Differentially Private Nonparametric Hypothesis Testing

Simon Couch
Reed College
Portland, Oregon
couchs@reed.edu

Zeki Kazan
Reed College
Portland, Oregon
kazanz@reed.edu

Kaiyan Shi
Reed College
Portland, Oregon
kaishi@reed.edu

Andrew Bray
Reed College
Portland, Oregon
abray@reed.edu

Adam Groce
Reed College
Portland, Oregon
agroce@reed.edu

ABSTRACT

Hypothesis tests are a crucial statistical tool for data mining and are the workhorse of scientific research in many fields. Here we study differentially private tests of independence between a categorical and a continuous variable. We take as our starting point traditional nonparametric tests, which require no distributional assumption (e.g., normality) about the data distribution. We present private analogues of the Kruskal-Wallis, Mann-Whitney, and Wilcoxon signed-rank tests, as well as the parametric one-sample t-test. These tests use novel test statistics developed specifically for the private setting. We compare our tests to prior work, both on parametric and nonparametric tests. We find that in all cases our new nonparametric tests achieve large improvements in statistical power, even when the assumptions of parametric tests are met.

CCS CONCEPTS

• **Mathematics of computing → Nonparametric statistics**; • **Security and privacy → Data anonymization and sanitization**.

KEYWORDS

differential privacy; hypothesis test; nonparametric

ACM Reference Format:
Simon Couch, Zeki Kazan, Kaiyan Shi, Andrew Bray, and Adam Groce. 2019. Differentially Private Nonparametric Hypothesis Testing. In *2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19), November 11–15, 2019, London, United Kingdom.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3319535.3339821

1 INTRODUCTION

In 2011, researchers in Switzerland began an investigation into the link between methylation levels of a given gene and the occurance of schizophrenia and bipolar disorder[4]. They recruited patients that suffered from psychosis as well as healthy controls and measured the level of methylation of the gene in each individual. They found that levels in the groups suffering from psychosis were higher than those in the healthy group. In order to rule out the possibility that this difference was due to sampling variability, they relied upon a suite of nonparameteric hypothesis tests to establish that this link likely exists in the general population.

While the results of these tests were published in an academic journal, the data itself is unavailable to preserve the privacy of the patients. This is a necessary consideration when working with sensitive data, but it hampers scientific reproducibility and the extension of this work by other researchers.

Our goal in this paper is to provide nonparametric hypothesis tests that satisfy differential privacy. The difficulty of developing a private test comes not just from the need to privately approximate a test statistic, but also from the need for an accurate reference distribution that will produce valid p-values. It is not sufficient to treat the approximate test statistic as equivalent to its non-private counterpart.

In this paper, we present several new hypothesis tests. In all cases we are considering data sets with a categorical explanatory variable (e.g., membership in the schizophrenic, bipolar, or control group) and a continuous dependent variable (e.g., methylation level). Our goal when designing a hypothesis test is to maximize the *statistical power* of the test, or equivalently to minimize the amount of data needed to detect a particular effect.

In the traditional public setting, there are two families of tests for these scenarios. The more commonly used are *parametric* tests that assume that within each group, the continuous variable follows a particular distribution (usually Gaussian). An alternative to these tests are *nonparametric* tests, which make no distributional assumption but, in exchange, have slightly lower power. Nonparametric tests generally rely upon substituting in, for each data point, the rank of the continuous variable relative to the rest of the sample. The test statistic is then a function of these ranks rather than the original values.

The private hypothesis tests we propose all use rank-based test statistics. Our overarching argument in this paper, beyond the individual value of each of the tests we introduce, is that in the private setting these rank-based test statistics are *more powerful* than the traditional parametric alternatives. This is contrary to the public setting, where the parametric tests (when their assumptions are met) perform better.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CCS '19, November 11–15, 2019, London, United Kingdom
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6747-9/19/11…$15.00
https://doi.org/10.1145/3319535.3339821

Our second, broader point is that as a research community we need to support the development of hypothesis tests specifically tailored to the private setting. Our private test statistics are not simply approximations of traditional test statistics from the public setting and as a result, we find that they can require an order of magnitude less data. Current tests used by statisticians in the public setting have been refined through decades of incremental improvement, and the same sort of development needs to happen in the private setting.

1.1 Our contributions

We introduce several new private hypothesis tests that mirror the three most commonly used rank-based tests. In one setting, this is a private analog of the traditional public test statistic, but for the remaining two settings it is a new statistic developed specifically for the private setting. The privacy of these statistics generally follows from non-trivial but reasonably straightforward applications of the Laplace mechanism. Our main contribution is not the method of achieving privacy but the construction of novel private hypothesis tests with high statistical power.

There are two components to the construction of each hypothesis test. The first is the creation of a test statistic to capture the effect of interest while remaining provably private. The second is a method to learn the distribution of the statistic under the null hypothesis in order to compute p-values, which are the object of primary interest to researchers.

In particular, we develop tests for the following cases:

Three or more groups. When the categorical variable divides the sample into three or more groups, the traditional public test is the one-way analysis of variance (ANOVA) in the parametric case and Kruskal-Wallace in the nonparametric case. ANOVA has been previously studied by Campbell et al. [3] and then by Swanberg et al. [28], who improved the power by an order of magnitude. We give the first private nonparametric test by modifying the rank-based Kruskal-Wallace test statistic for the private setting. We provide experimental evidence that this statistic has dramatically higher power than a simple privatized version of the public Kruskal-Wallace statistic. Moreover, we provide evidence that even in the parametric setting (i.e., when the data is normally distributed) the private rank-based statistic outperforms the private ANOVA test, in one representative case requiring only 23% as much data to reach the same power.

Two groups. In the two group setting, the most common public nonparametric test is the Mann-Whitney test. We provide an algorithm to release an approximation of this statistic under differential privacy and a second algorithm to conduct the test and release a valid p-value.

In the public setting, the analogous parametric test is the two sample t-test which is equivalent to an ANOVA test when there are two groups. We compare therefore to the private ANOVA test and show experimental evidence that our Mann-Whitney analog has significantly higher power.

Paired data. We also consider the case where the two sets of data are in correspondence with each other (e.g., before-and-after measurements). The nonparametric test in this case is the Wilcoxon signed-rank test, and it is the only nonparametric test that has previously been studied in the private setting[30]. We provide two improvements to the prior work. First, we change the underlying statistic to the less well-known Pratt variant, which we find conforms more easily to the addition of noise. Second, we show that our simulation method for computing reference distributions is more precise than the upper bounds used in the prior work (which contained an error which we identify and correct). The result is a significant improvement in the power of the test over the the prior work.

In parallel with the previous two scenarios, we then compare our private nonparametric test with a private version of the analogous parametric test, the paired t-test. A direct private implementation of this test does not exist in the literature, so we propose one here.[1] In alignment with the previous results, we show experimental evidence that the rank-based test has superior power.

For all our tests we give not just private test statistics but also precise methods of computing a reference distribution and a p-value, the final output practitioners actually need. We also experimentally verify that the probability of Type 1 errors (incorrectly rejecting the null hypothesis) is acceptably low. We give careful power analyses and use these to compare tests to each other. All our tests and experiments are implemented with publicly available code.[2]

We find that rank-based statistics are very amenable to the private setting. We also repeatedly find that what is optimal in the public setting is no longer optimal in the private setting. We hope that this work contributes to the development of a standard set of powerful hypothesis tests that can be used by scientists to enable inferential analysis while protecting privacy.

2 BACKGROUND

In this section, we begin by discussing hypothesis testing in general and outlining the formalities of differential privacy. We then discuss the difficulties of hypothesis testing within the privacy framework and the previous work done in this area. Each of our main results requires a more detailed discussion of prior work on that particular test or use case; we leave those discussions for Sections 3-5.

2.1 Hypothesis Testing

The key inferential leap that is made in hypothesis testing is the claim that not only is the *sample* of data incompatible with a particular scientific theory, but that the incompatiblility holds in a broader *population*. In the study on psychotic disease, the researchers used this technique to generalize from their 165 subjects to the population of causasian-descended Swiss. The scientific theory that they refuted, that there is no link between methylation levels and psychosis, is called the *null hypothesis* (H_0).

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

[1] In simultaneous work Gaboardi et al. [13] propose such a test, but our test is still higher power. See Section 5.5 for more details.

[2] Our source code is available at: *github.com/simonpcouch/non-pm-dpht*

that the observed test statistic or more extreme would occur under H_0.

Definition 2.1. For a given test statistic $t = f(\mathbf{x})$ and null hypothesis H_0, the *p-value* is defined as

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

If the function f is well-chosen, the underlying distribution of $\mathbf{X}$ will differ more from the distribution under H_0 and a low p-value becomes more likely. Typically a threshold value α is chosen, such that we reject H_0 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.

We define the *critical value* t^* to be the value of the test statistic t where $p = \alpha$. We use this to define the *statistical power*, a measure of how likely a hypothesis test is to pick up on a given effect (i.e. the chance of rejecting when the null hypothesis is false). 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* of a hypothesis test is

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

Statistical power is the accepted metric by which the statistics community judges the usefulness of a hypothesis test. It provides a common scale upon which to evaluate different tests for the same use case.

2.2 Differential Privacy

To persuade people to allow their personal data to be collected, data owners must protect information about specific individuals. Historically, ad-hoc database anonymization techniques have been used (i.e. changing names to numeric IDs, rounding geospatial coordinates to the nearest block, etc.), but these methods have repeatedly been shown to be ineffective [14, 19, 29].

Differential privacy, proposed by Dwork et al. in 2006 [9], is a mathematically robust definition of privacy preservation, which guarantees that a query does not reveal anything about an individual as a consequence of their presence in the database. When the condition of differential privacy is satisfied, there is not *much* difference between the output obtained from the original database, and that obtained from a database that differs by only one individual's data. Here we present (ϵ, δ)-differential privacy [8], which allows the closeness of output distributions to be measured with both a multiplicative and an additive factor. However, for most of our hypothesis tests $\delta = 0$.

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 in one row:

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

We call databases $\mathbf{x}$ and $\mathbf{x}'$ *neighboring* if they differ only in that a single row is altered (but not added or deleted), and we will use this notation in the following sections.[3]

[3]This is one of two roughly equivalent variants of differential privacy. The key difference is that under this definition the size of the database is public knowledge.

Differential privacy, like any acceptable privacy definition, is resistant to post processing. That is, if an algorithm is differentially private, an adversary with no access to the database will be unable to violate such privacy through further analysis (e.g., attempted deanonymization) of the query output [9].

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.*

Theorem 2.4 has another useful consequence. It allows us to develop private algorithms by first computing some private output, and then carrying out further computation on that output without accessing the database. The additional computation need not be analyzed carefully—the final output of the additional analysis is automatically known to retain privacy.

Differential privacy requires the introduction of some randomness to any query output. A frequently used method is the *Laplace mechanism*, introduced by Dwork et al. [9]. When given an arbitrary algorithm f with real-valued output, this mechanism will add some noise drawn from the Laplace distribution to the output of the algorithm and release a noisy output.

Definition 2.5 (Laplace Distribution)**.** The Laplace Distribution centered at 0 with scale b is the distribution with probability density function:

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

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

The scale of the Laplace Distribution used to produce the noisy output depends on the global sensitivity of the given algorithm f, which is the maximum change on the output of f that could result from the alteration of a single row.

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 neighbouring databases.

With computed sensitivity GS_f and privacy parameter ϵ, the Laplace mechanism applied to f ensures $(\epsilon, 0)$-differential privacy [9].

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 $\text{Lap}(GS_f/\epsilon)$, and GS_f is the global sensitivity of f.

Theorem 2.8. *(Laplace Mechanism) The Laplace mechanism preserves $(\epsilon, 0)$ differential privacy.*

Global sensitivity is the maximum effect that can be caused by changing a single row of any database. Sometimes it is helpful to talk about local sensitivity *for a given database $\boldsymbol{x}$* [21]. This is the maximum effect that can be caused by changing a row of that particular database.

Definition 2.9 (Local Sensitivity)**.** The sensitivity of a function f at a particular database $\mathbf{x}$ is:

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

where $\mathbf{x}'$ is a neighboring database.

Note that $GS_f = \max_{\mathbf{x}} LS_f(\mathbf{x})$. Local sensitivity cannot simply be used in the Laplace mechanism in place of global sensitivity, because local sensitivity itself is a function of the database and therefore cannot be released. But private upper bounds on local sensitivity can be used to create similar mechanisms that do preserve privacy, and one of our algorithms uses just such a technique.

Choosing ϵ is an important consideration when using differential privacy. We consider several values of ϵ throughout our power analyses. The lowest, .01 is an extremely conservative privacy parameter and allows for safe composition with many other queries of comparable ϵ value. We also use ϵs of .1 and 1, which, while higher, still provide very meaningful privacy protection. Ultimately, the choice of ϵ is a question of policy and depends on the relative importance with which privacy and utility are regarded. We also measure, for comparison, the power of the public versions of each test (equivalent to an ϵ of ∞). As one might expect, the amount of data needed to detect a given effect often scales roughly with the inverse of ϵ.

2.3 Differentially Private Hypothesis Testing

Performing hypothesis tests within the framework of differential privacy introduces new complexity. A function to compute a private test statistic (be it a private version of a standard test statistic or an entirely new test statistic) is not useful on its own. We need a p-value or other understandable output, and that means understanding the reference distribution (i.e., the distribution of the statistic given H_0).

In classical statistics, test statistics are computed with deterministic functions. The randomness added to the test statistic in order to privatize it introduces new complexity. Most importantly, it causes the reference distribution to change. One *cannot* simply compare the private test statistic to the usual reference distribution, as the addition of noise can inflate the type I error well above acceptable levels [3].

Because of this, a complete differentially private hypothesis test requires not only a function for computing a private test statistic, but also a method for determining its null distribution. Often the exact reference distribution cannot be determined, so worst-case reference distributions or upper bounds on the resulting critical value must be used, and the precision of this reference distribution can have a large effect on the resulting power.

The goal of differentially private hypothesis test design is to develop a test with power as close as possible to the public test.

2.4 Related Work

There is a substantial and growing literature on differentially private hypothesis testing. One area of research is the study of the rate of convergence of private statistics to the distributions of their public analogues [25, 26, 34]. These papers do not offer practical, implementable tests and discussion of reference distributions when the noise is not yet negligible is often limited or entirely absent. Further, the results are often entirely asymptotic, without regard for constants that may prove to be problematic.

The chi-squared test, which tests the independence of two categorical variables,[4] has been the subject of much study, resulting in the development of many private variants. One of these works, that of Vu and Slavkovic [32], provides methods for calculation of accurate p-values adjusted for the addition of Laplace noise for differentially private single proportion and chi-squared tests specifically for clinical trial data. Several other papers, though they make asymptotic arguments on the uniformity of their p-values, have developed frameworks for private chi-squared tests specifically for the intent of genome-wide association study (GWAS) data [11, 15, 31]. For these same tests, Monte Carlo simulation has been shown to offer more precise analysis in some cases [12, 33]. There has also been work, like that of Rogers and Kifer [23], that proposes entirely new test statistics with asymptotic distributions more similar to their public counterparts.

While the development of private test statistics has achieved much attention, careful evaluations of statistical power of these new test statistics is not always demonstrated. This is unfortunate, as the cost of privacy (utility loss) must be accurately quantified in order for the widespread adoption or implementation of any of these methods. Fortunately, rigorous power analysis seems to be more common in recent work. Awan and Slavkovic recently presented a test for simple binomial data [1]. While the setting is the simplest possible, their paper gives what we believe is the first private test to come with a proof of optimality, something normally very difficult to achieve even in the public setting.

The body of work on numerical (rather than categorical) methods is less extensive but has been growing quickly in recent years. In 2017, Nguyen and Hui proposed algorithms for survival analysis methods [20]. There have been frameworks developed for testing the difference in means of normal distributions [6, 7], and for testing whether a sample is consistent with a normal distribution with a particular mean [27]. Differentially private versions of linear regressions, a class of inference that is extremely common in many fields both within and outside of academia, have received a notable level of attention, but the treatment of regression coefficients as test statistics has come about only recently [2, 24]. Two works have studied differentially private versions of one-way analysis of variance (ANOVA) [3, 28]. The only prior work done on nonparametric hypothesis tests, as far as we are aware, is on the Wilcoxon signed-rank test by Task and Clifton in 2016 [30]. Prior work specifically relevant to the tests we are proposing will be discussed in more detail in the relevant section.

3 MANY GROUPS

We first consider the most general case, where we wish to distinguish whether many groups share the same distribution on a continuous variable. The standard parametric test in the public setting is the one-way analysis of variance (ANOVA), which tests the equality of means across many groups. Private ANOVA has been studied previously first by Campbell et al. [3] and then by

[4] This is the chi-squared test of independence. There are several related tests that use the same statistic, the chi-squared.

Swanberg et al. [28], who improved the power by an order of magnitude. The standard nonparametric test in the public setting is the Kruskal-Wallis test, which was used by the pschosis research group to determine that subjects in the schizophrenia, bipolar, and control groups had different methylation levels at a particular gene site[4]. As is standard for nonparametric statistics in the public setting, it sacrifices some power compared to ANOVA but no longer assumes normally distributed data. [10]

In this section we present two tests. The first is a straightforward privatization of the standard Kruskal-Wallis test statistic. The second modifies the statistic, essentially by linearizing the implied distance metric. We find first that our modified statistic has much higher power. We then further show that our modified statistic has much higher power than the ANOVA test of Swanberg et al. *even when the data is normally distributed.*

3.1 The Kruskal-Wallis test

The Kruskal-Wallis test, proposed by William Kruskal and W. Allen Wallis in 1952 [17], is used to determine if several groups share the same distribution in a continuous variable. The only assumptions are that the data are drawn randomly and independently from a distribution with at least an ordinal scale.

Take a database $\mathbf{x}$ with g groups[5] and n rows. Let n_i be the size of each group and r_{ij} be the rank of the j^{th} element of group i. (If values are equal for several elements, all are given a rank equal to the average rank for that set.) We define $\bar{r}_i$ to be $\frac{1}{n_i}\sum_{j=1}^{n_i} r_{ij}$, the mean rank of group i, and $\bar{r}$ to be $\frac{n+1}{2}$, the average of all the ranks. Then, the Kruskal-Wallis h-statistic is defined to be

$$h = (n-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}.$$

If there are no ties in the database, the denominator is constant and the formula can be simplified to

$$h = \frac{12}{n(n+1)}\sum_{i=1}^{g} n_i \bar{r}_i^2 - 3(n+1).$$

For clarity and consistency with later sections, we present this calculation as an algorithm. In general we use a subscript "stat" to label the algorithm computing a test statistic and a subscript "p" to denote the fully hypothesis test that outputs a p-value. We use tildes to indicate private algorithms.

Algorithm $\mathrm{KW}_{\text{stat}}$ **:** Kruskal-Wallis Test Statistic

Input: $\mathbf{x}$
for *group i of $\mathbf{x}$* **do**
  $\bar{r}_i \longleftarrow \left(\sum_{j=1}^{n_i} r_{ij}\right)/n_i$
$h \longleftarrow \frac{12}{n(n+1)}\sum_{i=1}^{g} n_i\bar{r}_i - 3(n+1)$
Output: h

[5]Throughout the paper we assume g is public and independent of the data, so we do not list it as a separate input. Because g is the number of *valid* groups, one or more of the g groups might not contain any observations. Allowing many valid groups that have no actual observations artificially increases the critical value, so it can reduce the power of our tests but does not affect the validity or privacy of the output.

3.2 Privatized Kruskal-Wallis

In this section, we bound the sensitivity of $\mathrm{KW}_{\text{stat}}$, allowing us to create a private version. We then present a complete algorithm for calculating a p-value and prove that it too is differentially private. We begin with the following sensitivity claim (see the the full version for the proof).

Theorem 3.1. *The sensitivity of* $\mathrm{KW}_{\text{stat}}$ *is bounded by 87.*

We are using the simplified formula that assumes there are no ties in the data, so our algorithm begins by adding a small amount of random noise to each data point to randomly order any ties. We may then compute the h-statistic as in the public setting and add noise proportional to the sensitivity.

Algorithm $\widetilde{\mathrm{KW}}_{\text{stat}}$ **:** Private Kruskal-Wallis Test Statistic

Input: $\mathbf{x}$, ϵ
Rank all data points, randomly breaking ties
$h \longleftarrow \mathrm{KW}_{\text{stat}}(\mathbf{x})$
$\widetilde{h} \longleftarrow h + \mathrm{Lap}\left(87/\epsilon\right)$
Output: $\widetilde{h}$

Theorem 3.2. *Algorithm* $\widetilde{\mathrm{KW}}_{\text{stat}}$ *is ϵ-differentially private.*

See the full version for the proof.

Algorithm $\widetilde{\mathrm{KW}}_{\text{p}}$ **:** Complete Kruskal-Wallis Test

Input: $\mathbf{x}$, ϵ, z
$\widetilde{h} \longleftarrow \widetilde{\mathrm{KW}}_{\text{stat}}(\mathbf{x}, \epsilon)$
for $k = 1$ to z **do**
  $\mathbf{x}^* \longleftarrow$ a database with independent uniform values from [0,1], divided almost equally into g groups
  $h_k \longleftarrow \widetilde{\mathrm{KW}}_{\text{stat}}(\mathbf{x}^*)$;
$p \longleftarrow$ fraction of h_k values greater than $\widetilde{h}$
Output: $\widetilde{h}, p$

Algorithm $\widetilde{\mathrm{KW}}_{\text{p}}$ is our complete algorithm to find a p-value given a database $\mathbf{x}$, privacy parameter ϵ. First a private test statistic $\widetilde{h}$ is computed. Then the reference distribution is approximated by simulating z databases under H_0 and computing the test statistic for each.[6] (The distribution of the test statistic is independent of the distribution of data between groups and the distribution of the i.i.d. data points, so our choice of equal-sized groups and uniform data from $[0, 1]$ is arbitrary.) The p-value is the percent of h_k more extreme than $\widetilde{h}$.

Theorem 3.3. *Algorithm* $\widetilde{\mathrm{KW}}_{\text{p}}$ *is ϵ-differentially private.*

Proof. By Theorem 3.2, the computation of $\widetilde{h}$ is ϵ-differentially private. All of the following steps (generating the reference distribution and calculating p-value) do not need to access to the database $\mathbf{x}$, and therefore by Theorem 2.4 (post processing), Algorithm $\widetilde{\mathrm{KW}}_{\text{p}}$ is ϵ-differentially private. □

[6]When we use the traditional Kruskal-Wallis test, the distribution of h-statistics asymptotically converges to the χ^2 distribution. Thus, for efficiency purposes, we sample h_k from $\chi^2(g-1) + \mathrm{Lap}(\Delta h/\epsilon)$

3.3 A New Test: Absolute Value Kruskal-Wallis

We now introduce our own new test, specifically designed for the private setting. Inspired by Swanberg et al. [28], we alter the Kruskal-Wallis statistic, measuring distance with the absolute value instead of the square of the differences. This statistic is now

$$h_{abs} = (n-1)\frac{\sum_{i=1}^{g} n_i|\bar{r}_i - \bar{r}|}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}|r_{ij} - \bar{r}|}.$$

As before, when there are no ties in the data, the statistic can be simplified. (See the full version for the calculation.) In this case, the form depends on the parity of n.

$$h_{abs} = \begin{cases} \frac{4(n-1)}{n^2}\sum_{i=1}^{g} n_i\left|\bar{r}_i - \frac{n+1}{2}\right|, & \text{if } n \text{ is even} \\ \frac{4}{n+1}\sum_{i=1}^{g} n_i\left|\bar{r}_i - \frac{n+1}{2}\right|, & \text{if } n \text{ is odd} \end{cases}$$

We call the algorithm to calculate the h_{abs} test statistic $\text{KWabs}_{\text{stat}}$. This statistic is preferable for two reasons. First, it has lower sensitivity. The following theorem is proved in the full version.

Theorem 3.4. *The sensitivity of* $\text{KWabs}_{\text{stat}}$ *is bounded by 8.*

Second, the actual values for h_{abs} are significantly higher than they are for h, so any given amount of noise is less likely to overwhelm the value.

Because of space constraints, we don't give pseudocode for this hypothesis test, but it follows exactly that of the previous test. The privatized statistic is computed by $\widetilde{\text{KWabs}}_{\text{stat}}$, which adds Laplace noise in the same way as for $\widetilde{\text{KW}}_{\text{stat}}$, but scaled to the lower sensitivity. The full hypothesis test, $\widetilde{\text{KWabs}}_{\text{p}}$, computes the p value in the same way as was done for $\widetilde{\text{KW}}_{\text{p}}$.[7]

Theorem 3.5. *Algorithm* $\widetilde{\text{KWabs}}_{\text{stat}}$ *and* $\widetilde{\text{KWabs}}_{\text{p}}$ *are* ϵ*-differentially private.*

Proof. The proof is identical to the proofs for Algorithms $\widetilde{\text{KW}}_{\text{stat}}$ and $\widetilde{\text{KW}}_{\text{p}}$ (Theorems 3.2 and 3.3). □

Unequal group sizes. The traditional h statistic (and therefore the noisy private analogue) has a reference distribution that is independent of the allocation of observations between groups. This is unfortunately not true for our new h_{abs} statistic. Fortunately, it seems that the worst-case distribution (i.e., the one resulting in the highest critical value) occurs when all groups are of equal size. (We present both theoretical and experimental evidence for this in the full version.) As a result, it is safe to always equal-sized groups when simulating a reference distribution, though for very unequal group sizes, there will be a significant loss in power compared to a hypothetical where group sizes were known. (If approximate group sizes are known publicly or released through other queries, those could be used instead when simulating the reference distribution.)

[7] Unlike before, a χ^2 approximation cannot be used.

3.4 Experimental Results

Power analysis. We now assess the power of our $\widetilde{\text{KWabs}}_{\text{p}}$ test on synthetic data (See the full version for an application to real-world data.) We generate many databases of data distributed with specified parameters and then run $\widetilde{\text{KWabs}}_{\text{p}}$ on each. The power of the test for a given set of parameters is the proportion of times $\widetilde{\text{KWabs}}_{\text{p}}$ returns a significant result (i.e. a p-value less than the significance level α, generally set at 0.05). We use three groups of normally-distributed data, separated by steps of one standard deviation (so the highest and lowest groups differ by two standard deviations). In our captions we denote the mean of group i with μ_i.

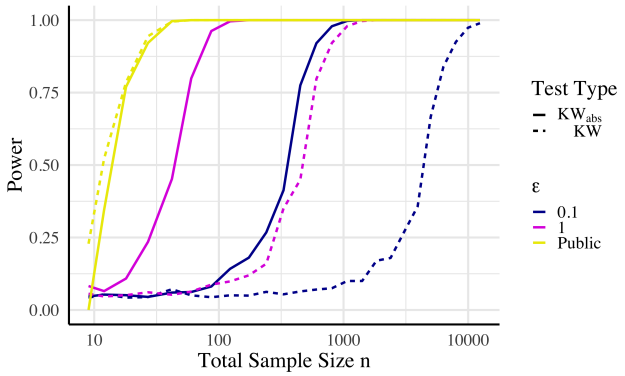

Figure 1: Power of $\widetilde{\text{KWabs}}_{\text{p}}$ at various values of ϵ and total sample size n. (Effect size: $max_i(\mu_i) - min_i(\{\mu_i\}) = 2\sigma$; $g = 3$; $\alpha = .05$; equal group sizes; normally distributed sample data)

As shown in Figure 1, our private absolute value test variant requires significantly less data points than the original private test to reach the same power. Thus, from here on, we only evaluate the power of the absolute value variant. Figure 1 also shows that, at an ϵ of 1, our private absolute value test only requires a database around a factor of 3 larger than the public test needs.

Uniformity of p-values. If a test is correctly designed, the probability of type 1 error (i.e., rejecting the null hypothesis when it is correct) should be less than or equal to α for any chosen value of α. Comparing the fit of a large number of simulated p-values generated from null distributions to the uniform distribution on the unit interval allows one to evaluate the uniformity of p-values for a given hypothesis test. A common tool to carry this procedure out, the quantile-quantile (or Q-Q) plot, plots the quantiles of the uniform, theoretical distribution against the quantiles of the p-values. The theoretical and emperical quantiles will be nearly equal at all quantiles when the p-values follow the theoretical distribution, resulting in a linear trend on the Q-Q plot. A convex Q-Q plot indicates an increase in the type II error rate (i.e. the test not rejecting the null hypothesis when it is indeed not true, causing a decrease in power) which is acceptable but undesirable, while a concave Q-Q plot indicates an exceedingly high type I error rate (i.e. the test rejecting the null hypothesis when it is true, causing undue increases in power) which is not acceptable. Figure 2 demonstrates the p-value uniformity of $\widetilde{\text{KWabs}}_{\text{p}}$. See the full version for a discussion of uniformity of p-values with unequal group sizes.

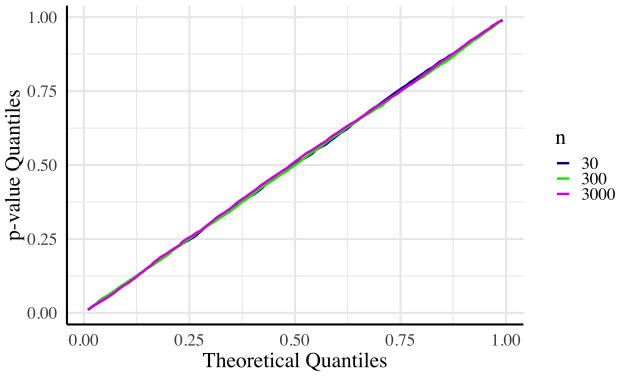

Figure 2: A quantile-quantile plot of $\widetilde{\text{KWabs}}_{\mathbf{p}}$ comparing the distribution of simulated p-values to the uniform distribution at varying n, all with equal group sizes. ($g = 3$; $\epsilon = 1$)

Comparison to previous work. The only prior work on hypothesis testing for independence of two variables, one continuous and one categorical, is that on ANOVA. The best private ANOVA analogue is that of Swanberg et al. [28]. In Figure 3 we compare $\widetilde{\text{KWabs}}_{\text{p}}$ to their test and we find its power to be much greater. To get 80% power with this effect size, our test requires only 23% as much data as the private ANOVA test. (The effect size used is the same as in [28].) We stress that this means our test is significantly higher-power, *in addition* to being usable for non-normal data. The test of Swanberg et al. also requires that the analyst issuing the query can accurately bound the range of the data—a bound that is too tight or too loose will reduce the power of the test. Our test works for data with unknown range.

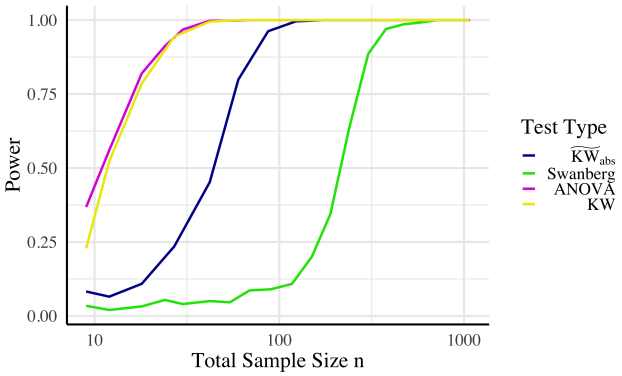

Figure 3: Power of $\widetilde{\text{KWabs}}_{\mathbf{p}}$, Swanberg et. al.'s test [28], and the public tests at various n. (Effect size: $max(\mu_n) - min(\mu_n) = 2\sigma$; $\epsilon = 1$; $g = 3$; $\alpha = .05$; equal group sizes; continuous sample data)

Robustness of results. Though it is unusual, it is possible that the relative power of different hypothesis tests could change when different effect sizes are considered. Therefore we repeat the analysis shown in Figure 1 with a variety of different effect sizes, group sizes, and number of groups. We also vary the frequency of tied values in the data, since the random ordering of tied values adds additional noise for our statistic. Finally, we run the comparison on real data comparing income and age. The results of these experiments are shown in the full version. We find that the results discussed above are consistent across these variations.

4 TWO GROUPS

We now consider the case of data with only two groups (e.g., restricting our comparison to the methylation levels of the bipolar subjects versus the healthy controls.) In the public nonparametric setting, one could simply use Kruskal-Wallis with $g = 2$, but one can also use the Mann-Whitney U-test (also called the Wilcoxon *rank-sum* test), proposed in 1945 by Frank Wilcoxon [35] and formalized in 1947 by Henry Mann and Donald Whitney [18]. In this section we construct a private version of the Mann-Whitney test and compare it to simply using $\widetilde{\text{KWabs}}_{\text{p}}$ with $g = 2$.

The standard parametric test in the public setting is the two-sample t-test. We know of three prior works that can, in some sense, be seen as providing an analogue of the two-sample t-test for the standard private setting. The only one for which this is an explicit goal is that of D'Orazio et al. [7]. This test releases private estimates of the difference in means and of the within-group variance and produces a confidence interval rather than a p-value. (The difference in means is done with simple Laplace noise, while the variance estimate uses a subsample-and-aggregate algorithm.) Most importantly, they assume that the size of the two groups is public knowledge, where we treat the categorical value of a data point (ex., schizophrenic or not) to be private data.

There are two other works we know of that provide a private analogue of the two-sample t-test as a result of a slightly different goal. The first is Ding et al. [6], who give a test under the more restrictive *local* differential privacy definition. This test is of course also private under the standard differential privacy definition. The other work is that of Swanberg et al. [28], who give a private analogue of the ANOVA test, as discussed previously. In the public setting, ANOVA with $g = 2$ is equivalent to the two-sample t-test. Based on (somewhat incomparable) experiments in their respective papers, it appears that the Swanberg et al. test is much higher power, which is unsurprising given that it was developed for the centralized database model of privacy. We therefore compare our work to this.

To our knowledge, there is no prior work specifically on a private version of the Mann-Whitney test. As before, we find that our rank-based nonparametric tests outperform the private parametric equivalent even when the data is normally distributed. We also find that, unlike in the public setting, the more generic Kruskal-Wallace analogue (used with $g = 2$) outperforms the more purpose-built test.

4.1 The Mann-Whitney test

The function used to calculate the Mann-Whitney U statistic is formalized in Algorithm MW_{stat}. As before, $\mathbf{x}$ is a database of size n, with r_{ij} being the rank of the j^{th} data point in group i. A statistic is first calculated for each group by summing the rankings in that group and subtracting a term depending on the group size. We

then take the minimum of the two statistics to get U. Compared to the other statistics we are considering, the directionality of U is reversed — low values are inconsistent with the null hypothesis and cause rejection, rather than high values.

Algorithm $\mathrm{MW}_{\mathrm{stat}}$ **:** Mann-Whitney Test Statistic

Input: $\mathbf{x}$

for $i \in \{1, 2\}$ **do**

  $R_i \longleftarrow \sum_j r_{ij}$

  $U_i \longleftarrow R_i - \frac{n_i(n_i+1)}{2}$

$U \longleftarrow \min\{U_1, U_2\}$

Output: U

4.2 A Differentially Private Algorithm

The global sensitivity of $\mathrm{MW}_{\mathrm{stat}}$ is n, but the local sensitivity is lower. We prove the following in the full version:

Theorem 4.1 (Sensitivity of $\mathrm{MW}_{\mathrm{stat}}$). *The local sensitivity is given by* $LS_{\mathrm{MW}_{\mathrm{stat}}}(\boldsymbol{x}) = \max\{n_1, n_2\}$*, where* n_1 *and* n_2 *are the sizes of the two groups in* $\mathbf{x}$*.*

We can now define our private test statistic, $\widetilde{\mathrm{MW}}_{\mathrm{stat}}$. This algorithm first uses a portion of its privacy budget (ϵ_m) to estimate the size of the smallest group. This value is then reduced to m^*, such that with probability $1 - \delta$ we have $n - m^* > LS_{\mathrm{MW}_{\mathrm{stat}}}(\mathbf{x})$. This means that we can then release U using noise proportional to $n - m^*$ (using the remaining privacy budget, ϵ_U. See the full version for proof that $\widetilde{\mathrm{MW}}_{\mathrm{stat}}$ is $(\epsilon_m + \epsilon_U, \delta)$-differentially private.

Algorithm $\widetilde{\mathrm{MW}}_{\mathrm{stat}}$ **:** Private Mann-Whitney Test Statistic

Input: $\mathbf{x}$, ϵ_m, ϵ_U, δ

$m \longleftarrow \min\{n_1, n_2\}$

$\widetilde{m} \longleftarrow m + \mathrm{Lap}\left(\frac{1}{\epsilon_m}\right)$

$c \longleftarrow -\frac{\ln(2\delta)}{\epsilon_m}$

$m^* \longleftarrow \max(\lceil \widetilde{m} - c \rceil, 0)$

$\widetilde{U} \longleftarrow \mathrm{MW}_{\mathrm{stat}}(\mathbf{x}) + \mathrm{Lap}\left(\frac{n-m^*}{\epsilon_U}\right)$

Output: $\widetilde{m}$, $\widetilde{U}$

As before, $\widetilde{\mathrm{MW}}_{\mathrm{stat}}$ is not meaningful on its own; we want an applicable reference distribution with which to calculate a corresponding p-value. This is shown below in algorithm $\widetilde{\mathrm{MW}}_{\mathrm{p}}$. It works similarly to the analogous algorithms $\widetilde{\mathrm{KW}}_{\mathrm{p}}$ and $\widetilde{\mathrm{KWabs}}_{\mathrm{p}}$. The key difference is that the reference distribution now depends on the group size estimate $\widetilde{m}$.[8]

[8]The algorithm given simulates full databases to compute the reference distribution. This is not particularly slow, but in the full version we show that one can also sample from a normal distribution with certain parameters to get an acceptable reference distribution more quickly.

Algorithm $\widetilde{\mathrm{MW}}_{\mathrm{p}}$ **:** Complete Mann-Whitney Test

Input: $\mathbf{x}$, ϵ_m,ϵ_U,δ, z

$(\widetilde{m}, \widetilde{U}) \longleftarrow \widetilde{\mathrm{MW}}_{\mathrm{stat}}(\mathbf{x}, \epsilon_{\mathrm{m}}, \epsilon_{\mathrm{U}}, \delta)$

$\widetilde{m} \longleftarrow \lceil \max(0, \widetilde{m}) \rceil$

for $k := 1$ to z **do**

  $\mathbf{x}^* \longleftarrow$ a database with n independent uniform values from [0,1] divided into 2 groups of size $\widetilde{m}$ and $n - \widetilde{m}$

  $U_k \longleftarrow \widetilde{\mathrm{MW}}_{\mathrm{stat}}(\mathbf{x}^*, \epsilon_{\mathrm{m}}, \epsilon_{\mathrm{U}}, \delta)$

$p \longleftarrow$ fraction of U_k less than $\widetilde{U}$

Output: $\widetilde{U}, p$

A note on design. In the case of $\widetilde{\mathrm{KWabs}}_{\mathrm{p}}$ we found that the highest possible critical value came from a reference distribution with equal-size groups. For this test that is not the case, so we cannot use equal-size groups when generating the reference distribution without unacceptable type 1 error. As a result, we need an estimate of group size. If we didn't need this estimate for the reference distribution, it is possible that $\widetilde{\mathrm{MW}}_{\mathrm{stat}}$ would be more accurate by simply using the global sensitivity bound on $\mathrm{MW}_{\mathrm{stat}}$. (It would be a slightly higher sensitivity, but no privacy budget would need to be expended on estimating m.) This is a good example of a point made in Section 1: simply acheiving an accurate of estimate of a test statistic is not enough. The ultimate goal of a hypothesis test is a p-value, which also requires an accurate reference distribution and high power in order to minimize decision error.

Type 1 error. The reference distribution in the $\widetilde{\mathrm{MW}}_{\mathrm{p}}$ algorithm depends on m, which is only estimated by $\widetilde{m}$, so we need to experimentally verify that type 1 error never exceeds α. See the full version for evidence that our estimate appears to be sufficiently accurate and for additional discussion.

Theorem 4.2. *Algorithm* $\widetilde{\mathrm{MW}}_{\mathrm{p}}$ *is* $(\epsilon_m + \epsilon_U, \delta)$*-differentially private.*

Proof. Since the computation of $(\widetilde{m}, \widetilde{U})$ is $(\epsilon_m + \epsilon_U, \delta)$-differentially private (see full version for the proof of this fact) and all of the steps following this computation do not require access to the database and are, thus, post processing, by Theorem 2.4, it follows that the complete algorithm is also $(\epsilon_m + \epsilon_U, \delta)$-differentially private. □

4.3 Experimental Results

Power analysis. We first assessed the power of our test on synthetic data.[9] We run $\widetilde{\mathrm{MW}}_{\mathrm{p}}$ on many simulated databases and report the percentage of the time that a significant result was obtained. For our first effect size, we have the two groups consist of normally distributed data with means one standard deviation apart. In all experiments we set $\delta = 10^{-6}$.

Our first step was to determine the optimal proportion of the total privacy budget, ϵ_{tot}, to allot to estimating m and the test statistic $\widetilde{\mathrm{MW}}_{\mathrm{stat}}$. We found that the optimal proportion of ϵ to allot to estimating m is roughly .65, experimentally confirmed at several choices of ϵ_{tot}, effect size, total sample size n, group size ratios n_1/n, and underlying distribution. (See the full version for more

[9]For application of our test to real-world data, see the full version

details.) We then fix the proportion of ϵ_{tot} allotted to ϵ_m as .65 and vary ϵ_{tot} and total sample size n to measure the power of our test.

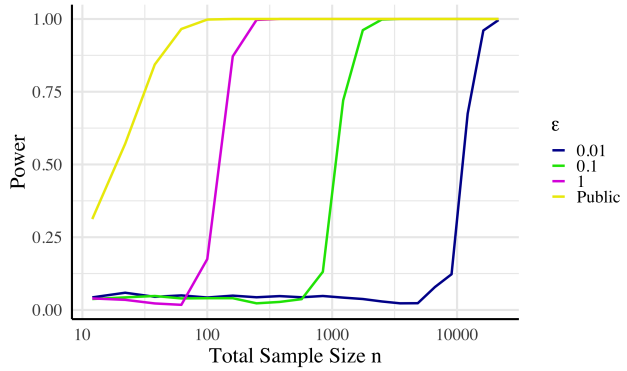

Figure 4: Power of $\widetilde{\text{MW}}_\text{p}$ at various values of ϵ_{tot} and total sample size n. (Effect size: $\mu_1 - \mu_2 = 1\sigma$; proportion of ϵ_{tot} to $\epsilon_m = .65$; $\alpha = .05$; m:$(n-m) = 1$)

As shown in Figure 4, the power loss due to privacy is not unreasonably large. At an ϵ_{tot} of 1, the test only requires a database that is approximately a factor of 3 larger than that needed for the public test to reach a power of 1. As one might expect, the database size needed to detect a given effect has a roughly inverse relationship with ϵ_{tot}. In the full version we perform a similar power analysis, varying effect size rather than sample size.

Uniformity of p-values. Algorithm $\widetilde{\text{MW}}_\text{p}$ uses the privatized group sample sizes $m^*, (n-m^*)$ in place of the true group sizes n_1, n_2 in order to simulate the reference distribution. Naturally, then, one may wonder *how* conservative our critical values are as a result of ensuring that the type 1 error rate does not exceed α. As shown in Figure 5, the type I error rate of our test does not exceed α when group sample sizes are equal. As total sample size n increases, the p-value quantiles asymptotically approach that of the theoretical distribution. In the full version, we also examine uniformity of p-values of $\widetilde{\text{MW}}_\text{p}$ with unequal group sizes and a variation of $\widetilde{\text{MW}}_\text{p}$ that assumes equal group sizes.

Comparison to previous work. The best existing test applicable in the same use case is that of Swanberg et al. [28]. Their differentially private ANOVA test can be used in the 2-group case to compare to our Mann-Whitney test. The results of this comparison, using the same paramater settings chosen for optimal power in their test, can be seen in Figure 6, where our test offers a substantial power increase.

Comparing $\widetilde{\text{MW}}_\text{p}$ and $\widetilde{\text{KWabs}}_\text{p}$. Both the Mann-Whitney and the Kruskal-Wallis can be used to compare the distributions of samples from two groups. As shown in Figure 7, we find that in the private setting, $\widetilde{\text{KWabs}}_\text{p}$ is more statistically powerful than $\widetilde{\text{MW}}_\text{p}$. This is perhaps surprising, since one might expect the test developed specifically for the two-group case to perform better. But this is an example of how some tests privatize more easily than others. $\widetilde{\text{MW}}_\text{p}$ requires knowledge of the group sizes, using up a fraction of the privacy budget, while the $\widetilde{\text{KWabs}}_\text{p}$ statistic is not dependent on group size.

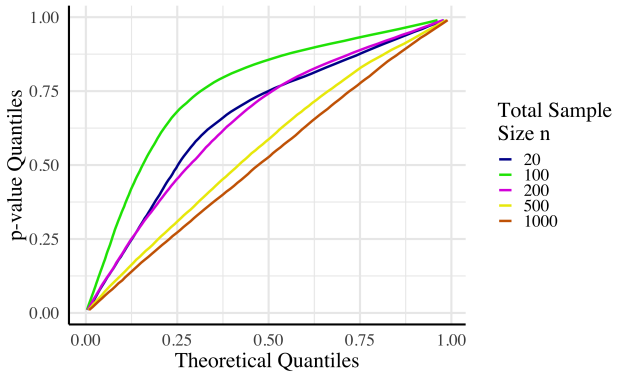

Figure 5: A quantile-quantile plot of $\widetilde{\text{MW}}_\text{p}$ varying n. ($\epsilon_{tot} = 1$; proportion of ϵ_{tot} to $\epsilon_m = .65$; m:$(n-m) = 1$; normally distributed sample data)

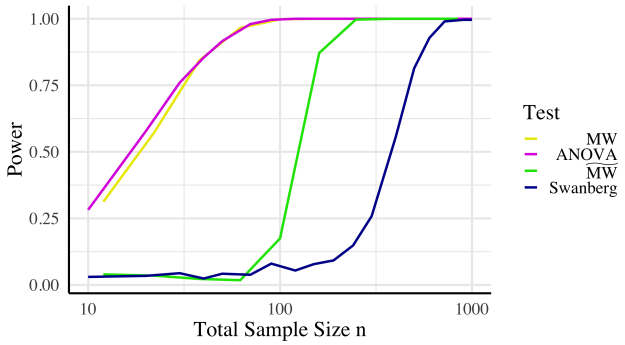

Figure 6: Power of $\widetilde{\text{MW}}_\text{p}$ and Swanberg et. al.'s test at various n. ($\epsilon_{tot} = 1$; Effect size: $\mu_1 - \mu_2 = 1\sigma$; proportion of ϵ_{tot} to $\epsilon_m = .65$; $\alpha = .05$; m:$(n-m) = 1$), normally distributed sample data

We did find one exception to this finding. If the analyst knows *a priori* that the two groups are of equal size (e.g., the data collection method guaranteed an equal number in each group) then $\widetilde{\text{MW}}_\text{p}$ can be run using an exact value of $n/2$ for the local sensitivity without the need to dedicate any privacy budget to estimating m. This increases the accuracy of $\widetilde{\text{MW}}_\text{stat}$ both by reducing the sensitivity used to add noise and by increasing the privacy budget allocated to U. We find that in this case $\widetilde{\text{MW}}_\text{p}$ *is* superior to $\widetilde{\text{KWabs}}_\text{p}$. See the full version for more details.

5 PAIRED DATA

We now consider a third situation, where there are two groups and the observations in those groups are paired. While this scenario did not exist in the original psychotic disease study, one can imagine recording the methylation levels of one of the groups before and after administering medication. Each subject then contributes a pair of data (u_i, v_i) that are highly correlated with one another. One

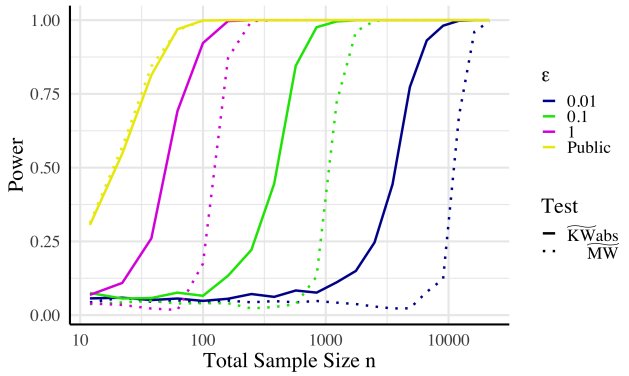

Figure 7: Power of $\widetilde{\text{MW}}_{\mathbf{p}}$ and $\widetilde{\text{KWabs}}_{\mathbf{p}}$ at various n and ϵ. (Effect size: $max(\mu_n) - min(\mu_n) = 1\sigma$; $g = 2$; $\alpha = .05$; equal group sizes; normally distributed sample data)

can assess the impact of the medication by considering whether the set of n differences, $\{v_i - u_i\}_i$, is plausibly centered at zero. The standard nonparametric hypothesis test for this situation is the Wilcoxon signed-rank test, proposed in 1945 by Frank Wilcoxon [36]. The parametric alternative is a simple one-sample t-test run on the set of differences.

This is the one setting where we are aware of prior work on a nonparametric test. Task and Clifton [30] gave the first private analogue of the Wilcoxon signed-rank test, referred to from here on as the *TC test*, in 2016. Our test makes two key improvements to theirs and exhibits higher power. We also correct some errors in their work. We discuss the differences in more detail in Section 5.3.

Despite its status as one of the most commonly used hypothesis tests, to our knowledge there is no practical, implementable private version of a one-sample t-test in the literature. In Section 5.4 we discuss some work that comes close, and then we give our own first attempt at a private t-test. We again find that our nonparametric test has significantly higher power than this parametric alternative.

5.1 The Wilcoxon signed-rank test

The function calculating the Wilcoxon test statistic is formalized in Algorithm $\mathcal{W}$. Given a database $\mathbf{x}$ containing sets of pairs (u_i, v_i), the test computes the difference d_i of each pair, drops any with $d_i = 0$, and then ranks them by magnitude. (If magnitudes are equal for several differences, all are given a rank equal to the average rank for that set.)

Under the null hypothesis that u_i and v_i are drawn from the same distribution, the distribution of the test statistic W can be calculated exactly using combinatorial techniques. This becomes computationally infeasible for large databases, but an approximation exists in the form of the normal distribution with mean 0 and variance $\frac{n_r(n_r+1)(2n_r+1)}{6}$, where n_r is the number of rows that were not dropped. Knowing this, one can calculate the p-value for any particular value of w.

Algorithm W_{stat} **:** Wilcoxon Test Statistic

Input: $\mathbf{x}$
for *row* i *of* $\mathbf{x}$ **do**
  $d_i \longleftarrow |v_i - u_i|$
  $s_i \longleftarrow \text{Sign}(v_i - u_i)$
Order the terms from lowest to highest d_i
Drop any $d_i = 0$
for *row* i *of* $\mathbf{x}$ **do**
  $r_i \longleftarrow$ rank of row i
$w \longleftarrow \sum_i s_i r_i$
Output: w

5.2 Our Differentially Private Algorithm

At a high level, our algorithm is quite straightforward and similar to prior work: we compute a test statistic as one might in the public case and add Laplacian noise to make it private. However, there are several important innovations relative to Task and Clifton that greatly increase the power of our test.

Our first innovation is to use a different variant of the Wilcoxon test statistic. While the version introduced in Section 5.1 is the one most commonly used, other versions have long existed in the statistics literature. In particular, we look at a variant introduced by Pratt in 1959 [22]. In this variant, rather than dropping rows with $d_i = 0$, those rows are included. When $d_i = 0$ we set $s_i = \text{Sign}(d_i) = 0$, so those rows contribute nothing to the resultant statistic, but they do push up the rank of other rows.

Algorithm WP_{stat}**:** Wilcoxon Test Statistic - Pratt Variant

Input: $\mathbf{x}$
for *row* i *of* $\mathbf{x}$ **do**
  $d_i \longleftarrow |v_i - u_i|$
  $s_i \longleftarrow \text{Sign}(v_i - u_i)$
Order the terms from lowest to highest d_i
for *row* i *of* $\mathbf{x}$ **do**
  $r_i \longleftarrow$ rank of row i
$w \longleftarrow \sum_i s_i r_i$
Output: w

In the public setting, the Pratt variant is not very different from the standard Wilcoxon, being slightly more or less powerful depending on the exact effect one is trying to detect [5]. In the private setting, however, the difference is substantial.

The benefit to the Pratt variant comes from how the test statistics are interpreted. In the standard Wilcoxon, it is known that the test statistic follows an approximately normal distribution, but the variance of that distribution is a function of n_r, the number of non-zero d_i values. In the private setting, this number is not known, and this has caused substantial difficulty in prior work. (See Section 5.3 for more discussion.) On the other hand, the Pratt variant produces a test statistic that is always compared to the same normal distribution, which depends only on n. The algorithm $\widetilde{\text{WP}}_{\text{stat}}$ that outputs a differentially private analogue is shown below.

Theorem 5.1. *Algorithm* $\widetilde{\text{WP}}_{\text{stat}}$ *is ϵ-differentially private.*

Algorithm $\widetilde{\text{WP}}_{\text{stat}}$ **:** Private Wilcoxon Test Statistic

Input: $\mathbf{x}$, ϵ

$w \longleftarrow \text{WP}_{\text{stat}}(\mathbf{x})$

$\widetilde{w} \longleftarrow w + \text{Lap}\left(\frac{2n}{\epsilon}\right)$

Output: $\widetilde{w}$

See the full version for proof of Theorem 5.1.

To complete the design of our test, we compute a reference distribution through simulation as was done in $\widetilde{\text{KWabs}}_{\text{p}}$ and $\widetilde{\text{MW}}_{\text{p}}$. Here we use the standard normal approximation for the distribution of the w test statistic, though one could simulate full databases as well. We call this algorithm $\widetilde{\text{WP}}_{\text{p}}$.

Algorithm $\widetilde{\text{WP}}_{\text{p}}$ **:** Complete Wilcoxon Test

Input: $\mathbf{x}$, ϵ, z

$\widetilde{w} \longleftarrow \widetilde{\text{WP}}_{\text{stat}}(\mathbf{x}, \epsilon)$

for $k = 1$ to z **do**

  $w_k \longleftarrow \text{Normal}(0, n(n+1)(2n+1)/6) + \text{Lap}(2n/\epsilon)$;

$p \longleftarrow$ fraction of w_k more extreme than $\widetilde{w}$

Output: $\widetilde{w}, p$

Theorem 5.2. *Algorithm* $\widetilde{\text{WP}}_{\text{p}}$ *is* ϵ*-differentially private.*

Proof. The computation of $\widetilde{w}$ was already shown to be private. The remaining computation needed to find the p-value does not need access to the database—it is simply post-processing. By Theorem 2.4, it follows that the $\widetilde{\text{WP}}_{\text{p}}$ algorithm is also private. □

5.3 Experimental Results

Power analysis. We assess the power of our differentially-private Wilcoxon signed-rank test first on synthetic data. (For tests with real data, see the full version.) In order to measure power, we must first fix an effect size. We chose to have the u_i and v_i values both generated according to normal distributions with means one standard deviation apart. We then measure the statistical power of Algorithm $\widetilde{\text{WP}}_{\text{p}}$ (for a given choice of n and ϵ) by repeatedly randomly sampling a database $\mathbf{x}$ from that distribution and then running $\widetilde{\text{WP}}_{\text{p}}$ on that database.[10] The power is the percentage of the time $\widetilde{\text{WP}}_{\text{p}}$ returns a p-value less than α. See the full version for a similar analysis of power, varying effect size rather than sample size.

Uniformity of p-values. In algorithm $\widetilde{\text{WP}}_{\text{p}}$ we draw our reference distribution samples (the w_k values) assuming there are no $d_i = 0$ rows. The distribution will technically differ slightly when there are many rows with $d_i = 0$, so we need to confirm experimentally that the difference is inconsequential or otherwise acceptable.

Figure 9 shows a Q-Q plot of $\widetilde{\text{WP}}_{\text{p}}$ on three sets of p-values, all generated under H_0, with $\epsilon = 1$, $n = 500$. When there are no ties in the original data (0% of $d_i = 0$), the Q-Q plot line is indistinguishable from the identity line, indicating that the test

[10] Our actual implementation differs slightly from this. To save time when running a huge number of tests with identical n and ϵ, we first generate the reference distribution W_k values, which can be reused across runs.

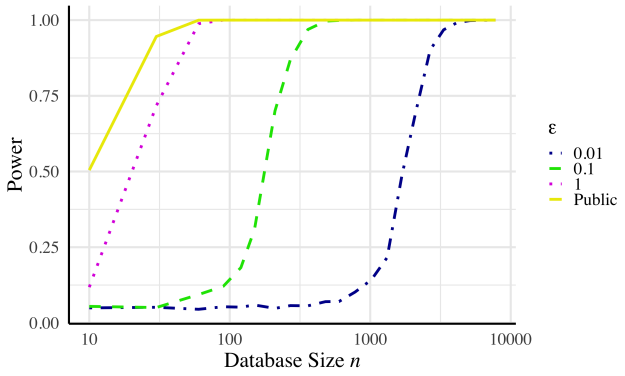

Figure 8: Power of $\widetilde{\text{WP}}_{\text{p}}$ **at various** ϵ **and** n**. (Effect size:** $\mu_u - \mu_v = 1\sigma$**;** $\alpha = .05$**; normally distributed sample data)**

is properly calibrated. Encouragingly, introducing a substantial number of ties into the data (30% of $d_i = 0$) has little noticeable effect. In order to induce non-uniformity in the p-values, one needs an extremely high proportion of rows with $d_i = 0$. The curve with 90% zero values is shown as an illustration. When the proportion of zeros is very high, the variance of the p-values will be narrower than the reference distribution, resulting in a lower critical value. Since the value we are using is higher, our test is overly conservative,[11] but this is acceptable as type I error is still below α.

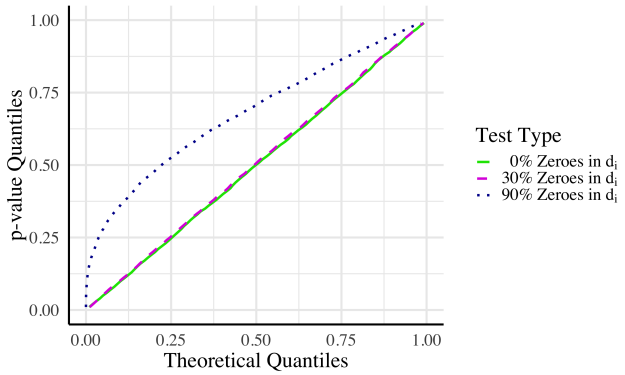

Figure 9: A quantile-quantile plot of $\widetilde{\text{WP}}_{\text{p}}$ **comparing the distribution of simulated p-values to the uniform distribution (**$\epsilon = 1$**,** $n = 500$**; normally distributed sample data).**

Comparison to previous work. In 2016, Task and Clifton [30] introduced the first differentially private version of the Wilcoxon signed-rank test, from here on referred to as the TC test. Our work improves upon their test in two ways. We describe the two key differences below, and then compare the power of our test to theirs. We also found a significant error in their work.[12] All comparisons

[11] One could try to estimate the number of zeros to be less conservative, but that would require allocating some of the privacy budget towards that estimate, which is not worth it in most circumstances.

[12] This error has been confirmed by Task and Clifton in personal correspondence.

are made to our implementation of the TC test with the relevant error corrected.

Task and Clifton compute an analytic upper bound on the critical value t^*. For a given n and ϵ, the private test statistic $\widetilde{w}$ under H_0 is sampled according to a sum $W+\Lambda$, where W is a random draw from a normal distribution (scaled according to n) and Λ is a Laplace random variable (scaled according to n and ϵ). In particular, say that b is a value such that $\Pr[W > b] < \beta$ and g is a value such that $\Pr[\Lambda > g] < \gamma$. Then we can compute the following bound. (The last line follows from the fact that the two events are independent.)

$$\begin{aligned}\Pr[\widetilde{W} > b+g] &< \Pr[W > b \text{ or } \Lambda > g] \\ &= \Pr[W > b] + \Pr[\Lambda > g] \\ &\quad - \Pr[W > b \text{ and } \Lambda > g] \\ &= \beta + \gamma - \beta\gamma\end{aligned}$$

Task and Clifton always set $\gamma = .01$ and then vary the choice of β such that they have $\alpha = \beta + \gamma - \beta\gamma$ for whatever α is intended as the significance threshold.[13]

The bound described above is correct but very loose, and our simulation method gives drastically lower critical values. Table 1 contains examples of the critical values achieved by each method for several parameter choices. More values can be found in the full version, where we also experimentally confirm that these values result in acceptable type 1 error.

Table 1: Critical Value Comparison for $n = 100$

ϵ	α	Public	New	TC
1	0.1	1.282	1.417	2.680
	0.05	1.645	1.826	3.091
	0.025	1.960	2.186	3.511
0.1	0.1	1.282	5.684	14.786
	0.05	1.645	8.063	15.197
	0.025	1.960	10.438	15.617
0.01	0.1	1.282	55.350	135.843
	0.05	1.645	79.233	136.254
	0.025	1.960	103.116	136.674

Critical values for $n = 100$ and several values of ϵ and α. To allow easy comparison, these values are for a normalized W statistic, i.e., W has been divided by the relevant constant so that it is (before the addition of Laplacian noise) distributed according to a standard normal. See the full version for the equivalent table at n = 1000.

Our second key change from the TC test, mentioned earlier, is that we handle rows with $d_i = 0$ according to the Pratt variant of the Wilcoxon, rather than dropping them completely as is more traditional. The reason the traditional method is so difficult in the private setting is that the reference distribution one must compare to depends on the number of rows that were dropped. If n_r is the number of non-zero rows (i.e., rows that weren't dropped), one is supposed to look up the critical value associated with n_r, rather than the original size n of the database.

Unfortunately, n_r is a sensitive value and cannot be released privately.[14] Task and Clifton show that it is *acceptable* (in that it does not result in type I error greater than α) to compare to a critical value for a value of n_r that is lower than the actual value. This allows them to give two options for how one might deal with the lack of knowledge about n_r.

High Utility This version of the TC test simply assumes $n_r \geq .3n$ and uses the critical value that would be correct for $n_r = .3n$. We stress that this algorithm is *not* actually differentially private, though it could easily be captured by a sufficiently weakened definition that limited the universe of allowable databases. Another problem is that for most realistic data, n_r is much greater than $.3n$ and using this loose lower bound still results in a large loss of power.

High Privacy This version adds k dummy values to the database with $d_i = \infty$ and k with $d_i = -\infty$.[15] Then one can be certain of the bound $n_r \geq 2k$. This is a guaranteed bound so this variant truly satisfies differential privacy. On the other hand it is a very loose lower bound in most cases, leading to a large loss of power.

Experimental comparison. We compare the statistical power of our test to that of the TC test. We begin by again measuring the power when detecting the difference between two normal distributions with means one standard deviation apart. The results can be seen in Figure 10. If we look at the database size needed to achieve 80% power, we find that the 32 data points we need, while more than the public test (14), are many fewer than the TC High Utility variant (80) or the TC High Privacy variant (122). The full version includes a figure for $\epsilon = .1$ as well. What we see is that, while all private tests require more data, our test (requiring $n \approx 236$) still requires about 40% as much data as the TC High Utility variant (588). The TC High Privacy variant, however, scales much less well to low ϵ and requires roughly 2974 data points.

The results in Figure 10 use a continuous distribution for the real data, so there are no data points with $d_i = 0$. Because one of the crucial differences between our algorithms is the method for handling these zero values, we also consider the effect when there are a large number of zeros in the full version. Overall, we see that both in situations with no zero values and situations with many, our test achieves the rigorous privacy guarantees of the TC High Privacy test while achieving greater utility than the TC High Utility test.

Relative contribution of improvements. Given that we make two meaningful changes to the TC test, one might naturally wonder whether both are truly useful or whether the vast majority of the improvement comes from one of the two changes. To test this, we compare to an updated variant of the TC test where we calculate critical values exactly through simulation, as we do in our algorithm, but otherwise run the TC test unchanged (referred to as "High

[13] This is where Task and Clifton make an error. This formula is correct, but they used an incorrect density function for the Laplace distribution and as a result calculated incorrect values of g.

[14] A private estimate could be released, but one would have to devote a significant portion of the privacy budget for the hypothesis test to this estimate, greatly decreasing the accuracy/power of $\widetilde{W}_{\text{stat}}$.

[15] Task and Clifton do not discuss how to choose k, and in our experimental comparisons we set $k = 15$, the same value they use.

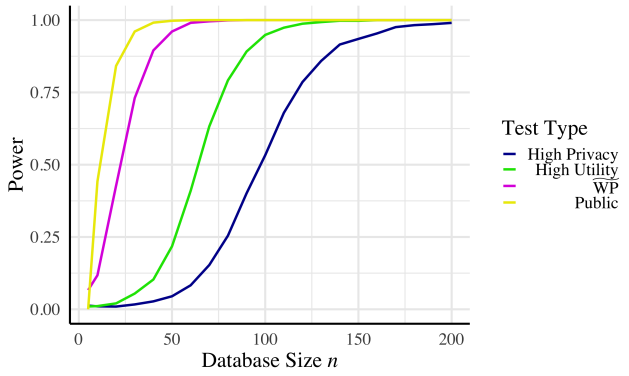

Figure 10: Power of the TC test, $\widetilde{\text{WP}}_{\mathbf{p}}$, and the public test at various n. (Effect size: $\mu_u - \mu_v = 1\sigma$, $\epsilon = 1$; $\alpha = .05$; normally distributed sample data)

Privacy +" and "High Utility +"). The result is presented in Figure 11, where we find the resulting algorithm to rest comfortably between the original TC test and our proposed test. This means that both the change to the critical value calculation and the switch to the Pratt method of handling $d_i = 0$ rows are important contributions to achieving the power of our test.

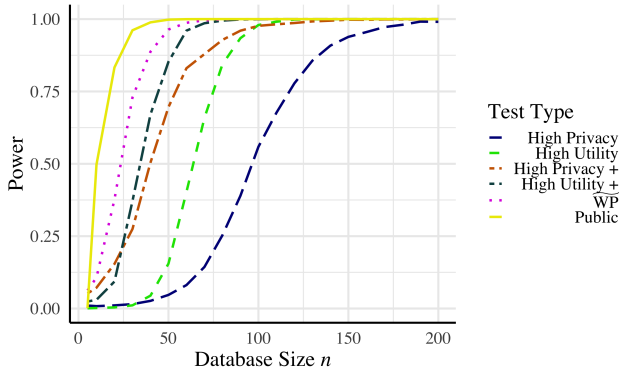

Figure 11: Power comparison of the TC algorithms, the TC algorithms with our critical values (denoted with a +), our new algorithm, and the public algorithm at various sample sizes n. (Effect size: $\mu_u - \mu_v = 1\sigma$; $\epsilon = 1$; $\alpha = .05$)

5.4 Parametric Alternative: A New T-test

The parametric analog to the Wilcoxon test is to run a one sample t-test on the set of differences $\{v_i - u_i\}_i$ to see if their mean is significantly different from zero (also called a paired t-test). There has been surprisingly little work on the creation of a private version of a one sample t-test. Karwa and Vadhan [16] study private confidence intervals, which are in a sense equivalent to a t-test. However, their analysis is asymptotic and they say that the algorithm does not give practical results with database size in the thousands. Sheffet [24] provides a method for calculating private coefficient estimates for linear regression and also transforms the t-distribution to provide an appropriate reference distribution for inference. In the public setting, one can convert a test on regression coefficients to a one sample t-test but choosing a constant independent variable and making the sample data the dependent variable. However, Sheffet's method only works when all variables are significantly spread out, so this method fails.

Here we propose what we believe is the first private version of a one sample t-test, with two arguable exceptions. The first is simultaneous work by Gaboardi et al. [13] in the local privacy model. We compare our results to theirs in more detail in Section 5.5. The other work is that of Solea [27], but according to Solea's own experiments that test often gives type 1 error rates well above the chosen α for many parameter choices, so we don't consider it a usable test.

The database for a one sample t-test has observations $x_1, \ldots, x_n$ assumed to come from a normal distribution with mean μ and standard deviation σ. (For paired data, each observation is the differences between the observation in the two groups). The test statistic is given by $\mathsf{T}_{\text{stat}}(\mathbf{x}) = \frac{\bar{x}}{s/\sqrt{n}}$, where $\bar{x}$ is the mean of the data and s is the standard deviation of the data.

A private t-test. As before, we achieve privacy through the addition of Laplacian noise, but the sensitivity of T_{stat} is unbounded, so we instead release separate private estimates of the numerator and denominator. For this analysis, similar to the private ANOVA tests [28], we assume that the data is scaled such that all observations are on the interval $[-1, 1]$. We first find the sensitivities of $\bar{x}$ and s^2 and then use post-processing, composition, and the Laplace Mechanism to combine these to obtain the private t-statistic. In the case where s^2 is estimated to be negative, the test statistic cannot be computed as normal, and we return 0, indicating an unwillingness to reject the null hypothesis.

Theorem 5.3. *The sensitivity of $\bar{x}$ is $\frac{2}{n}$.*

Theorem 5.4. *The sensitivity of s^2 is $\frac{5}{n-1}$.*

See the full version for proof of Theorem 5.3 and 5.4.

Algorithm $\widetilde{\mathsf{T}}_{\text{stat}}$: Private t-Test Statistic

Input: $\mathbf{x}$, $\epsilon_{\bar{x}}$, ϵ_{s^2}

$\widetilde{\bar{x}} = \bar{x} + \text{Lap}(\frac{1/n}{\epsilon_{\bar{x}}})$

$\widetilde{s^2} = s^2 + \text{Lap}(\frac{5/(n-1)}{\epsilon_{s^2}})$

if $\widetilde{s^2} < 0$ **then**

  $\widetilde{T} = 0$

else

  $\widetilde{T} = \frac{\widetilde{\bar{x}}/n}{\sqrt{\widetilde{s^2}}/\sqrt{n}}$

Output: $\widetilde{T}$

Theorem 5.5. *Algorithm $\widetilde{\mathsf{T}}_{\text{stat}}$ is $(\epsilon_{\bar{x}} + \epsilon_{s^2})$-differentially private.*

Proof. By the Laplace mechanism, the computation of $\widetilde{\bar{x}}$ is $\epsilon_{\bar{x}}$-differentially private and the computation of $\widetilde{s^2}$ is ϵ_{s^2}-differentially private. Since the computation of $\widetilde{T}$ does not require access to the database, it is only post-processing and its release is $(\epsilon_{\bar{x}} + \epsilon_{s^2})$-differentially private. □

To carry out the full paired t-test, we estimate the reference distribution through simulation and release a private p-value.

Algorithm $\widetilde{\mathrm{T}}_{\mathrm{p}}$ **:** Complete t-Test

Input: $\mathbf{x}$, $\epsilon_{\bar{x}}$, ϵ_{s^2}, z

$\widetilde{t} := \widetilde{\mathrm{T}}_{\mathrm{stat}}(\mathbf{x}, \epsilon_{\bar{x}}, \epsilon_{s^2})$

for $k = 1$ to z **do**

  $\mathbf{x}^* \longleftarrow$ a database with n independent draws from $N(\mu = 0, \sigma \approx 0.3)$, each truncated to $[-1, 1]$

  $t_k \longleftarrow \widetilde{\mathrm{T}}_{\mathrm{stat}}(\mathbf{x}^*)$

$p \longleftarrow$ fraction of t_k more extreme than $\widetilde{t}$

Output: $\widetilde{t}, p$

Theorem 5.6. *Algorithm* $\widetilde{\mathrm{T}}_{\mathrm{p}}$ *is* $\epsilon_{\bar{x}} + \epsilon_{s^2}$*-differentially private.*

Proof. The computation of $\widetilde{t}$ was already shown to be private. The remaining computation needed to find the p-value does not need access to the database—it is simply post-processing. By Theorem 2.4, it follows that the $\widetilde{\mathrm{T}}_{\mathrm{p}}$ algorithm is also private. □

5.5 Experimental t-Test evaluation

We first must set a parameter in our $\widetilde{\mathrm{T}}_{\mathrm{p}}$ algorithm. In particular, for a given total ϵ, we must decide how to allocate the budget between $\epsilon_{\bar{x}}$ and ϵ_{s^2}. We choose this allocation experimentally, deciding to allocate 50% of the budget towards each value. This is nontrivial, and the full version contains experimental results and further discussion. Luckily, the exact choice of this allocation does not seem to have a large effect on the power of the test.

We then evaluate the power and validity of the final $\widetilde{\mathrm{T}}_{\mathrm{p}}$ test.

Comparison to other work. Simultaneous to our work, Gaboardi et al. [13] developed a private one sample t-test under the more restrictive local differential privacy model. As one might expect, our test in the more standard setting is much higher power. They develop both a t-test and a z-test, which is equivalent to the t-test except that the variance of the data is assumed to be already known. Only the z-test is given experimental evaluation, but with an effect size three times the size we use in our experiments, their test (at $\epsilon = 1$) requires roughly 4000 data points to reach 80% power, while our test requires roughly 100. Their t-test would presumably require even more data.

Comparison to nonparametric test. Since we have already developed a test for the paired-data use case, we assessed the power of $\widetilde{\mathrm{T}}_{\mathrm{p}}$ in comparison to $\widetilde{\mathrm{WP}}_{\mathrm{p}}$ by simulating synthetic data as described in Section 5.3. Just as in the many groups and two groups scenarios, the nonparametric test substantially outperforms its parametric counterpart, as shown in Figure 12. In this case, $\widetilde{\mathrm{WP}}_{\mathrm{p}}$ needs 8% of the data required by $\widetilde{\mathrm{T}}_{\mathrm{p}}$ to reach the same power.

Uniformity of p-values. As with all of our tests, we experimentally ensure that type I error rate is bounded by α in Figure 13. This figure confirms the fact that our type I error rate is bounded above by α. For small sample sizes, the line on the quantile-quantile plot goes above the diagonal. This is the acceptable direction, the sign of a conservative test. In this case it occurs because some test statistics

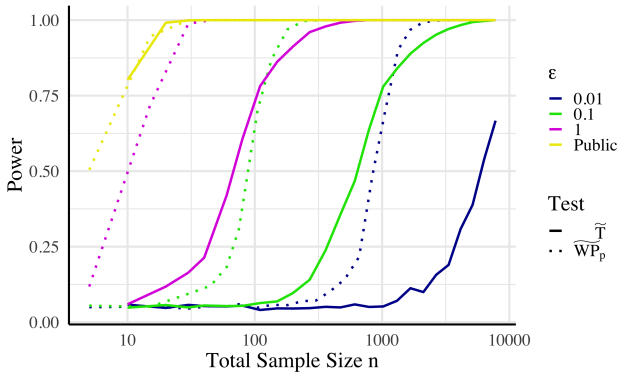

Figure 12: Power of $\widetilde{\mathrm{T}}_{\mathrm{p}}$ and $\widetilde{\mathrm{WP}}_{\mathrm{p}}$ at various ϵ and n. (Effect size: $\mu_u - \mu_v = 1\sigma$; $\alpha = .05$; normally distributed sample data)

in the reference distribution are set to zero (as a result of noise added for privacy overwhelming $\widetilde{s^2}$). If, for example, 10% of the reference distribution samples are at zero, then p values below 10% are impossible. As shown by the $n = 1000$ line, at sufficiently large sample sizes this effect essentially vanishes.

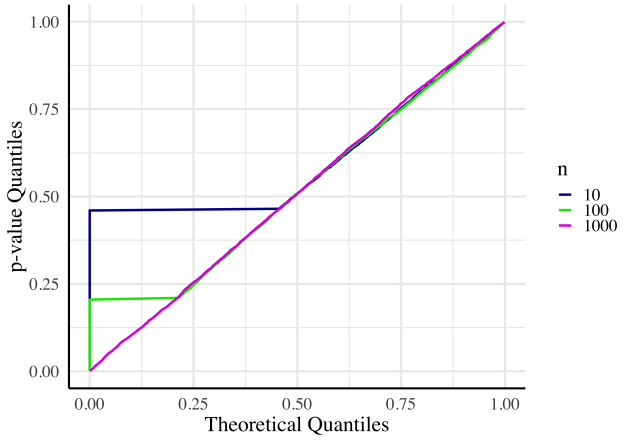

Figure 13: A quantile-quantile plot of $\widetilde{\mathrm{T}}_{\mathrm{p}}$ at various n. ($\epsilon = 1$; equal ϵ allotment to each statistic)

6 CONCLUSION

We have introduced several new tests, of which three ($\widetilde{\mathrm{KWabs}}_{\mathrm{p}}$, $\widetilde{\mathrm{MW}}_{\mathrm{p}}$, and $\widetilde{\mathrm{WP}}_{\mathrm{p}}$) are improvements on the state of the art. These allow researchers to address inferential questions using nonparametric methods while preserving the privacy of the data. More broadly, we found that the basic idea of using ranks in the private setting is potent. Not only do they remove the need to assume a bound on the data, they also directly increase statistical power. When working with many groups, two group, or with paired data, rank-based tests are more powerful than their parametric analogues and can be made yet more powerful through sensible adaptations.

We hope others will push this technique forward — we have no reason to believe that our tests are optimal.

ACKNOWLEDGMENTS

We would like to thank Christine Task and Chris Clifton for generous and enlightening discussions regarding their previous work. This material is based upon work supported by the National Science Foundation under Grant No. SaTC-1817245 and the Richter Funds.

REFERENCES

[1] Jordan Awan and Aleksandra Slavković. 2018. Differentially private uniformly most powerful tests for binomial data. In *Advances in Neural Information Processing Systems*. 4208–4218.

[2] Andrés F Barrientos, Jerome P Reiter, Ashwin Machanavajjhala, and Yan Chen. 2019. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics* (2019), 1–24.

[3] Zachary Campbell, Andrew Bray, Anna Ritz, and Adam Groce. 2018. Differentially Private ANOVA Testing. In *Data Intelligence and Security (ICDIS), 2018 1st International Conference on*. IEEE, 281–285.

[4] Anthony Carrard, Annick Salzmann, Alain Malafosse, and Felicien Karege. 2011. Increased DNA methylation status of the serotonin receptor 5HTR1A gene promoter in schizophrenia and bipolar disorder. *Journal of Affective Disorders* 132(3) (2011), 450–453.

[5] William Jay Conover. 1973. On Methods of Handling Ties in the Wilcoxon Signed-Rank Test. *J. Amer. Statist. Assoc.* 68, 344 (1973), 985–988.

[6] Bolin Ding, Harsha Nori, Paul Li, and Joshua Allen. 2018. Comparing population means under local differential privacy: with significance and power. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[7] Vito D'Orazio, James Honaker, and Gary King. 2015. Differential Privacy for Social Science Inference. (2015).

[8] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our Data, Ourselves: Privacy Via Distributed Noise Generation, In Advances in Cryptology (EUROCRYPT 2006). 4004, 486–503. https://www.microsoft.com/en-us/research/publication/our-data-ourselves-privacy-via-distributed-noise-generation/

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*. Springer, 265–284.

[10] Morten Fagerland, Leiv Sandvik, and Petter Mowinckel. 2011. Parametric Methods Outperformed Non-Parametric Methods in Comparisons of Discrete Numerical Variables. *BMC Medical Research Methodology* 11 (04 2011), 44.

[11] Stephen E Fienberg, Aleksandra Slavkovic, and Caroline Uhler. 2011. Privacy preserving GWAS data sharing. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 628–635.

[12] Marco Gaboardi, Hyun-Woo Lim, Ryan M Rogers, and Salil P Vadhan. 2016. 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.

[13] Marco Gaboardi, Ryan Rogers, and Or Sheffet. 2018. Locally private mean estimation: Z-test and tight confidence intervals. *arXiv preprint arXiv:1810.08054* (2018).

[14] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics* 4, 8 (2008), e1000167.

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

[16] Vishesh Karwa and Salil Vadhan. 2017. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908* (2017).

[17] William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621.

[18] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[19] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 111–125.

[20] Thông T Nguyên and Siu Cheung Hui. 2017. Differentially Private Regression for Discrete-Time Survival Analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1199–1208.

[21] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 75–84.

[22] John W Pratt. 1959. Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures. *J. Amer. Statist. Assoc.* 54, 287 (1959), 655–667.

[23] Ryan Rogers and Daniel Kifer. 2017. A new class of private Chi-square hypothesis tests. In *Artificial Intelligence and Statistics*. 991–1000.

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

[25] Adam Smith. 2008. Efficient, differentially private point estimators. *arXiv preprint arXiv:0809.4794* (2008).

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

[27] Eftychia Solea. 2014. Differentially Private Hypothesis Testing For Normal Random Variables. (2014).

[28] Marika Swanberg, Ira Globus-Harris, Iris Griffith, Anna Ritz, Adam Groce, and Andrew Bray. 2019. Improved Differentially Private Analysis of Variance. *Proceedings on Privacy Enhancing Technologies* (2019).

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

[30] Christine Task and Chris Clifton. 2016. Differentially Private Significance Testing on Paired-Sample Data. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 153–161.

[31] Caroline Uhlerop, Aleksandra Slavković, and Stephen E Fienberg. 2013. Privacy-preserving data sharing for genome-wide association studies. *The Journal of privacy and confidentiality* 5, 1 (2013), 137.

[32] Duy Vu and Aleksandra Slavkovic. 2009. Differential privacy for clinical trial data: Preliminary evaluations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*. IEEE, 138–143.

[33] Yue Wang, Jaewoo Lee, and Daniel Kifer. 2015. Revisiting Differentially Private Hypothesis Tests for Categorical Data. *arXiv preprint arXiv:1511.03376* (2015).

[34] Larry Wasserman and Shuheng Zhou. 2010. A statistical framework for differential privacy. *J. Amer. Statist. Assoc.* 105, 489 (2010), 375–389.

[35] Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin* 1, 6 (1945), 80–83.

[36] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83.