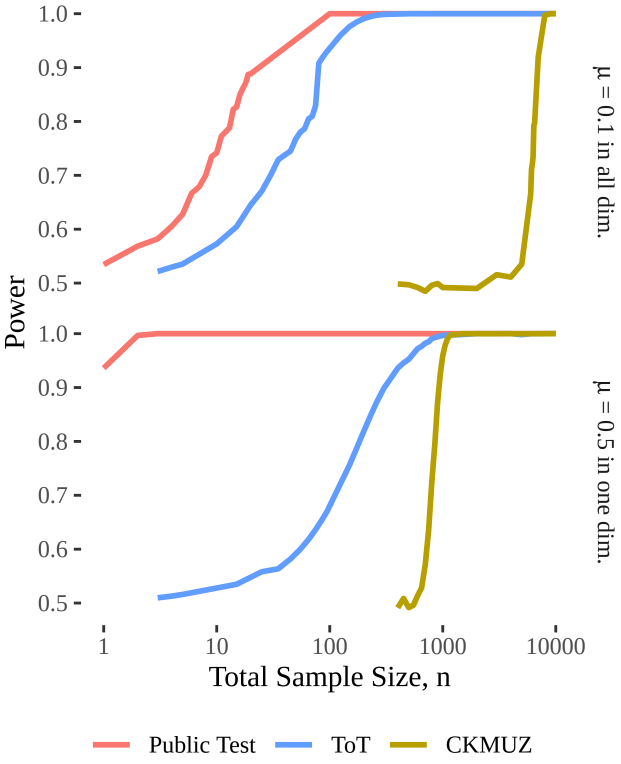

Figure 3. Power comparison between CKMUZ and the tests of tests for various sample sizes n. For the top panel, $\mu_i = 0.1$ for all i; for the bottom, $\mu_1 = 0.5$ and $\mu_i = 0$ for $i \neq 1$. We set $d = 100$, $\varepsilon = 1$, and α for the test of tests and public test is set to match the Type I Error of the CKMUZ test. We optimize m, α_0 for the test of tests as discussed in Section 3.4 with target $\rho = 0.9$ for $n \leq 500$ and $\rho = 0.99$ for $n > 500$.

In the first case, with the larger effect size, we reach 80% power at $n = 65$, while the CKMUZ test isn't even valid until $n = 359$ and doesn't reach 80% power until $n = 6500$. In the second case, with a smaller effect size, the difference is smaller though still substantial. ToT requires only 20% as much data to reach 80% power — 190 data points compared to 850. (At 99% power, the gap is smaller, with ToT needing 87% of the data needed by CKMUZ.) Additional comparisons are provided in Appendix B.

Of course, the CKMUZ test *does* have better asymptotic performance, so there is some sufficiently small ε and effect size (and sufficiently large δ) such that it becomes the higher-power test. However, for a wide variety of practical situations, those superior asymptotics have not yet come into play, and ToT is the better choice.

5.2. One-way Analysis of Variance

As our second example we consider a one-way ANOVA, which examines whether groups of data have the same mean. Formally, each of g groups has a mean μ_g. Data within each group is drawn from $\mathcal{N}(\mu_i, \sigma^2)$ for a fixed, unknown σ. Under H_0 all groups have equal mean, while in H_A some means differ. The classical test for this setting uses the F statistic, which follows a known distribution under H_0. (For a more thorough introduction, see (Rice, 2007).) For this analysis, we focus on the case of equal-sized groups. We call the ratio of the between-group variance and the within-group variance $\eta = \mathsf{Var}(\mu_1, \ldots, \mu_g)/\sigma^2$ the effect size. The power of an ANOVA in this setting has a known solution available in most statistical software.

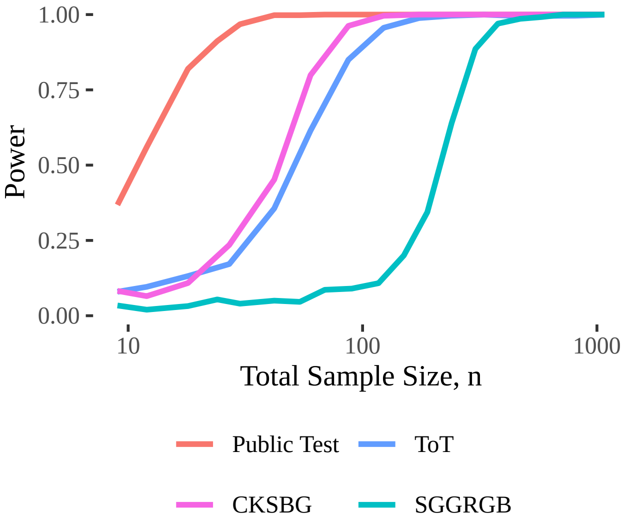

Figure 4. Power comparison between CKSBG. SGGRGB, and the test of tests for various sample sizes n. The effect size is $\eta = 1$, $\varepsilon = 1$, and number of groups $g = 3$. All groups are of equal size and $\alpha = 0.05$. All non-public power curves are estimated via simulation. We optimize m, α_0 for the test of tests as discussed in Section 3.4 with target $\rho = 0.9$.

This setting is the subject of a significant line of work. (Campbell et al., 2018) give the first private test, which was later improved upon by (Swanberg et al., 2019) (henceforth, SGGRGB) and then (Couch et al., 2019) (henceforth CKSBG). To the best of our knowledge, the private nonparametric test of CKSBG is the most powerful private test available in this setting and thus will serve as a benchmark for the performance of the test of tests.

For comparison to the test of CKSBG, we choose the setting in Figure 3 of (Couch et al., 2019) which examines privacy level $\epsilon = 1$ and effect size $\eta = 1$. As shown in the top panel of Figure 4, the test of tests is slightly worse, but the difference is small. (CKSBG require 22% less data to reach 80% power.) It performs much better than the test of SGGRGB. Importantly, unlike the tests tailored to this setting, the test of tests does not require the estimation of a reference distribution via simulation. Thus, the test of tests is faster to run (and p-values are arguably more accurate).

Varying the setting shows that these two tests are incomparable. We have included additional examples in Appendix B. With a smaller effect size, the gap between ToT and prior work increases, while a large effect size and/or smaller ε actually results in ToT becoming the state of the art most powerful test, though by a small margin. Regardless of the specifics of the comparison, we find it exciting that our general framework is at all comparable to a highly-refined test carefully developed for a specific situation.

Acknowledgments

We would like to thank Andrés Barrientos for sharing code that implements his parallel work and Canyon Foot for sharing some code that appears in our implementations. All authors were supported by the National Science Foundation under Grant No. SaTC-1817245. Kaiyan Shi acknowledges additional support from the U.S. Army Research Office under Grant No. W911NF-20-1-0015 and Zeki Kazan from NSF Grant No. SES-2217456. We would also like to thank the three anonymous reviewers who provided excellent feedback.

References

Alabi, D. and Vadhan, S. Hypothesis testing for differentially private linear regression. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

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

Awan, J. and Slavković, A. Differentially private uniformly most powerful tests for binomial data. In *Advances in Neural Information Processing Systems*, pp. 4208–4218, 2018.

Barrientos, A. F., Reiter, J. P., Machanavajjhala, A., and Chen, Y. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, pp. 1–24, 2019.

Brenner, H. and Nissim, K. Impossibility of differentially private universally optimal mechanisms. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 71–80. IEEE, 2010.

Cai, B., Daskalakis, C., and Kamath, G. Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning*, pp. 635–644. PMLR, 2017.

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

Canonne, C. L., Kamath, G., McMillan, A., Smith, A., and Ullman, J. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 310–321, 2019.

Canonne, C. L., Kamath, G., McMillan, A., Ullman, J., and Zakynthinou, L. Private identity testing for high-dimensional distributions. *Advances in Neural Information Processing Systems*, 33:10099–10111, 2020.

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

Ding, B., Nori, H., Li, P., and Allen, J. Comparing population means under local differential privacy: with significance and power. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

D'Orazio, V., Honaker, J., and King, G. Differential privacy for social science inference. *Sloan Foundation Economics Research Paper*, 2015.

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

Fienberg, S. E., Slavkovic, A., and Uhler, C. Privacy preserving gwas data sharing. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 628–635. IEEE, 2011.

Gaboardi, M., Lim, H.-W., Rogers, R. M., and Vadhan, S. P. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR, 2016.

Gatignon, H. and Xuereb, J.-M. Strategic orientation of the firm and new product performance. *Journal of marketing research*, 34(1):77–90, 1997.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

Jaworski, B. J. and Kohli, A. K. Market orientation: antecedents and consequences. *Journal of marketing*, 57(3): 53–70, 1993.

Johnson, A. and Shmatikov, V. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1079–1087. ACM, 2013.

Kramer, B. An ordered logit model for the evaluation of dutch non-life insurance companies. *De Economist*, 144 (1):79–91, 1996.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large datasets (how to break anonymity of the netflix prize dataset). *University of Texas at Austin*, 2008.

Narayanan, S. Private high-dimensional hypothesis testing. In *Conference on Learning Theory*, pp. 3979–4027. PMLR, 2022.

Nguyên, T. T. and Hui, S. C. Differentially private regression for discrete-time survival analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1199–1208. ACM, 2017.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84. ACM, 2007.

Peña, V. and Barrientos, A. F. Differentially private hypothesis testing with the subsampled and aggregated randomized response mechanism. *arXiv preprint arXiv:2208.06803*, 2022.

Rice, J. *Mathematical Statistics and Data Analyss*. Brooks/Cole, Belmont, CA, 2007.

Rogers, R. and Kifer, D. A new class of private chi-square hypothesis tests. In *Artificial Intelligence and Statistics*, pp. 991–1000, 2017.

Sheffet, O. Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3105–3114. JMLR. org, 2017.

Smith, A. Efficient, differentially private point estimators. *arXiv preprint arXiv:0809.4794*, 2008.

Smith, A. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 813–822. ACM, 2011.

Solea, E. Differentially private hypothesis testing for normal random variables. Master's thesis, Pennsylvania State University, 2014.

Swanberg, M., Globus-Harris, I., Griffith, I., Ritz, A., Groce, A., and Bray, A. Improved differentially private analysis of variance. *Proceedings on Privacy Enhancing Technologies*, 2019(3):310–330, 2019.

Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Uddin, M. and Huynh, N. Factors influencing injury severity of crashes involving hazmat trucks. *International journal of transportation science and technology*, 7(1):1–9, 2018.

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

Vu, D. and Slavkovic, A. Differential privacy for clinical trial data: Preliminary evaluations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pp. 138–143. IEEE, 2009.

Wang, Y., Lee, J., and Kifer, D. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

Wang, Y., Kifer, D., Lee, J., and Karwa, V. Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, 8(1), 2018.

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

A. Proofs

Here we include the proofs that were excluded from the main body.

A.1. Proofs for Section 3

First, we prove that our test is private and valid. These are straightforward results using known results/techniques.

Theorem 3.1. *Algorithm 1 is ε-differentially private.*

Proof. By Subsample and Aggregate (Nissim et al., 2007) and Theorem 6.1 in Awan and Slavković (Awan & Slavković, 2018), which shows the release of the statistic with Tulap noise satisfies privacy, the release of z is ε-differentially private. By Theorem 2.4 (post processing), the release of the p-value is also ε-differentially private. □

Theorem 3.2. *Algorithm 1 is valid. That is, when the data is drawn from H_0 the probability of rejection is at most α.*

Proof. Each of the m subgroups will reject (i.e., be included in the count a) with probability at most α. For most this follows from the validity of the public test τ. In cases when τ can't be run, it follows from the uniform selection of p_j. From there, the validity follows immediately from the results of (Awan & Slavković, 2018). □

Next, we prove that the sample complexity of our test is an $\mathcal{O}(1/\varepsilon)$ factor more than that of the public test.

Theorem 3.3. *The sample complexity required for our test to achieve ρ power is*

$$n = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$$

Proof. Consider an alternative test with two changes: we add Laplace noise instead of Tulap noise and we use the proportion below the threshold as our test statistic, rather than the count. In line 6 of Algorithm 1, we use the alternative test statistic

$$\tilde{z} = \frac{a + \text{Lap}\left(\frac{1}{\varepsilon}\right)}{m} = \frac{a}{m} + \frac{\text{Lap}\left(\frac{1}{\varepsilon}\right)}{m}.$$

This output is guaranteed to be ε-differentially private by Theorems 2.4 and 2.8. It has sensitivity 1 since changing a row in the dataset can only change the p-value in one group and therefore can change a by at most 1.

Let $L = \frac{\text{Lap}\left(\frac{1}{\varepsilon}\right)}{m}$. Then there exists some $\varepsilon = \varepsilon^*$, some number of subgroups m, and some subgroup size n such that $L = \frac{\text{Lap}\left(\frac{1}{\varepsilon^*}\right)}{m}$ is small enough that power ρ can be achieved with sample size mn, where n is the number of datapoints in each group and m is the number of groups.

Now let $\varepsilon' = \frac{\varepsilon^*}{k}$ and $m' = km$. By the scaling property of Laplace distribution, we have

$$\frac{\text{Lap}\left(\frac{1}{\varepsilon'}\right)}{m'} = \frac{\text{Lap}\left(\frac{1}{\varepsilon^*/k}\right)}{km} = \frac{k\text{Lap}\left(\frac{1}{\varepsilon^*}\right)}{km} = \frac{\text{Lap}\left(\frac{1}{\varepsilon^*}\right)}{m} = L.$$

This means that if the number of datapoints in each group, n, is unchanged, then for $\varepsilon' = \frac{\varepsilon^*}{k}$ the noise added is still L if there are k times as many groups.

We then need to consider how the additional groups affect the term $\frac{a}{m}$ (this term is independent of ε). Note that $\text{E}[\frac{a}{m}] = \theta$, a constant, and so changing the number of groups will have no effect on this expectation. But $\text{Var}[\frac{a}{m}] = \frac{\theta(1-\theta)}{m}$, so increasing the number of groups by a factor of k will decrease the variance by a factor of k. Thus, the distribution of $\tilde{z}$ will have unchanged mean, and will still be distributed according to a binomial distribution, but it will now have lower variance. This means the power of the test must necessarily increase.

This analysis shows that if power ρ can be achieved with ε at sample size mn, then it can also be achieved with ε/k at sample size smaller than or equal to kmn. In other words, the sample complexity is inversely related to ε. This shows that this alternate test has sample complexity $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$. Our test, which uses the uniformly most powerful ε-differentially private binomial test instead of simple Laplace noise, must also have sample complexity $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$. □

Now we compute the exact (rather than asymptotic) power of the test.

Let $f_X(x)$ and $F_X(x)$ refer to the probability density function (PDF) and cumulative density function (CDF), respectively, of a random variable X. We now establish two lemmas about the CDFs of relevant variables.

Lemma A.1. *Let θ be the power of the public test τ in each of the m subsamples with significance level α_0. Let $A \sim \text{Binomial}(m, \theta)$ and let $Z|A \sim \text{Tulap}(A, e^{-\varepsilon})$. Then the cumulative distribution function of Z is*

$$F_Z(z) = \sum_{i=0}^{M} F_{Z|A}(z|i)\; f_A(i).$$

Proof. Let $f(a, z)$ be the joint probability density function of A and Z. Then the CDF of Z is

$$\begin{aligned}
F_Z(z) &= \int_{-\infty}^{z} f_Z(t)\; dt \\
&= \int_{-\infty}^{z} \sum_{i=0}^{M} f(i, t)\; dt \\
&= \sum_{i=0}^{M} \int_{-\infty}^{z} f_{Z|A}(t|i)\; f_A(i)\; dt \\
&= \sum_{i=0}^{M} F_{Z|A}(z|i)\; f_A(i)
\end{aligned}$$

□

Lemma A.2. *Let $B \sim \text{Binomial}(m, \alpha_0)$ and $N \sim \text{Tulap}(0, e^{-\varepsilon})$. Then the cumulative distribution function of $B + N$ is*

$$F_{B+N}(t) = \sum_{i=0}^{M} f_B(i) F_N(i - t).$$

Proof. By convolution, the CDF of $B + N$ is

$$\begin{aligned}
F_{B+N}(t) &= P(B + N \leq t) \\
&= \sum_{i=0}^{M} \int_{t-i}^{\infty} f_B(i) f_N(x)\; dx \\
&= \sum_{i=0}^{M} f_B(i)(1 - F_N(t - i)) \\
&= \sum_{i=0}^{M} f_B(i) F_N(i - t)
\end{aligned}$$

□

Theorem 3.4. *Let θ be the power of the public test τ in each of the m subsamples with significance level α_0. Let $A \sim \text{Binomial}(m, \theta)$, $Z|A \sim \text{Tulap}(A, e^{-\varepsilon})$, $B \sim \text{Binomial}(m, \alpha_0)$, and $N \sim \text{Tulap}(0, e^{-\varepsilon})$. Then the power of our test is*

$$\mathcal{P}(\varepsilon, \alpha, m, \alpha_0, \theta) = (1 - F_Z(F_{B+N}^{-1}(1 - \alpha))).$$

Proof. Let W_i be a random variable which outputs 1 if $p_i < \alpha_0$ and 0 otherwise. It is thus distributed $W_i \sim \text{Bernoulli}(\theta)$. Then $A = \sum_{i=1}^{m} W_i \sim \text{Binomial}(m, \theta)$ is the number of p-values less than α_0 and $Z \mid A \sim \text{Tulap}(A, e^{-\varepsilon})$ is the differentially private estimate of A.

$B \sim$ Binomial (M, α_0) is the number of p-values less than α_0 under the null hypothesis and $N \sim$ Tulap$(0, e^{-\varepsilon}, 0)$ is the required amount of Tulap noise to maintain ε-differential privacy. For any valid hypothesis test, under the null hypothesis, $\theta = P(p_i < \alpha_0) \leq \alpha_0$. Testing the hypothesis of interest is thus equivalent to testing

$$H_0 : \theta \leq \alpha_0 \quad \text{and} \quad H_A : \theta > \alpha_0.$$

The p-value for this test is

$$p(Z) = P(B + N \geq Z \mid Z).$$

The power of this test is then

$$\begin{aligned} P(p(Z) \leq \alpha) &= P(1 - F_{B+N}(Z) \leq \alpha) \\ &= P(Z \geq F_{B+N}^{-1}(1 - \alpha)) \\ &= (1 - F_Z(F_{B+N}^{-1}(1 - \alpha))). \end{aligned}$$

By Lemmas A.1 and A.2, $F_Z(z) = \sum_{i=0}^{M} F_{Z|A}(z|i)\ f_A(i)$ and $F_{B+N}(t) = \sum_{i=0}^{M} f_B(i) F_N(i - t)$. This completes the proof. □

Fixing the public test, we then get this corollary, from which we can calculate bounds on the cost of privacy, thought of as the increase in the amount of data needed compared to the non-private test.

Corollary 3.5. *Suppose that a public hypothesis test τ requires at most n data points to achieve θ power at a significance level α_0 for any choice of the data. Then, in order for the private test with privacy parameter ε to achieve ρ power at a significance level α, the necessary number of data points is bounded above by $n\tilde{m}$, where $\tilde{m}$ is the smallest m such that*

$$\rho \leq \mathcal{P}(\varepsilon, \alpha, m, \alpha_0, \theta)$$

Proof. Consider a database partitioned into $\tilde{m}$ subsets, each of size n. When running the public hypothesis test on each subset, the true probability that the p-value is over the threshold α_0 is some $\theta^* \geq \theta$. Consider $\mathcal{P}(\varepsilon, \alpha, \tilde{m}, \alpha_0, \theta^*)$. Since the distribution of Z under the alternative hypothesis will shift further away from the null distribution (its center is now $\tilde{m}\theta^* \geq \tilde{m}\theta \geq \tilde{m}\alpha_0$), it follows that

$$\mathcal{P}(\varepsilon, \alpha, \tilde{m}, \alpha_0, \theta^*) \geq \mathcal{P}(\varepsilon, \alpha, \tilde{m}, \alpha_0, \theta).$$

Now let m' be the smallest m such that

$$\rho \leq \mathcal{P}(\varepsilon, \alpha, m, \alpha_0, \theta^*).$$

Since $\mathcal{P}$ is strictly increasing as a function of m, it follows that $m' \leq \tilde{m}$.

Now consider the true number of datapoints required for the private test to achieve ρ power. I.e., the minimum number of datapoints the test can achieve full power over all choices of α_0 and m. Formally, we define

$$m^* = \arg\min_m \{\rho \leq \mathcal{P}(\varepsilon, \alpha, m, \alpha_0, \theta^*) \mid m \in \mathbb{N}, \alpha_0 \in [0, 1]\}.$$

Then since the test can achieve ρ power with $m'n$ datapoints and m^*n is the minimum number of datapoints required to achieve ρ power, it follows that $m'n \geq m^*n$. Combining this with the early inequality gives $\tilde{m}n \geq m^*n$. □

A.2. Proofs for Section 4

The results of (Awan & Slavković, 2018) can be used to show that ToT has higher power than the PB framework.

Theorem 4.1. *For any choice of public test τ and privacy parameter ϵ, the statistical power of the private test resulting from the test of tests framework will be higher than that resulting from the PB framework.*

Proof. Fix a number of subtests m and subtest significance threshold α_0. Then the higher power for test of tests is an immediate consequence of the main result of Awan and Slavković (Awan & Slavković, 2018). Up until the end of the for-loop in Algorithm 1, the two frameworks are identical, and the remainder of the algorithm can be viewed as a binomial test for whether the proportion of "reject" decisions in subtests is greater than α_0. Because the Awan and Slavković binomial test is proven to be the uniformly most powerful test in this situation, it must be higher power than the actions performed by the PB framework.

Allowing test of tests to use it's own optimal m and α_0 (rather than matching that of PB) can only increase the gap between the two frameworks, since the optimal values might differ. □

Here we compute an analytic expression for the power of the test proposed by (Peña & Barrientos, 2022).

Theorem A.3. *Let θ be the power of the public test τ with significance level α_0. Let $f_{i,p,m}$ be the probability mass function of a Poisson-binomial distribution with a success probability vector of p repeated i times and $1-p$ repeated $m-i$ times. Then the power of the PB test is*

$$\mathcal{P}_{PB}(\varepsilon, \alpha, m, p, \alpha_0, \theta) = \sum_{i=0}^{m} \sum_{j=\frac{m+1}{2}}^{m} f_{i,p,m}(j) \binom{m}{i} \theta^i (1-\theta)^{m-i}.$$

Proof. Let W be the number of sub-samples in which the public test is rejected and let $\tilde{W}$ be the number in which the sub-test is rejected after the randomized response mechanism is applied. We begin by considering the probability H_0 is rejected conditional on $W = i$, which occurs if and only if $\tilde{W} > \frac{m-1}{2}$. The probability that $\tilde{W} = j$ is given by a Poisson-Binomial distribution with the vector of success probabilities

$$(\underbrace{p, \ldots, p}_{i \text{ times}}, \underbrace{1-p, \ldots, 1-p}_{m-i \text{ times}}).$$

We let $f_{i,p,m}$ denote this distribution.

Now consider the overall test. The probability that $W = i$ is given by a binomial distribution with size m and probability θ. Applying the Law of Total Probability then gives,

$$\begin{aligned} Pr\left(\tilde{W} > \frac{m-1}{2}\right) &= \sum_{i=0}^{m} Pr\left(\tilde{W} > \frac{m-1}{2} \mid W = i\right) Pr\left(W = i\right) \\ &= \sum_{i=0}^{m} \sum_{j=\frac{m+1}{2}}^{m} Pr\left(\tilde{W} = j \mid W = i\right) \binom{m}{i} \theta^i (1-\theta)^{m-i} \\ &= \sum_{i=0}^{m} \sum_{j=\frac{m+1}{2}}^{m} f_{i,p,m}(j) \binom{m}{i} \theta^i (1-\theta)^{m-i}. \end{aligned}$$

Note that the PB optimization will only select odd m, ensuring that $\frac{m+1}{2}$ is an integer. This completes the proof. □

A.3. Proofs for Section 5

Here we compute an analytic expression for the power of the public test for deviation of d-dimensional Gaussian from a given mean.

Theorem 5.1. *Let F_0 be the CDF of $\chi^2(df = d)$ and F_A be the CDF of $\chi^2\left(df = d, \lambda = n\sum_{j=1}^{d} \mu_j^2\right)$, where n is the sample size and μ_j is the j^{th} entry of $\boldsymbol{\mu}$. Then the power of the public test with significance level α is $1 - F_A(F_0^{-1}(1-\alpha)$.*

Proof. First, note that $\bar{X}_j \sim \mathcal{N}\left(\mu_j, \sigma = \frac{1}{\sqrt{n}}\right)$. Thus, $\sqrt{n}\bar{X}_j \sim \mathcal{N}\left(\sqrt{n}\mu_j, \sigma = 1\right)$. It follows that, for the test statistic,

$$Z = n\sum_{j=1}^{d} \bar{X}_j^2 = \sum_{j=1}^{d} (\sqrt{n}\bar{X}_j)^2 \sim \chi^2(df = d, \lambda).$$

Under the null hypothesis, $\lambda = 0$. But if $\boldsymbol{\mu} \neq \mathbf{0}$,

$$\lambda = \sum_{j=1}^{d} (\sqrt{n}\mu_j)^2 = n\sum_{j=1}^{d} \mu_j^2.$$

Let Z be the observed test statistic and P be the corresponding p-value. For significance level α, the power of the test is thus

$$\begin{aligned} Pr(P \leq \alpha) &= Pr(F_0(Z) \geq 1 - \alpha) \\ &= Pr(Z \geq F_0^{-1}(1 - \alpha) \\ &= 1 - F_A(F_0^{-1}(1 - \alpha)). \end{aligned}$$

□

Here we compute a lower bound on the Type 1 error of the CKMUZ test.

Theorem A.4. *Let* $L \sim \text{Laplace}(b)$*, where*

$$b = \left(5\Delta_\delta^G + \frac{432d}{\varepsilon} \ln \frac{nd}{\delta} \sqrt{\ln \frac{n}{\delta} \cdot \ln \frac{5}{4\delta}}\right) / \varepsilon$$

and Δ_δ^G *is as defined in Algorithm 4 of CKMUZ. Let* $Z \sim \mathcal{N}(0, \sigma = n\sqrt{2d})$*. Then the Type I Error of the CKMUZ algorithm is bounded below by*

$$1 - F_{Z+L}\left(\frac{n^2\gamma^2}{324}\right).$$

Proof. We begin with Stage 2 of CKMUZ's Algorithm 4. In Stage 1, any condition that fails results in a rejection of the null hypothesis, which implies that the Type I Error resulting from the final steps is in a lower bound on the overall Type I Error. In Stage 2, for each row $j \in \{1, \ldots, n\}$, the algorithm will either draw $\hat{X}^{(j)}$ from $\mathcal{N}_d(\mathbf{0}, \mathbb{I}_d)$ or set $\hat{X}^{(j)} = X^{(j)}$, the original row. But under the null hypothesis, $X^{(j)} \sim \mathcal{N}_d(\mathbf{0}, \mathbb{I}_d)$, so either way $\hat{X}^{(j)} \sim \mathcal{N}_d(\mathbf{0}, \mathbb{I}_d)$.

Under the null hypothesis, then, $T(\hat{X}) = T(X)$. As a consequence of the author's Theorem B.2, under the null hypothesis $T(\hat{X}) \sim \mathcal{N}(0, \sigma = n\sqrt{2d})$. In stage 3, the algorithm then adds noise from $L \sim \text{Laplace}(b)$ to $T(X)$, where

$$\begin{aligned} b &= \left(5\Delta_\delta^G + \frac{432d}{\varepsilon} \ln \frac{nd}{\delta} \sqrt{\ln \frac{n}{\delta} \cdot \ln \frac{5}{4\delta}}\right) / \varepsilon \\ \Delta_\delta^G &= 144\left(d \ln \frac{d}{\delta} + \frac{d}{n\varepsilon^2} \ln^2 \frac{1}{\delta} + \sqrt{nd}\sqrt{\ln \frac{d}{\delta} \cdot \ln \frac{n}{\delta}} + \frac{\sqrt{d}}{\varepsilon} \ln \frac{1}{\delta} \sqrt{\ln \frac{n}{\delta}}\right) \ln \frac{nd}{\delta}. \end{aligned}$$

The null hypothesis is then rejected if and only if $T(\hat{X}) + L > \frac{n^2\gamma^2}{324}$. Letting $Z \sim \mathcal{N}(0, \sigma = n\sqrt{2d})$, the Type I Error of the test is then bounded below by

$$\begin{aligned} P\left(T(\hat{X}) + L > \frac{n^2\gamma^2}{324} \mid H_0\right) &= 1 - P\left(Z + L \leq \frac{n^2\gamma^2}{324}\right) \\ &= 1 - F_{Z+L}\left(\frac{n^2\gamma^2}{324}\right). \end{aligned}$$

□

B. Additional Figures

Here we include additional figures.

B.1. Peña-Barrientos Framework

Figure 5 presents a comparison of the theoretical power of the binomial tests proposed by (Awan & Slavković, 2018) and PB. Figures 6 and 7 present comparisons of the PB framework and the test of tests in additional settings.

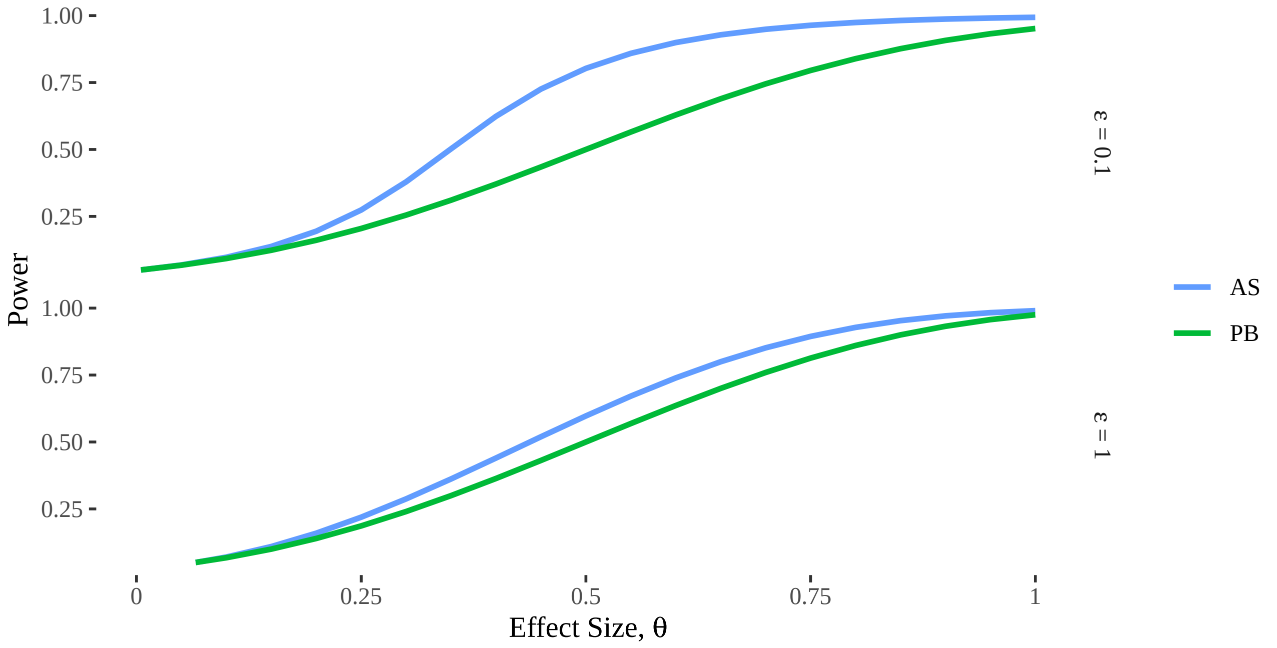

Figure 5. Let m and α_0 be the parameters selected for the PB test using the "low-power" setting recommendations for $\alpha = 0.05$ and a given ε. The plot presents the power of a test of the hypotheses $H_0 : \theta \leq \alpha_0$ and $H_A : \theta > \alpha_0$ for data $\mathbf{x} \sim \text{Binom}(m, \theta)$ as a function of θ for the binomial tests proposed in (Awan & Slavković, 2018) (AS) and PB. The left panel presents $\varepsilon = 0.1$, and the right panel presents $\varepsilon = 1$.

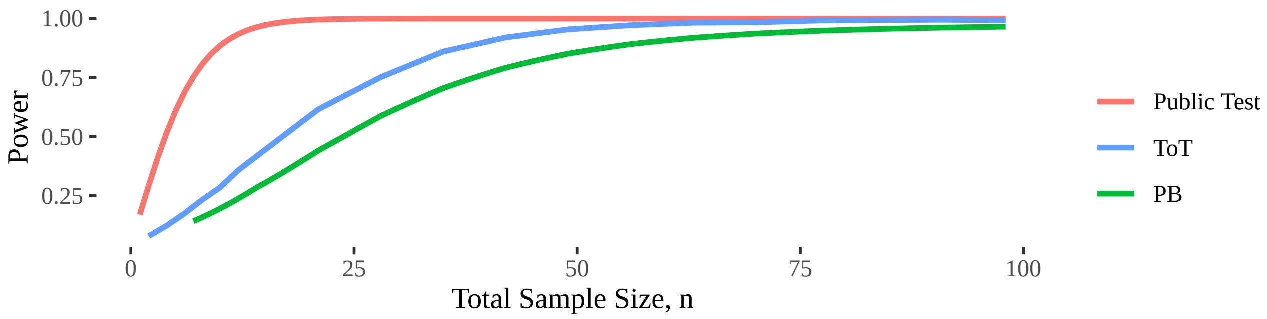

Figure 6. Power comparison between PB and the test of tests for various choices of database size n with $\alpha = 0.05$ and a z-test. The effect size is $\mu/\sigma = 1$ and privacy parameter is $\varepsilon = 1$. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target power $\rho = 0.9$.

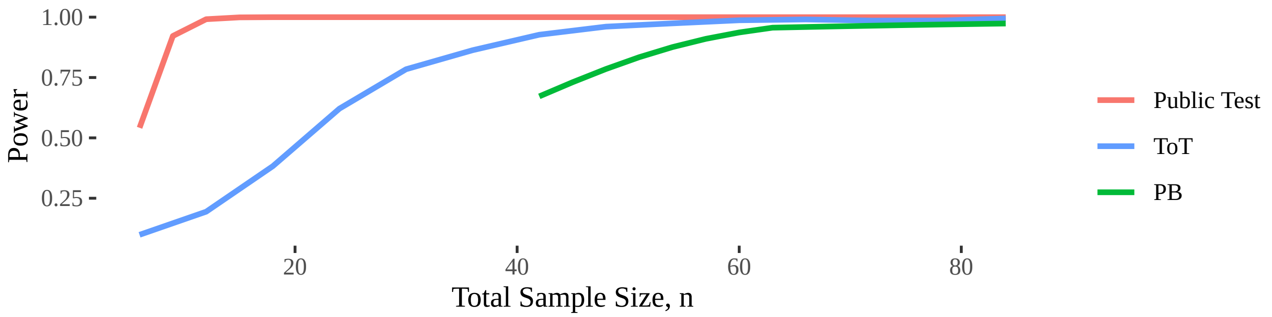

Figure 7. Power comparison between PB and the test of tests for various choices of database size n with $\alpha = 0.05$ and a one-way ANOVA with non-private groups. The effect size is $\eta = 4$ and privacy parameter is $\varepsilon = 1$. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target power $\rho = 0.9$.

B.2. Multivariate Normal Data

Figure 8 compares the empirical Type I Error of the two tests in Figure 3. Figures 9 to 12 provide additional power comparisons between CKMUZ and ToT with various dimensions d, effect sizes $\boldsymbol{\mu}$, and privacy parameters ε.

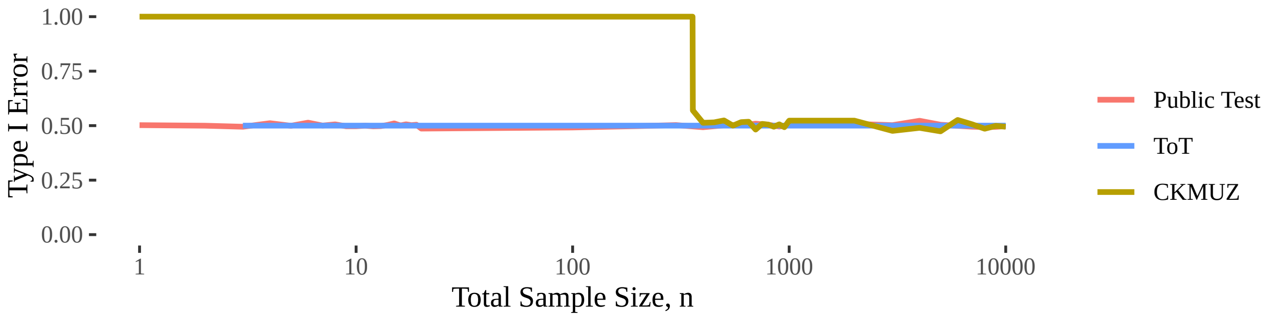

Figure 8. Type I error check between CKMUZ and the tests of tests for various sample sizes n. The dimension is $d = 100$, $\varepsilon = 1$, and α for the test of tests and public test is set to be 0.5. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$.

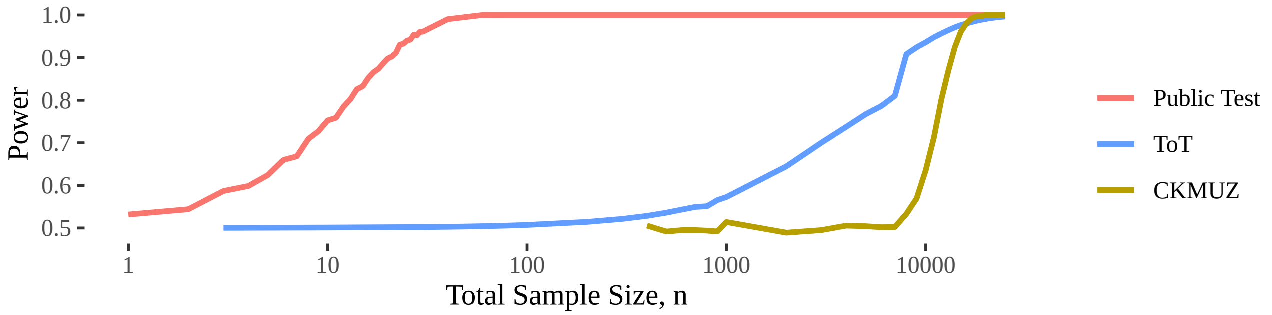

Figure 9. Power comparison between CKMUZ and the tests of tests for various sample sizes n. The true mean is $\mu_1 = 0.1$ and $\mu_i = 0$ for all $i \neq 1$. The dimension is $d = 100$, $\varepsilon = 1$, and α for the test of tests and public test is set to match the Type I Error of the CKMUZ test. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$.

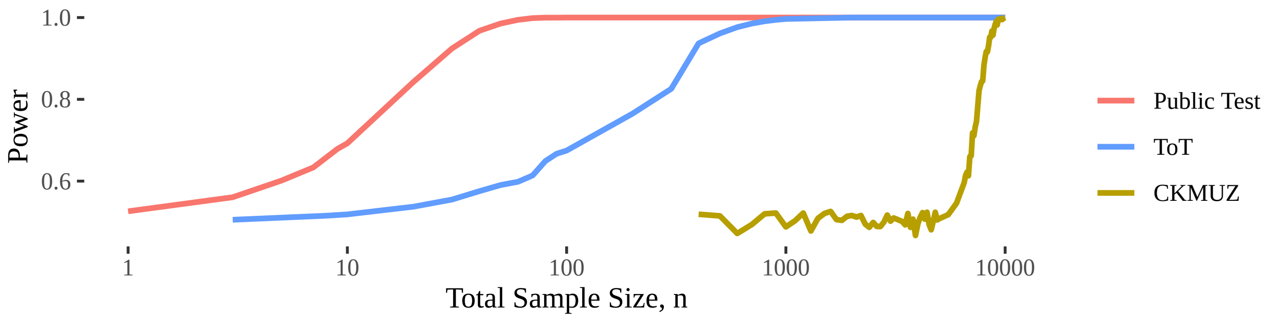

Figure 10. Power comparison between CKMUZ and the tests of tests for various sample sizes n. The true mean is $\mu_i = 0.1$ for $i <= 20$ and $\mu_i = 0$ otherwise. The dimension is $d = 60$, $\varepsilon = 1$, and α for the test of tests and public test is set to match the Type I Error of the CKMUZ test. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$.

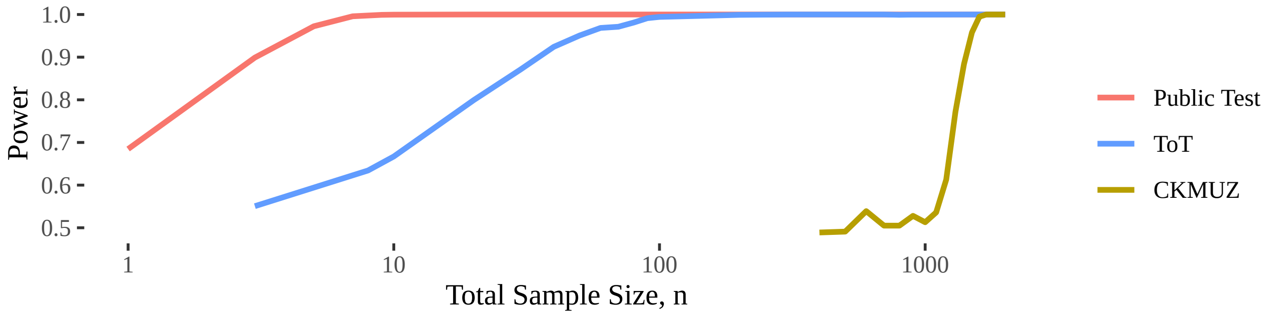

Figure 11. Power comparison between CKMUZ and the tests of tests for various sample sizes n. The true mean is $\mu_i = 0.3$ for $i <= 20$ and $\mu_i = 0$ otherwise. The dimension is $d = 60$, $\varepsilon = 1$, and α for the test of tests and public test is set to match the Type I Error of the CKMUZ test. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$ for $n \leq 50$ and $\rho = 0.99$ for $n > 50$.

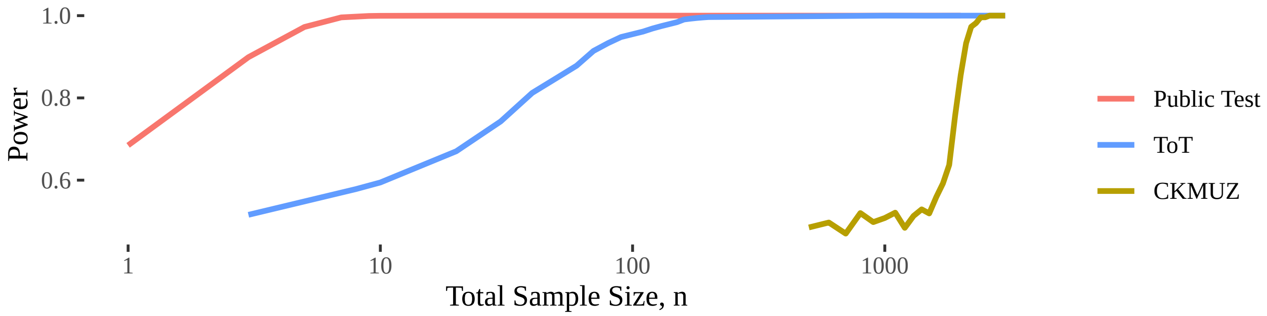

Figure 12. Power comparison between CKMUZ and the tests of tests for various sample sizes n. The true mean is $\mu_i = 0.3$ for $i <= 20$ and $\mu_i = 0$ otherwise. The dimension is $d = 60$, $\varepsilon = 0.5$, and α for the test of tests and public test is set to match the Type I Error of the CKMUZ test. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$ for $n \leq 100$ and $\rho = 0.99$ for $n > 100$.

B.3. One-way Analysis of Variance

Figures 13 to 16 give more comparisons between CKSBG, SGGRGB, and ToT for various choices of parameters. Figures 13 and 14 are comparisons with $g = 2$ groups with both effect sizes, while Figures 15 and 16 are comparisons with $g = 3$ groups.

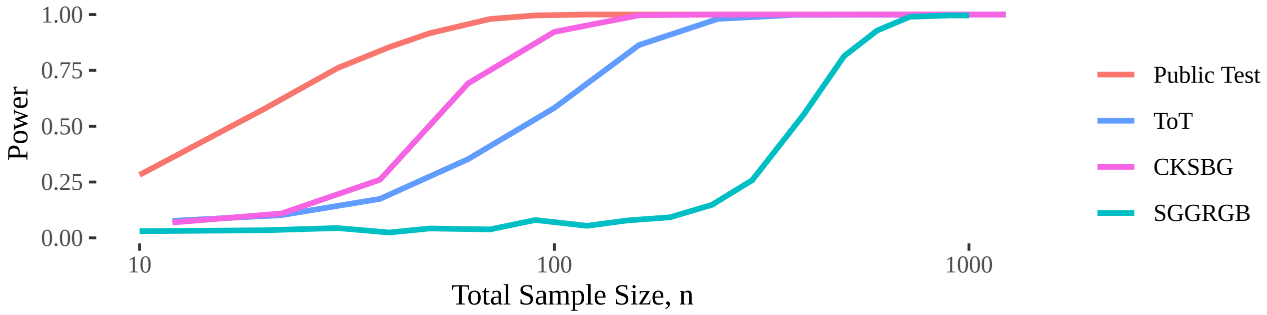

Figure 13. Power comparison between CKSBG,SGGRGB, and the test of tests for various choices of database size n. The effect size is $\eta = 0.5$, privacy parameter $\varepsilon = 1$, and number of groups $g = 2$. All groups are of equal size and $\alpha = 0.05$. All power curves (except the public test) are estimated via simulation. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$.

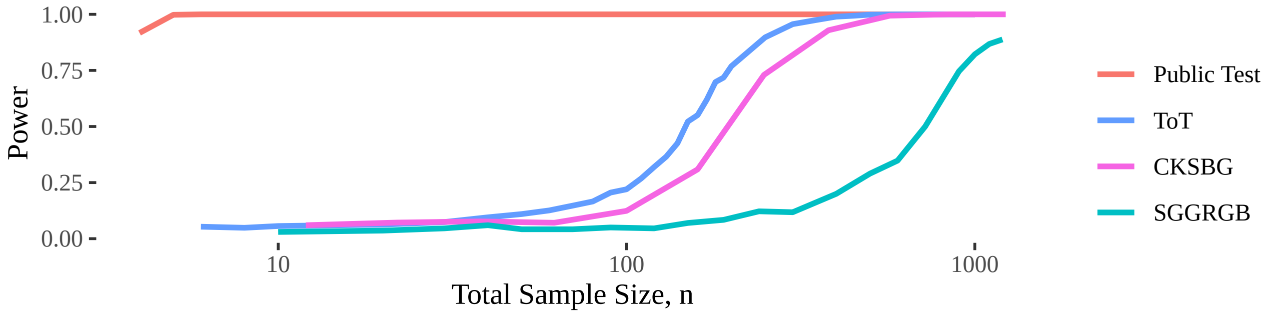

Figure 14. Power comparison between CKSBG, SGGRGB, and the test of tests for various choices of database size n. The effect size is $\eta = 25$, privacy parameter $\varepsilon = 0.1$, and number of groups $g = 2$. All groups are of equal size and $\alpha = 0.05$. All power curves (except the public test) are estimated via simulation. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$.

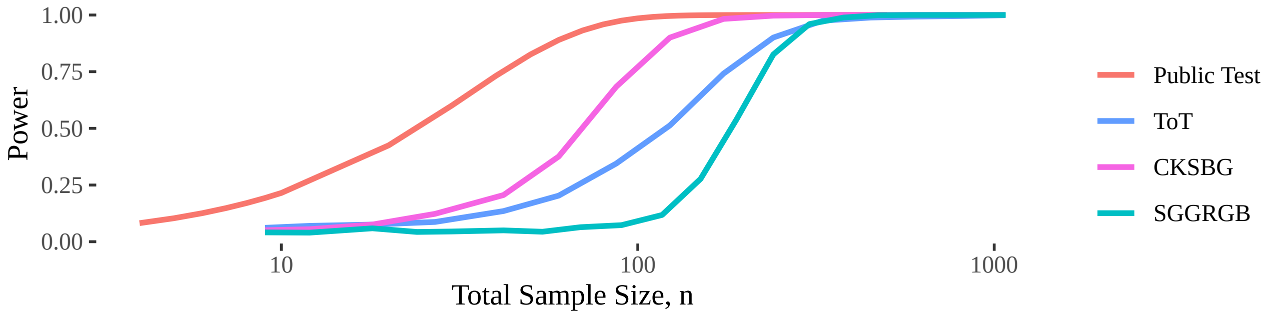

Figure 15. Power comparison between CKSBG, SGGRGB, and the test of tests for various choices of database size n. The effect size is $\eta = 0.35$, privacy parameter $\varepsilon = 1$, and number of groups $g = 3$. All groups are of equal size and $\alpha = 0.05$. All power curves (except the public test) are estimated via simulation. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$.

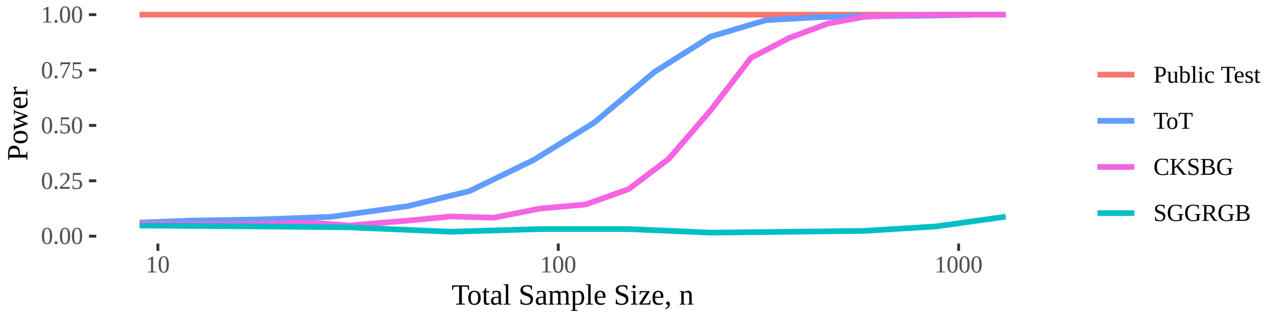

Figure 16. Power comparison between CKSBG, SGGRGB, and the test of tests for various choices of database size n. The effect size is $\eta = 25$, privacy parameter $\varepsilon = 0.1$, and number of groups $g = 3$. All groups are of equal size and $\alpha = 0.05$. All power curves (except the public test) are estimated via simulation. We optimize m and α_0 for the test of tests at each n as discussed in Section 3.4 with target $\rho = 0.9$.