sciendo

Proceedings on Privacy Enhancing Technologies ; 2019 (3):310–330

Marika Swanberg, Ira Globus-Harris, Iris Griffith, Anna Ritz, Adam Groce, and Andrew Bray*

Improved Differentially Private Analysis of Variance

Abstract: Hypothesis testing is one of the most common types of data analysis and forms the backbone of scientific research in many disciplines. Analysis of variance (ANOVA) in particular is used to detect dependence between a categorical and a numerical variable. Here we show how one can carry out this hypothesis test under the restrictions of differential privacy. We show that the F-statistic, the optimal test statistic in the public setting, is no longer optimal in the private setting, and we develop a new test statistic F_1 with much higher statistical power. We show how to rigorously compute a reference distribution for the F_1 statistic and give an algorithm that outputs accurate p-values. We implement our test and experimentally optimize several parameters. We then compare our test to the only previous work on private ANOVA testing, using the same effect size as that work. We see an order of magnitude improvement, with our test requiring only 7% as much data to detect the effect.

Keywords: Differential privacy, hypothesis testing, ANOVA

DOI 10.2478/popets-2019-0049

Received 2018-11-30; revised 2019-03-15; accepted 2019-03-16.

1 Introduction

A universal and recurring challenge in scientific research is determining whether a measured effect is real. That is, researchers wish to determine if the effect observed in a particular dataset indicates a similar effect in the broader world from which the sample was drawn. The most common statistical tool to make this determination is a hypothesis test. The particular form of the hypothesis test is driven by the scientific question and the data at hand.

Hypothesis testing is a common tool in population association studies, where the goal is to identify whether genetic variation is associated with disease risk [1]. Consider a study looking at the effect of a particular gene's mutation on some health outcome (e.g., blood pressure or weight). The data may include the mutation status of that gene (which may harbor one or more mutations on one or both copies of DNA). The researcher's goal is to determine if the gene's mutation status has an impact on that health outcome. When the health outcome is measured using a numerical variable, the natural hypothesis test to use is the one-way ANOVA (analysis of variance) test, treating the gene's mutation status as a categorical variable.

The first step in conducting the ANOVA is to calculate a single number, the F-statistic, which measures the variation in group means compared to the variation in individual data points. The F-statistic is constructed so that if the expected value of the health outcome is the same in all groups, the expected value of F is 1. If this is not the case, the value can be dramatically larger. Seeing a high value of F, a researcher will compare this value to the *reference distribution*, i.e., the distribution of F that would occur if the gene had no impact on the health outcome. The result of this comparison is a p-value, the probability that the observed F could occur by chance. If the p-value is low, the analyst can conclude that this gene must indeed affect the given health outcome. (For more detail on how ANOVA is used in this setting, see [12].)

The analysis described above assumes that the researcher has full access to the database. However, there are many settings in medicine, psychology, education, and economics (not to mention private-sector data analysis) where the database is not available to the analyst due to privacy concerns. A well-established solution is to allow the researcher to issue queries to the data which are proven to satisfy differential privacy. Differential privacy requires the addition of random noise to statistical

Marika Swanberg: Mathematics Department, Reed College, E-mail: marlswanb@reed.edu
Ira Globus-Harris: Mathematics Department, Reed College, E-mail: irglobush@reed.edu
Iris Griffith: Mathematics Department, Reed College, E-mail: irisrose@reed.edu
Anna Ritz: Biology Department, Reed College, E-mail: aritz@reed.edu
Adam Groce: Mathematics Department, Reed College, E-mail: agroce@reed.edu
***Corresponding Author: Andrew Bray:** Mathematics Department, Reed College, E-mail: abray@reed.edu

BY-NC-ND

queries and guarantees that the results reveal very little about any individual's data.

In this paper we propose a new statistic for ANOVA, called F_1, that is specifically tailored to the differentially private setting. This statistic measures the same variations as the F statistic, but uses $|a-b|$ instead of $(a-b)^2$ to measure the distance between a and b. In the public setting the F_1 is a worse test statistic than the traditional F-statistic, but we show that in the private setting it has much higher power than the previously published differentially private F-statistic. That is, we show that it can detect effects with a little as 7% of the data that was previously required. (In one example, an effect that took 5300 data points to detect 90% of the time with $\epsilon = 1$ in the prior work takes only 350 data points to detect using our new hypothesis test.)

1.1 Contributions and organization

We first review differential privacy, hypothesis testing, and the body of work that lies at the intersection of the two fields (Section 2). In Section 3 we then present a new test statistic, F_1, for ANOVA in the private setting. While there is some work on differentially private hypothesis testing, designing a new test statistic explicitly tailored for compatibility with differential privacy has been done by few others [14].

In Section 3.2 we give a private algorithm for computing an approximation of F_1 by applying the Laplacian mechanism to the computation of several intermediate values. Section 3.3 then describes how to compute the correct reference distribution for F_1 to in order to compute accurate p-values, which are the end result used by practitioners. Computing the reference distribution is complicated by the fact that, unlike the traditional F-statistic, F_1 is not scale-free.

We implement the private F_1-statistic and apply the method to different simulated datasets in Section 4. The computational experiments allow us to optimize ρ, a parameter that determines the allocation of our privacy budget between the two important intermediate values. We also compare our method to prior work [3], and show an order of magnitude improvement in statistical power.

Finally, in Section 5 we present a generalization of the F_1 statistic that allows for an arbitrary exponent in the distance measure (besides the absolute value from F_1 and the L^2-norm from the traditional F). We find that the L^1-norm used in the F_1 statistic is the most powerful across a wide range of scenarios.

2 Background

We begin by discussing hypothesis testing in general, the one-way analysis of variance (ANOVA) test in particular, and differential privacy. Readers familiar with one or more of these topics should be able to skip the relevant sections. We then discuss how these topics come together in private hypothesis testing and related work in this area.

2.1 Hypothesis testing

Hypothesis tests are common tools for making statistical inferences from data. The end goal of a hypothesis test is to determine whether a data set is consistent with a proposed model. This model is called the *null hypothesis*, denoted H_0, and it suggests a mechanism by which the data could have been generated. The mechanism is chosen to be scientifically meaningful, for example: the variable of interest has the same distribution across all of the treatment groups.

The comparison between H_0 and an observed data set is made using a *test statistic.* A test statistic f is simply a function from the data set to the real numbers. The goal is to design a test statistic with a known distribution when the data comes from H_0, but which will follow a markedly different distribution under other scenarios. The question then becomes, for a given database $\mathbf{x}$ with $f(\mathbf{x}) = t$, how likely is a value at least as extreme as t to occur if $\mathbf{x}$ was drawn from H_0. To compute this probability, we need to compare t to the reference distribution.

Definition 1 (Reference Distribution)**.** *Suppose f is a function that computes a test statistic. The reference distribution for f is the probability distribution of the statistic T when $T = f(\mathbf{X})$ and $\mathbf{X}$ is drawn from a distribution consistent with H_0.*

This reference distribution is used to calculate a p-value. A p-value is the probability, under the reference distribution, of drawing a statistic at least as extreme as the observed statistic.

Definition 2 (p-value)**.** *For a given test statistic $t = f(\mathbf{x})$ and null hypothesis H_0, the p-value is defined as*

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

The p-value provides context for the observed statistic by positioning it in the range of statistics that could be observed under H_0.

Typically, researchers choose a significance threshold α and reject the null hypothesis when their calculated p-value is less than α. The α-level determines the probability of a *type I error*, which occurs when an analyst rejects the null hypothesis despite it being true. The value of the statistic that demarcates this rejection region is called the *critical value*, denoted by t^*. That is, $\Pr[T \geq t^* \mid T = f(\mathbf{X}) \text{ and } \mathbf{X} \leftarrow H_0] = \alpha$.

When one develops a test statistic, a primary goal is to maximize *statistical power*. The power of a test quantifies how effectively it can detect a deviation from H_0. It is the probability of rejecting when H_0 is false. Generally, the power depends on both the amount of data and the effect (i.e., how different the true distribution H_A is from H_0).

Definition 3 (Statistical Power)**.** *For a specific alternate hypothesis H_A, the statistical power of a hypothesis test is*

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

2.2 One-way ANOVA

Consider again our example from earlier, where a researcher has, for a set of individuals, both blood pressure measurements and the mutation status of a particular gene. This is a classic setting for a one-way analysis of variance (ANOVA) test.

Index each individual person or observation in a database $\mathbf{x}$ with $i \in \{1, \ldots, N\}$. Each i is associated with a group or category c_i (e.g., mutation status) and a numerical value $y_i \in \mathbb{R}$ (e.g., blood pressure). We use $\bar{y}$ to represent the mean of all N numerical values. Index each group with $j \in \{1, \ldots, k\}$. Each j is associated with $C_j = \{i \mid c_i = j\}$, the set of indices of observations in group j. Denote the size of the set n_j and the mean of the values in that set $\bar{y}_j$. (Note that this means $\bar{y}_{c_i}$ is the mean of numerical values in the same group as observation i.)

The null hypothesis in a one-way ANOVA is that the y_i follow identical normal distributions regardless of their group. This motivates the test statistic used in an ANOVA test, the F-statistic, which measures the ratio of the variation between group means (weighted by the size of the groups) and the variation of individuals within each group. If all groups had equal means, the variation between group means would be proportional to the variation between individual observations.[1]

Definition 4 (F-Statistic)**.** *Given a database $\mathbf{x}$ with k groups and N total entries, the F-statistic is the ratio of two values, traditionally called $SSA(\mathbf{x})$ and $SSE(\mathbf{x})$. The Sum of Squared errors of All category means (SSA) is a measure of variance between group means, weighted by group size:*

$$SSA(\mathbf{x}) = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2.$$

The Sum of Squared Errors of all observations (SSE) is a measure of variance within groups:

$$SSE(\mathbf{x}) = \sum_{i=1}^{N} (y_i - \bar{y}_{c_i})^2.$$

The F-statistic is the ratio of SSA and SSE, each divided by their respective degrees of freedom. These adjusted values are called the Mean Sum of All category errors (MSA) and Mean of Sum of Squared Errors (MSE). We can now finish defining the F-statistic.

$$F(\mathbf{x}) = \frac{SSA(\mathbf{x})/(k-1)}{SSE(\mathbf{x})/(N-k)} = \frac{MSA(\mathbf{x})}{MSE(\mathbf{x})}.$$

Under the null hypothesis that the y_i follow identical normal distributions regardless of their group, the reference distribution of the F-statistic is known exactly. This comes from recognizing that $SSA(\mathbf{x})$ is drawn from $\sigma^2\chi^2_{k-1}$, the chi-squared distribution with $k-1$ degrees of freedom scaled by within-group variance, and $SSE(\mathbf{x})$ is drawn from $\sigma^2\chi^2_{N-k}$. The ratio of these values, therefore, has a reference distribution that is scale-free (not dependent on σ) and can be calculated knowing only N and k.

2.3 Differential privacy

Differential privacy is a security definition for the release of information about a database of private records. Here we outline the foundational definitions and theorems in differential privacy; everything below first appeared in the seminal paper of Dwork et al. [8].

---

[1] Our approach relies upon simulating normally distributed data in correspondence with the traditional normality assumption. The one-way ANOVA test is known to be robust to deviations from normality [15], so our approach should be applicable even in settings where the normality assumption is suspect. Readers interested in more about ANOVA generally are referred to [5].

Suppose we have a database $\mathbf{x}$ containing sensitive information that we want to study. In particular, we want to publish the output $f(\mathbf{x})$ of a function f (also sometimes called a *mechanism*) on our database while protecting the privacy of the individuals whose data was collected. Differential privacy promises that an adversary will learn approximately nothing about an individual as a result of their presence in $\mathbf{x}$. Informally, this is done by requiring that the probability of seeing any particular output is roughly the same regardless of what information a given individual submitted to the database.

As above, we use N to represent the number of rows in $\mathbf{x}$, where a "row" is simply the set of data associated with a single individual. We now define *neighboring databases*, which differ in only one row.

Definition 5 (Neighboring Databases). *Two databases $\mathbf{x}$ and $\mathbf{x}'$ are neighboring if $\mathbf{x}$ can be transformed to $\mathbf{x}'$ by changing only one individual's data (where a change is an in-place modification, not a full addition or removal).*

To protect the privacy of individuals in a database, differentially private requires that the output of a query on any two neighboring databases should look nearly identical.

Definition 6 (Differential Privacy). *A (randomized) mechanism f with range R is* ε-differentially private *if for all $S \subseteq R$ and for all neighboring databases $\mathbf{x}$ and $\mathbf{x}'$*

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

The parameter ε is called the *privacy parameter*, and its choice is a policy decision. The lower the chosen ε, the stronger the privacy guarantee. Note that because neighboring databases are the same size, N can always be released without compromising privacy.[2]

Differential privacy has several useful properties. One of the most useful is composition:

Theorem 1 (Composition). *Suppose f and g are respectively ε_1- and ε_2-differentially private mechanisms. Then, a mechanism h that returns the results of applying f and g to $\mathbf{x}$, $h(\mathbf{x}) = (f(\mathbf{x}), g(\mathbf{x}))$, is $(\varepsilon_1 + \varepsilon_2)$-differentially private.*

In other words, the privacy guarantee decreases, but does not disappear, when a database is queried multiple times. Composition allows database administrators to issue researchers a privacy budget, which researchers can then divide up as they wish between any number of different queries.

Another defining feature of differential privacy is its resistance to *post-processing*.

Theorem 2 (Post-Processing). *Let f be an ε-differentially private mechanism, and let g be an arbitrary function. Then, $h(\mathbf{x}) = g(f(\mathbf{x}))$ is also ε-differentially private.*

This theorem allows us to do any computation we desire on the output of our differentially-private mechanism without diminishing the privacy guarantees. We will utilize this property to compute p-values of the private F-statistic. The p-values will be automatically private without additional argument.

Our algorithms are constructed by taking building blocks and combining them with composition and post-processing, but the fundamental building blocks are made private using the Laplace mechanism, the oldest and maybe simplest method for achieving differential privacy. The Laplace mechanism allows the conversion of any function f into a private approximation $\hat{f}$. One must first compute (or bound) the *sensitivity* of the function, the maximum effect on the output that a single row can have.

Definition 7 (Sensitivity). *The sensitivity of a (deterministic) real-valued function f on databases is the maximum of $|f(\mathbf{x}) - f(\mathbf{x}')|$ taken over all pairs $(\mathbf{x}, \mathbf{x}')$ of neighboring databases.*[3]

The Laplace mechanism will use random noise drawn from the Laplace distribution.

Definition 8 (Laplace Distribution). *The Laplace Distribution (centered at 0) with scale b is the dis-*

---

[2] Differential privacy can also be defined in terms of databases that differ by an addition/deletion, rather than by a change in a row. For most applications these definitions are equivalent except for a change in ϵ by a factor of two (with the version here being the more stringent interpretation of ϵ). The one significant difference is that the other definition does not result in N being public, which is important for our work here.

[3] Sensitivity and the Laplace mechanism can be defined on functions with output in $\mathbb{R}^n$, but we only need the one-dimensional version.

tribution with probability density function

$$\mathsf{Lap}(z \mid b) = \frac{1}{2b}\exp\left(-\frac{|z|}{b}\right).$$

We use $\mathsf{Lap}(b)$ *to represent the Laplace distribution with scale* b*.*

We can now present the Laplace mechanism.

Theorem 3 (Laplacian Mechanism). *Let* f *be a function with sensitivity bounded above by* s*. Let* L *be a random variable drawn from* $\mathsf{Lap}(s/\varepsilon)$*. Then the function* $\hat{f}(\mathbf{x}) = f(\mathbf{x}) + L$ *is* ε*-differentially private.*

2.4 Differentially private hypothesis testing

In order to create a differentially private hypothesis test, we need a private function f of a database to serve as our test statistic. This could be a differentially private estimate of an existing test statistic, or it could be a new test statistic altogether. Because randomization is essential to differential privacy, f will be randomized. The same statistic on the same database may yield different outputs each time it is computed.

In addition to a test statistic f, we require a suitable reference distribution to calculate the corresponding p-value. While it may be tempting to compute the p-value using the reference distribution for the non-private statistic one is estimating, this may yield wildly inaccurate results [3], because adding noise to the statistic increases the probability of outlier output values. Instead, we must compute the reference distribution for the noisy statistic. Only then can we calculate an accurate p-value.

The goal of differentially private hypothesis testing is to create a private test statistic and method of computing the p-value that maximizes statistical power, ideally approaching the power of the equivalent test in the classical non-private setting.

2.5 Related work

There has been a moderate amount of work on differentially private hypothesis testing, but because there are many hypothesis tests most individual tests have received only a small amount of attention, and some very common tests have not seen a private analogue developed at all.

Several papers have addressed testing the value of a mean or the difference of means [6, 7, 19]. Hypothesis tests using coefficients of a linear regression to test for dependence between continuous variables is extremely common in many academic disciplines, but only recently has a method for carrying this analysis out privately been developed [2, 16]. and Nguyên and Hui propose a test for surival analysis data [13]. There is one prior work on private ANOVA testing, that of Campbell et al. [3]. We will discuss this result in greater depth in the next section.

The chi-squared test, which tests for the independence of two categorical variables, has received the most study. Vu and Slavković [21] give an analogue to the test and also compute accurate p-values. Many private chi-squared tests have been specifically motivated by genome-wide association studies (GWAS) [9, 11, 20]. These give p-value calculations, but they are only accurate in the limit as N grows large. Other work has used Monte Carlo simulations (as we do in this work) to give more accurate reference distributions for small N [10, 22]. Rogers and Kifer [14] instead propose a new statistic with an asymptotic distribution more similar to its non-private analogue. We note that this is one of few papers that, like the present work, proposes test statistics intended for the private setting, rather than simply approximating the accepted test statistic from the classic public setting. Very few of these papers carefully measure the power of the test they develop. Rogers and Kifer [14] and Gaboardi et al. [10] are notable exceptions, giving power curves for several different approaches.

There is also a significant body of work looking at how quickly private approximations of test statistics converge to their limiting distributions (e.g., [17, 18, 23]). These are important theoretical results, but they do not usually yield practical tests. Unless N is very large (in which case the details of the test do not matter very much anyway) the distribution of the test statistic is not close enough to that of the standard public to allow accurate computation of p-values.

2.6 Prior work on private ANOVA

The only previous work on differentially private ANOVA testing that the authors are aware of is Campbell et al. [3] Using the ANOVA test as defined above, they analyze the sensitivity of the *SSA* and *SSE* with the assumption that all data was normalized to be between 0 and 1 and add Laplacian noise proportional to these sensitivities to the public computation of the *SSA* and *SSE*. Their algorithm then uses post-processing to calculate the noisy F-statistic, and returns this in addition to the noisy *SSA* and *SSE* (Algorithm 1).

Algorithm 1 private_F($\mathbf{x}, \epsilon$)

```
Compute $\widehat{\mathrm{SSA}} = \mathrm{SSA} + Z_1$ where $Z_1 \sim \mathrm{Lap}\left(\frac{7-9/N}{\varepsilon/2}\right)$
Compute $\widehat{\mathrm{SSE}} = \mathrm{SSE} + Z_2$ where $Z_2 \sim \mathrm{Lap}\left(\frac{5-4/N}{\varepsilon/2}\right)$
Compute $\widehat{F} = \frac{\widehat{\mathrm{SSA}}/(k-1)}{\widehat{\mathrm{SSE}}/(N-k)}$
return $\widehat{F}, \widehat{\mathrm{SSA}}, \widehat{\mathrm{SSE}}$
```

Normally, the F distribution is used to calculate a p-value for the F-statistic. However, Campbell et al. find that the distribution of the private estimate $\widehat{F}$ differs too much from the F distribution for this to be acceptable. Furthermore, they find that it is no longer scale-free, meaning that the distribution depends on the within-group variance σ^2.

Fortunately, the *SSE* is an estimate of σ, so using this estimate they computed an estimated distribution on $\widehat{F}$ through simulation. They could then compare a given value of $\widehat{F}$ to this distribution to obtain a p-value.

To assess the power of their private ANOVA algorithm, they simulate databases with three equal-sized groups with values drawn from $\mathcal{N}(0.35, 0.15)$, $\mathcal{N}(0.5, 0.15)$, and $\mathcal{N}(0.65, 0.15)$ respectively. For several (N, ε)-pair choices, they generate many sets of data, apply the private ANOVA test, calculate the p-value, and record the percentage of simulations with p-values less than 0.05. They find that when $\varepsilon = 1$, they need over five thousand data points to detect this effect (compared to two or three dozen data points in the public setting). Our goal in this paper is to reduce the gap between the public and private setting.

3 A new test statistic

In this section we describe our hypothesis test. This begins with the introduction in Section 3.1 of F_1, a new test statistic for the ANOVA setting. In Section 3.2 we then show how to privately compute a private approximation of F_1. Finally, in Section 3.3 we calculate p-values for the private F_1 statistic. This means simulating a correct reference distribution against which we can compare our output.

3.1 The F_1 statistic

Our goal is to define a statistic that releases similar information as the F-statistic, but has higher power (Definition 3) for reasonable privacy guarantees. We focused on two approaches to improve the power of the ANOVA calculation: reducing the amount of Laplacian noise by decreasing sensitivity, and making *SSA* and *SSE* numerically larger so that the noise has less influence over the total value of the statistic. We achieved both goals by taking the absolute values of the summand terms in the *SSA* and *SSE*, rather than squaring them.

As before, let k denote the number of categories in the database, c_i be the category and y_i be the numerical value associated with observation i, n_j be the size of category j, and N be the total size of the data set. Additionally, $\bar{y}$ is the grand mean of the entries in database $\mathbf{x}$ and $\bar{y}_j$ is the mean of the entries in group j.

Definition 9 (F_1-statistic)**.** *Given a database* $\mathbf{x}$ *with* k *groups and* N *total entries, the* $SA(\mathbf{x})$ *and* $SE(\mathbf{x})$ *calculations are defined as follows:*

$$SA(\mathbf{x}) = \sum_{j=1}^{k} n_j |\bar{y} - \bar{y}_j|$$

$$SE(\mathbf{x}) = \sum_{i=1}^{N} |y_i - \bar{y}_{c_i}|.$$

The F_1*-statistic is the ratio of SA and SE, each divided by their respective degrees of freedom.*

$$F_1(\mathbf{x}) = \frac{SA(\mathbf{x})/(k-1)}{SE(\mathbf{x})/(N-k)}.$$

The F_1-statistic measures variation between group means compared to variation within groups in essentially the same way as the original F-statistic. The *SA* grows as the group means diverge. What constitutes a "large" variation between group means depends on the variation between individual items, so *SE*, which measures this individual-level variation, provides a sense of scale for the *SA* value.

In the next section, we show that the sensitivities of *SA* and *SE* in the F_1-statistic are less than half as large as the sensitivities of *SSA* and *SSE* in the original F-statistic. Further, because the summand terms are restricted to $[0, 1]$, the *SA* and *SE* values are larger, meaning they can tolerate the addition of more noise before losing their usefulness.

3.2 A private approximation of F_1

The sensitivity of F_1 is very high. (In the worst case, *SE*($\mathbf{x}$) is almost zero and very small changes can have huge effects on $F_1(\mathbf{x})$.) As a result, we can't simply apply the Laplace mechanism to F_1. Instead, we choose to apply it individually to the *SA* and *SE* functions, and then use composition and post-processing to compute an estimate of F_1. We must therefore bound the sensitivities of *SA* and *SE*.

We assume that the number of valid category values, k, is fixed and public, but the number of entries in each group is not. (This includes the possibility that one or more categories exist as valid entries but do not appear in the actual database.) We also assume that there are maximum and minimum possible values for the data, and that the computation first uses these to normalize the data, mapping it to the interval $[0,1]$.

Theorem 4 (SE Sensitivity)**.** *The sensitivity of the SE calculation in Definition 9 is bounded above by 3.*

Proof. Suppose neighboring databases $\mathbf{x}$ and $\mathbf{x}'$ differ by some row r Say that in $\mathbf{x}$, $c_r = a$, and in $\mathbf{x}'$, $c_r = b$. Let n_a be the size of category a excluding r and let n_b be the size of category b excluding r. We begin by expressing the *SE* calculation as nested summations indexing over group size and entries within each group.

$$SE(\mathbf{x}) = \sum_{j=1}^{k} \sum_{i \in C_j} |y_i - \bar{y}_{c_i}|.$$

Call $t_i = |y_i - \bar{y}_{c_i}|$, and let $c_i \neq a, b$. Then, t_i does not change between $\mathbf{x}$ and $\mathbf{x}'$. Now, suppose $i \neq r$ but $c_i = a$. It follows that $\Delta t_i \leq 1/(n_a+1)$, since the only change comes from $\bar{y}_a$. There are n_a such terms, so the total contribution from these terms is at most $n_a/(n_a+1)$. If $a \neq b$, we must also consider t_i where $i \neq r$ but $c_i = b$, for which we have $\Delta t_i \leq 1/(n_b+1)$. Thus, the terms in groups a and b excluding row r together contribute

$$\frac{n_a}{n_a+1} + \frac{n_b}{n_b+1} \leq 2$$

if $a \neq b$ and just $n_a/(n_a+1) < 1$ otherwise.

Now, consider Δt_r. Since y_r, $\bar{y}_a$, and $\bar{y}_b$ are all in the interval $[0,1]$, the difference between t_r in database $\mathbf{x}$ and in $\mathbf{x}'$ is at most 1. Thus, the total sensitivity of *SE* is bounded above by 3.

□

Theorem 5 (SA Sensitivity)**.** *The sensitivity of the SA calculation in Definition 9 is bounded above by 4.*

Proof. Again, suppose neighboring databases $\mathbf{x}$ and $\mathbf{x}'$ differ by some row r in both the categorical and numerical values, with $c_r = a$ in $\mathbf{x}$, and $c_r = b$ in $\mathbf{x}'$. Denote the number of entries in groups a and b not including row r by n_a and n_b, respectively. We begin by expressing the *SA* calculation as two sums indexing over groups and entries within each group.

$$SA(\mathbf{x}) = \sum_{j=1}^{k} \sum_{i \in C_j} |\bar{y} - \bar{y}_{c_i}|$$

Consider the change from $SA(\mathbf{x})$ to $SA(\mathbf{x}')$ as if it occurred in two steps. In the first step, the grand mean $\bar{y}$ is updated. Note that in the worst case, $\bar{y}$ can change by at most $1/N$ between $\mathbf{x}$ and $\mathbf{x}'$. Then, since there are N summands, each including the grand mean, this step changes the value by a maximum of $N(1/N) = 1$ to the overall sensitivity.

In the second step we change the group means for groups a and b. There will be n_a terms containing $\bar{y}_a$, each of which will change by at most $1/n_a$, changing the overall value by at most 1. Similarly updating $\bar{y}_b$ changes n_b terms each by at most $1/n_b$ for a total contribution of 1.

Finally, the change of r between $\mathbf{x}$ and $\mathbf{x}'$ contributes 1 to the overall sensitivity, and thus the sensitivity of the *SA* is bounded above by 4. □

Having proven these sensitivities, we can now introduce our private algorithm (Algorithm 2) for approximating F_1. The algorithm first estimates *SA* and *SE*, allocating part of the privacy budget to each one. We introduce a parameter $\rho \in (0,1)$ that determines the relative amount of the privacy budget spent on each intermediate value. The optimal value of ρ will be experimentally determined in Section 4.2. We note that in addition to using a different test statistic, the prior work by Campbell et al. did not consider ρ values other than 0.5.

Algorithm 2 private_F1($\mathbf{x}, \varepsilon, \rho$)

```
$\widehat{\mathrm{SA}} = \mathrm{SA}(\mathbf{x}) + L_1$ where $L_1 \sim \mathrm{Lap}\left(\frac{4}{\rho\varepsilon}\right)$
$\widehat{\mathrm{SE}} = \mathrm{SE}(\mathbf{x}) + L_2$ where $L_2 \sim \mathrm{Lap}\left(\frac{3}{(1-\rho)\varepsilon}\right)$
$\widehat{F_1} = \frac{\widehat{\mathrm{SA}}/(k-1)}{\widehat{\mathrm{SE}}/(N-k)}$
return $\widehat{F_1}, \widehat{\mathrm{SA}}, \widehat{\mathrm{SE}}$
```

Theorem 6. *Algorithm 2 is ε-differentially private.*

Proof. By the sensitivity bounds of SE and SA in Theorems 4 and 5 and the Laplace mechanism (Theorem 3), $\widehat{SA}$ is $\rho\varepsilon$-differentially private and $\widehat{SE}$ is $(1-\rho)\varepsilon$-differentially private. By the composition theorem (Theorem 1), outputting both is ε-differentially private. Since k and N are both public information, computing $\widehat{F_1}$ is post-processing (Theorem 2). □

3.3 Reference distribution and p-values

As discussed previously, the test statistic on its own is not useful; we need a p-value to provide a sense of scale in the context of the null hypothesis. Computing a p-value begins with an accurate reference distribution. We numerically approximate this distribution through simulation. The intermediate values of the F statistic, SSA and SSE, are drawn from $\sigma^2\chi^2_{k-1}$ (for SSA) and $\sigma^2\chi^2_{n-k}$ (for SSE). Campbell et al. [3] used this to easily sample from the correct distributions for SSA and SSE.

The SA and SE values needed to compute F_1 follow no similarly tractable distribution, so instead we simulate full databases according to the null hypothesis and for each calculate $\widehat{F_1}$. The distribution of these $\widehat{F_1}$ values approximates the reference distribution. The goal, given a database $\mathbf{x}$, is to simulate databases with the same size N and number of groups k, same standard deviation σ, and same expected value μ. The values N and k are public, so we can use those values. The expected value μ is not, but as long as it is safely inside the $[0,1]$ interval, its value has no effect on the distribution, so we simply always use 0.5.

Unfortunately, using an accurate σ is more difficult. Unlike in the non-private setting, the reference distribution depends on the choice of σ, so an inaccurate value can cause incorrect results. We had two choices: either we could use some of our ε-budget to directly estimate the standard deviation of the y_i in database $\mathbf{x}$, or we could devise an indirect method of estimating the standard deviation given $\widehat{SA}$ and $\widehat{SE}$. See Appendix D for further discussion of the first option. (It has higher power for low n but takes longer to approach full power.) Here we focus on the second option by deriving an unbiased estimator $\hat{\sigma}$ for σ that can be computed from $\widehat{SE}$.

$$\hat{\sigma} = \frac{\widehat{SE}}{\tilde{N}}\sqrt{\frac{\pi}{2}}, \text{ where } \tilde{N} = \sum_{j=1}^{k} n_j \sqrt{\left(1 - \frac{1}{n_j}\right)}$$

See Appendix A for the proof that this is indeed an unbiased estimator.

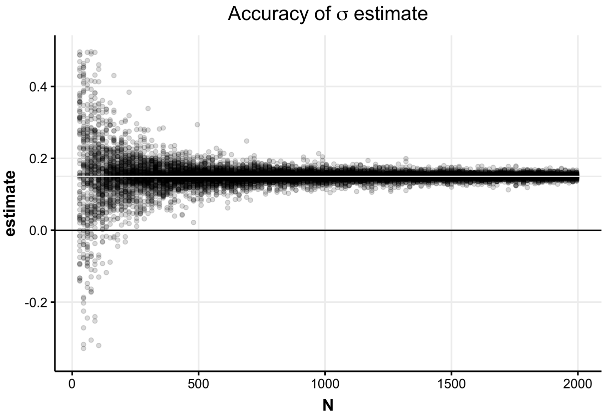

Fig. 1. $\hat{\sigma}$ is unbiased ($\sigma = .15$) with shrinking variance as N grows large. Each point represents the estimate from a simulated null database. At small N, there is a non-zero probability of returning a negative estimate.

Computing this estimator requires knowledge of each of the group sizes, which are private. Fortunately, $\tilde{N}$ is closely approximated by $N-k$, which we know. At the smallest database sizes that we considered (around $N = 100$ with three equal-sized groups), this approximation has $< 1\%$ error. As the size of the database grows, this error shrinks to zero. Figure 1 visualizes the unbiasedness and asymptotically shrinking variance of $\hat{\sigma}$ (using the $\tilde{N} = N-k$ approximation). We further confirm in Section 4 that it is precise enough to compute valid p-values.

Another issue presented by the private estimation of σ is that the Laplacian noise can be large enough to make the estimate negative. Negative standard deviation estimates are more likely to occur when the SE is small, i.e., when the database is small or when the within-group standard deviation is small.

This problem is unique to the private setting and we do the most conservative possible thing — we never reject the null hypothesis when the estimated standard deviation was negative. Our reasoning was that a negative standard deviation has no statistical meaning and any calculations made from such an estimate would be uninformative. This method for dealing with negative standard deviation estimates results in a type I error rate lower than α, which means that our results are more conservative. As was mentioned earlier, the SA increases as the database size and effect size between groups increases. Thus, negative standard deviation estimates tend to occur for database sizes and effect sizes that are so small that the effect is undetectable anyway.

As a result, this conservative choice does little to reduce the power of our test.

A formal description of the private F_1 ANOVA test in presented in Algorithm 3, which returns a Boolean indicating whether the null hypothesis H_0 is rejected.

Algorithm 3 ANOVA_test($\mathbf{x}$, ε, α, *reps*)

```
$\widehat{F_1}, \widehat{SA}, \widehat{SE}$ = private_F1($\mathbf{x}$,$\varepsilon$)
if $\widehat{\mathrm{SE}} < 0$ then
    return False
end if
$\widehat{\sigma} = \sqrt{\pi/2}\frac{\widehat{\mathrm{SE}}}{(N-k)}$
$significant = 0$
for $i = 1$ to $reps$ do
    $\mathbf{x}^{ref}$ = $N$ draws from $\mathcal{N}(0.5, \widehat{\sigma})$ divided into $k$ equal-sized groups
    $\widehat{F_1}^{ref}, \widehat{SA}^{ref}, \widehat{SE}^{ref}$ = private_F1($\mathbf{x}^{ref}, \varepsilon$)
    if $\widehat{F_1}^{ref} > \widehat{F_1}$ then
        $significant = significant + 1$
    end if
end for
$p$-value = $significant/reps$
if $p$-value $< \alpha$ then
    return True
else
    return False
end if
```

Theorem 7. *Algorithm 3 is ε-differentially private.*

Proof. This follows immediately from Theorem 6 (privacy of the test statistic) and Theorem 2 (post-processing). Using the notation of Theorem 2, the call on line 1 to private_F1 (Algorithm 2) is the function f, which is proven ϵ-differentially private by Theorem 6. The rest of the computation is the function g, and uses only the output of f. (Note that later calls to private_F1 call it on simulated data.) The whole of Algorithm 3 is therefore the composition of these two functions and as a result is itself private. □

4 Experimental results

In this section we assess the properties and performance of Algorithm 3 through simulation.

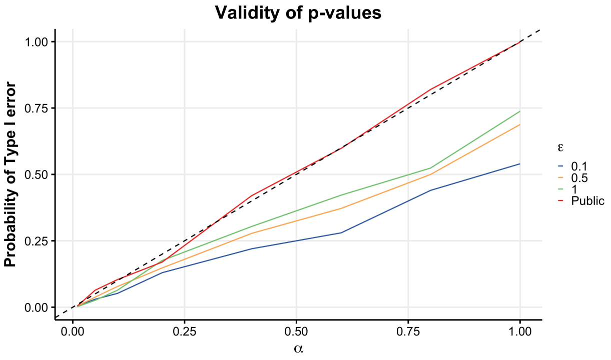

Fig. 2. The empirical type I error rate under three private scenarios is less than α while the public error rate is exactly α (within MC variability). Each point in a line represents 500 simulated tests, each with $N = 180$, $k = 3$, and equal-size groups.

4.1 Properties of the test

A p-value is considered valid if [4]

$$\Pr[p\text{-value} \leq \alpha \mid \mathbf{X} \leftarrow H_0] \leq \alpha.$$

In other words, the actual type I error rate must be less than or equal to α. This can be assessed by conducting many tests on simulated data sampled according to H_0 at various α and checking if the proportion of rejections is less than or equal to α.

Figure 2 presents these simulations for several choices of ϵ and demonstrates that the F_1 test produces valid p-values. The type I error rate is lowest at low ϵ values because the high privacy guarantee requires a large amount of noise be added to SE, which can lead to a negative estimate for σ, which in turns leads to an automatic decision to retain H_0 (see section 3.3).

We note also that when we simulate data for calculating a reference distribution, we always simulate data with equal-size groups. I.e., we must confirm that the critical value of the reference distribution is highest when groups are of equal size. Fortunately, this appears to be the case. Appendix B contains both experimental and theoretical arguments for this claim, though not a complete analytic proof.

4.2 Optimal ρ

The computation of the private F_1 statistic requires the specification of $\rho \in (0, 1)$, the parameter that determines the proportion of the privacy budget that is allocated to SA relative to SE. We determined an optimal value for ρ by constructing power plots comparing database size to

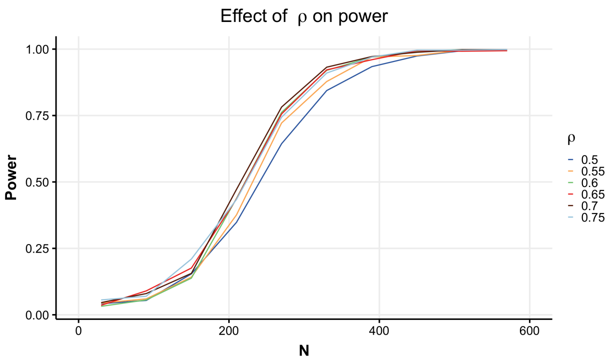

Fig. 3. Power curves at varying ρ in a setting where $\epsilon = 1$, $k = 3$, $\sigma = .15$, and effect size: 1σ. Power is experimentally maximized when $\rho = .7$.

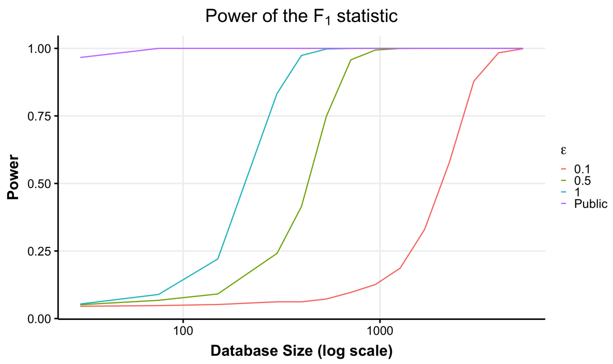

Fig. 4. Power curves at varying privacy levels in a simulation setting where $k = 3$, $\sigma = .15$, and effect size: 1σ.

power for different ρ values. We began by exploring the full range from 0.1 to 0.9 by 0.1-increments to get a sense of the range of variability in power. After that initial pass, we tuned in to the value with higher precision. We considered many effect sizes and found that in every case $\rho \approx 0.7$ was the most powerful.

Figure 3 is an example of one of the many scenarios that were considered, which identifies .7 as the optimal value. The scale of the effect of ρ on power is not dramatic but it was persistent across scenarios.

4.3 Performance of the test: power

There can be many tests for a given scientific setting that generate valid p-values and have identical type I error rates. What distinguishes them is their statistical power, or the probability that they reject H_0 when the database comes from a distribution under H_A.

The most common way to assess the power of a test is to generate a plot of power as a function either of database size or of effect size. Figure 4 fixes the effect size and shows power curves as a function of database size for four choices of ϵ. An ideal test would very quickly develop power near 1 with very little data. In our private setting, it is clear that the cost of high privacy ($\epsilon = .1$) is roughly an order of magnitude more data than modest privacy ($\epsilon = 1$) to achieve high power. Our private test (for reasonable values of ϵ) still requires much more data than the public version, hundreds of data points as opposed to dozens.

Each point in a line of a power curve was computed from 10,000 simulations, each based on a synthetic database of a set sample size and a set effect size (distance between the group means). We considered the effect H_A where $k = 3$ groups, each distributed $\mathcal{N}(0.35, 0.15), \mathcal{N}(0.5, 0.15), \mathcal{N}(0.65, 0.15)$ and equal group sizes, which was the same scenario as Campbell et al. [3]. Each simulation starts with a draw from H_A and computes a p-value as described in Algorithm 3.

The most relevant comparison, however, is between the private F_1 statistic and the privatized version of the classical F_2 statistic proposed in Campbell et al., the only prior private version of ANOVA. This comparison can be seen in Figure 5, and the improvement is substantial. For example, at $\epsilon = 1$, if one wanted to collect enough data to detect this effect with 80% probability, one would need 4500 observations if using the prior best test, whereas with the F_1 test presented here one would need only 300 data points, a 15-fold reduction in the necessary amount of data.

Figure 5 also demonstrates the degree to which the F_1 is well-suited to the private setting; the greatest improvement in power occurs under high privacy ($\epsilon = 0.1$). As ϵ grows large, the difference between the two statistics shrinks. In the public setting, they are nearly indistinguishable, though the F_2 is narrowly more powerful at every database size.[4]

The improvement in power is the greatest practical contribution of our work: the ability to conduct a private ANOVA with an order of magnitude less data than the existing approach. This improvement can be attributed to two key characteristics of our new test. The first and most important is the notion of measuring distance using the L^1 norm. The second is the unequal

---

[4] This empirical result is consistent with theoretical results in the classical statistics literature that discuss conditions in which the F_2 is a most powerful test [5].

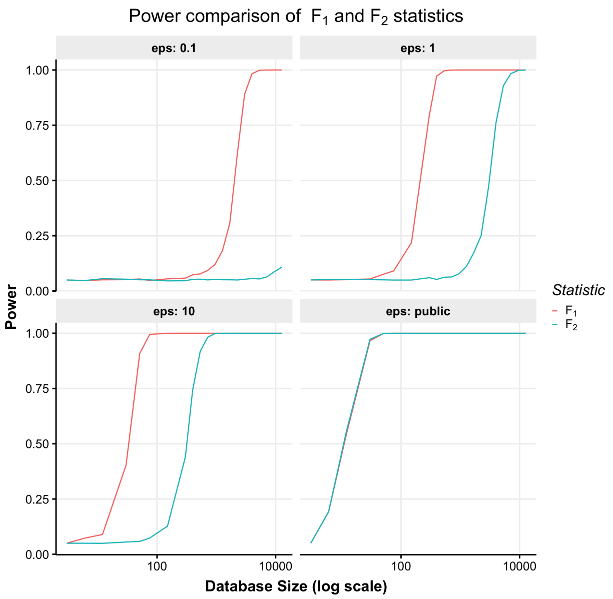

Fig. 5. Comparison of the power of the new F_1-statistic and the prior state of the art test using F_2, for three values of ϵ and in the public setting.

apportionment of the privacy between the *SA* and *SE* terms.

5 Other considerations

In this section we present a straightforward generalization of the classic F and F_1 to allow other exponents. In Section 5.1 we show that in this generalized class of statistics F_1 is indeed optimal. In Section 5.2 we discuss the experimental exploration of the full parameter space that guides the formulation of the F_1 statistic.

5.1 Varying the exponent

As seen is the previous sections, the change from squaring the differences in the original F-test to taking their absolute value with the F_1-statistic improved power significantly in the differentially private setting. It is not obvious that switching the exponent from 2 to 1 is optimal — perhaps some other exponent is superior. An exponent of 0 is clearly horrible, so there must be a local maximum in the power of the statistic for some exponent between 0 and 2.

In order to determine which exponent is in fact optimal, we further generalize the notion of an F-test. We define SQA and SQE, which are equivalent to SSA and SSE except that the summand is raised to the q^{th} exponent, and we call the resulting statistic F_q. Note that F_1 (as defined earlier) is a special case of F_q for $q = 1$, and F_2 is the standard F-statistic.

Definition 10 (F_q)**.** *Given a database* $\mathbf{x}$ *with* k *groups and* N *total entries, define* SQA *and* SQE *as follows:*

$$SQA(\mathbf{x}) = \sum_{j=1}^{k} n_j \left|\overline{y}_j - \overline{y}\right|^q$$

$$SQE(\mathbf{x}) = \sum_{i=1}^{N} \left|y_i - \overline{y}_{c_i}\right|^q$$

Then, F_q *is defined as*

$$F_q(\mathbf{x}) = \frac{SQA/(k-1)}{SQE/(N-k)}$$

We must now create a private approximation of F_q for arbitrary q. To do this, we first bound the sensitivity of the SQA and SQE with the following two theorems, the proofs of which can be found in Appendix C.

Theorem 8 (SQE Sensitivity)**.** *The sensitivity of* SQE *is bounded above by*

$$2\left(\frac{N}{2}\right)^{(1-q)} + 1$$

when $q \in (0, 1)$ *and*

$$N - N\left(1 - \frac{2}{N}\right)^q + 1$$

when $q \geq 1$*. Note that both give an upper bound of 3 when* $q = 1$*.*

Theorem 9 (SQA Sensitivity)**.** *The sensitivity of* SQA *is bounded above by*

$$N\left(\frac{3}{N}\right)^q + 1$$

when $q \in (0, 1)$ *and*

$$N - N\left(1 - \frac{3}{N}\right)^q + 1$$

when $q \geq 1$*. Note that both give an upper bound of 4 when* $q = 1$*.*

Given these sensitivity bounds, we can calculate a private approximation of F_q for simulated data using the same algorithm as for F_1, but with the sensitivities altered according to the choice of q. We can also simulate

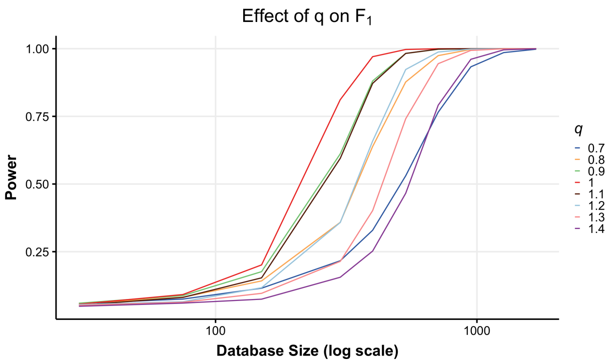

Fig. 6. Power curves at varying exponents in a simulation setting where $k = 3$, $\sigma = .15$, and effect size: 1σ. Power is experimentally maximized when $q = 1$.

a reference distribution by adapting Algorithm 3, which was done to construct Figure 6. As is clear from these results, in terms of power, the optimal value of q is 1.

We note that in the computation shown in Figure 6, we don't estimate σ using SQE. This is because we have not developed an estimator for σ that can be computed from SQE (for $q \neq 1, 2$). If another value of q had indeed been optimal, the next step would have been to find such an estimator and confirm it was accurate enough to produce acceptable p-values. But since the power cannot possibly improve when switching to an estimated σ value, this result is sufficient to show that other values of q need not be considered.

5.2 Parameter tuning

With the generalization of the F-statistic, we add q to the list of parameters that determine the power of a testing procedure. The parameters can be organized as follows:

Data Generation: N, k, σ, effect size
Private Algorithm: ϵ, q, ρ

The analyst gets to select the parameters corresponding to the private algorithm. While ϵ is set based on privacy concerns, q and ρ should be set to maximize power, which our work suggest occurs at roughly 1 and 0.7, respectively. These conclusions are based upon extensive exploration of the parameter space, a selection of which can be found in Appendix E.

The salient feature of these plots is that the choice of q is much more consequential than the choice of ρ. In the setting where $\epsilon = .1$, we found that the result seen in Figure 5 – a greater than 10-fold reduction in database size to get equivalent power – holds across a range of difference data generation parameters.

By contrast, the effect of ρ on power is much smaller; the database size reduction is closer to 1.1- or 1.2-fold when moving from $\rho = .5$ to $\rho = .7$ when $\epsilon = .1$.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. SaTC-1817245 and by the Gillespie Family Student Research Fund.

References

[1] David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781, 2006.
[2] Andrés F Barrientos, Jerome P Reiter, Ashwin Machanavajjhala, and Yan Chen. Differentially private significance tests for regression coefficients. *arXiv preprint arXiv:1705.09561*, 2017.
[3] Zachary Campbell, Andrew Bray, Anna M. Ritz, and Adam Groce. Differentially private anova testing. *1st International Conference on Data Intelligence and Security (ICDIS)*, pages 281–285, 2018.
[4] George Casella and Roger Berger. *Statistical Inference*. Brooks/Cole, Belmont, CA, 2 edition, 2002.
[5] D. R. Cox. *Theoretical statistics*. Chapman and Hall, London, 1974.
[6] Bolin Ding, Harsha Nori, Paul Li, and Joshua Allen. Comparing population means under local differential privacy: with significance and power. *arXiv preprint arXiv:1803.09027*, 2018.
[7] Vito D'Orazio, James Honaker, and Gary King. Differential privacy for social science inference. 2015.
[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876, pages 265–284. Springer, 2006.
[9] Stephen E Fienberg, Aleksandra Slavkovic, and Caroline Uhler. Privacy preserving gwas data sharing. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 628–635. IEEE, 2011.
[10] Marco Gaboardi, Hyun-Woo Lim, Ryan M Rogers, and Salil P Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *ICML*, pages 2111–2120, 2016.
[11] Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1087. ACM, 2013.

[12] Jerome Meyers and Arnold Well. *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, London, 2 edition, 2003.

[13] Thông T Nguyên and Siu Cheung Hui. Differentially private regression for discrete-time survival analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1199–1208. ACM, 2017.

[14] Ryan Rogers and Daniel Kifer. A New Class of Private Chi-Square Hypothesis Tests. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 991–1000, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[15] Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Buhner. Is it really robust? reinvestigatin the robustness of anova against violations of the normal distribution assumption. *Methodology*, pages 147–151, 2010.

[16] Or Sheffet. Differentially private ordinary least squares. *arXiv preprint arXiv:1507.02482*, 2015.

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

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

[19] Eftychia Solea. Differentially private hypothesis testing for normal random variables. 2014.

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

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

[22] Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

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

A Appendix: An unbiased estimator for σ

We are concerned with finding an unbiased estimator for σ given $\widehat{SE}$.

$$\widehat{SE} = \left(\sum_{j=1}^{k} \sum_{i \in C_j} |y_i - \bar{y}_{c_i}| \right) + L_2 \tag{1}$$

Our first goal is to find $E(\widehat{SE})$. We begin by noting several relevant distributions.

$$\begin{aligned} y_i &\sim N(\mu, \sigma^2) \\ \bar{y}_{c_i} &\sim N(\mu, \sigma^2/n_{c_i}) \\ y_i - \bar{y}_{c_i} &\sim N(0, \tau^2). \end{aligned}$$

The term τ^2 will be useful in determing the expected value of each term in $\widehat{SE}$, so we seek to express it in terms of known quantities.

$$\begin{aligned} \tau^2 &= Var(y_i - \bar{y}_{c_i}) \\ &= Var(y_i) + Var(\bar{y}_{c_i}) - 2Cov(y_i, \bar{y}_{c_i}) \\ &= \sigma^2 + \frac{\sigma^2}{n_{c_i}} - 2Cov(y_i, \bar{y}_{c_i}). \end{aligned} \tag{2}$$

To solve for the covariance, noting that $Cov(X, Y) = E(XY) - \mu^2$, we start by finding

$$\begin{aligned} E(y_i \bar{y}_{c_i}) &= \frac{1}{n_{c_i}} E(y_i(y_1 + y_2 + \ldots + y_i + \ldots + y_{n_{c_1}})) \\ &= \frac{1}{n_{c_i}} E(y_i^2 + y_i S_{-i}) \\ &= \frac{1}{n_{c_i}} \left(E\left(y_i^2\right) + E\left(y_i S_{-i}\right) \right), \end{aligned} \tag{3}$$

where $S_{-i} = y_1 + y_2 + \ldots + y_{i-1} + y_{i+1} + \ldots + y_{n_{c_i}}$. The term y_i^2 will be chi-square distributed when standardized as follows:

$$Z^2 = \left(\frac{y_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \left(y^2 - 2y_i\mu + \mu^2 \right).$$

Therefore we can write

$$\begin{aligned} E(y^2) &= E\left(\frac{\sigma^2}{\sigma^2} \left(y^2 - 2y_i\mu + \mu^2 \right) + 2y_i\mu - \mu^2 \right) \\ &= E\left(\sigma^2 Z^2 + 2y_i\mu - \mu^2 \right) \\ &= \sigma^2 E(Z^2) + 2\mu E(y_i) - \mu^2 \\ &= \sigma^2 + 2\mu^2 - \mu^2 \\ &= \sigma^2 + \mu^2. \end{aligned}$$

Continuing from Eq. (3),

$$\begin{aligned} \frac{1}{n_{c_i}} \left(E\left(y_i^2\right) + E\left(y_i S_{-i}\right) \right) &= \frac{1}{n_{c_i}} \left(\sigma^2 + \mu^2 + E(y_i)E(S_{-i}) \right) \\ &= \frac{1}{n_{c_i}} \left(\sigma^2 + \mu^2 + (n_{c_i} - 1)\mu^2 \right) \\ &= \frac{1}{n_{c_i}} \left(\sigma^2 + n_{c_i}\mu^2 \right) \\ &= \frac{\sigma^2}{n_{c_i}} + \mu^2. \end{aligned}$$

Now we can return to the covariance term in Eq. (2),

$$\begin{aligned} Cov(y_i, \bar{y}_{c_i}) &= E(y_i \bar{y}_{c_i}) - \mu^2 \\ &= \frac{\sigma^2}{n_{c_i}} + \mu^2 - \mu^2 \\ &= \frac{\sigma^2}{n_{c_i}}. \end{aligned}$$

Now we can finish the calculation of τ^2.

$$\begin{aligned} \tau^2 &= \sigma^2 + \frac{\sigma^2}{n_{c_i}} - 2Cov(y_i, \bar{y}_{c_i}) \\ &= \sigma^2 + \frac{\sigma^2}{n_{c_i}} - 2\frac{\sigma^2}{n_{c_i}} \\ &= \sigma^2 - \frac{\sigma^2}{n_{c_i}} \\ &= \sigma^2 \left(1 - \frac{1}{n_{c_i}}\right). \end{aligned}$$

The distribution of the absolute value of a normal random variable with mean 0 and variance τ^2 is half normal with a single parameter τ, which should properly be indexed by the observation. For ease of notation, call this random variable W_i.

$$W_i = |y_i - y_{c_i}| \sim HN(\tau_i),$$

where $E(W_i) = \tau_i \sqrt{\frac{2}{\pi}}$. To find $E(\widehat{SE})$, we can write it using the double sum notation as in Eq. (1). Since L_2 is a Laplace distribution centered at zero, we can use linearity of expectation to further simplify the expectation:

$$\begin{aligned} E(\widehat{SE}) &= E\left(\sum_{j=1}^{k} \sum_{i \in C_i} W_i\right) + E(L_2) \\ &= E\left(\sum_{j=1}^{k} \sum_{i \in C_i} W_i\right) + 0 \\ &= \sum_{j=1}^{k} E\left(n_j W_j\right). \end{aligned}$$

The change in indices is justified by realizing that W_i is the same for all n_{c_i} elements in C_i (in expectation).

Continuing,

$$\begin{aligned} \sum_{j=1}^{k} E\left(n_j W_j\right) &= \sum_{j=1}^{k} n_j E\left(W_j\right) \\ &= \sum_{j=1}^{k} n_j \tau_j \sqrt{\frac{2}{\pi}} \\ &= \sqrt{\frac{2}{\pi}} \sum_{j=1}^{k} n_j \sqrt{\sigma^2 \left(1 - \frac{1}{n_j}\right)} \\ &= \sigma \sqrt{\frac{2}{\pi}} \underbrace{\sum_{j=1}^{k} n_j \sqrt{\left(1 - \frac{1}{n_j}\right)}}_{\tilde{N}}. \end{aligned}$$

Denote the sum $\tilde{N}$, which is a number that approaches N as the group sizes get large. This allows us to express the expected value more concisely as

$$E(\widehat{SE}) = \sigma \sqrt{\frac{2}{\pi}} \tilde{N}.$$

The final step is to correct for this bias in our final estimator:

$$\begin{aligned} \hat{\sigma} &:= \frac{\widehat{SE}}{\tilde{N}} \sqrt{\frac{\pi}{2}}, \text{where} \\ E\left(\hat{\sigma}\right) &= \sigma. \end{aligned}$$

Using this exact estimator requires knowledge of each of the group sizes, which are private. Instead of dedicating part of the ϵ budget to this estimation, we used $N - k$ in place of $\tilde{N}$. At the smallest database sizes that we considered (around $N = 100$), this approximation accounts for $< 1\%$ error. As the size of the database grows, this error shrinks to zero.

B Appendix: Validity under unequal n_j

Consider a specific setting in which $N = 800$, $\sigma = 0.1$, and $k = 4$. Figure 7 shows the reference distributions of F_1 in four scenarios, each with a different group allocation of the 800 observations. The distribution in s_0, the equal group size scenario, generally takes the highest values; indeed it exceeds the other scenarios at every quantile. This represents the distribution that we use to calculate p-values and reject H_0 whenever an observed statistic is greater than the vertical dotted line (when $\alpha = .05$). This means that in the other three scenarios, featuring unequal allocation, the actual type I error rate

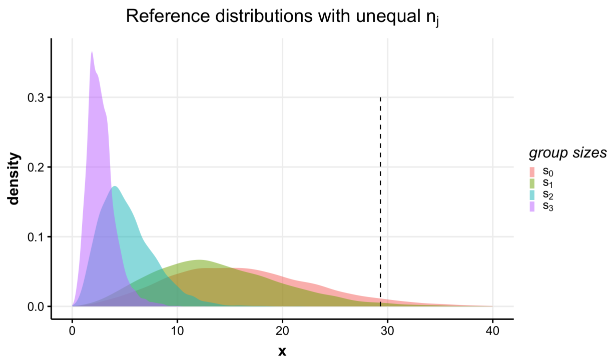

Fig. 7. Distribution of $\hat{F}_1$ under H_0 in four group size allocation scenarios, each with 10,000 simulations. $s_0\colon \{200, 200, 200, 200\}$, $s_1\colon \{100, 100, 100, 500\}$, $s_2\colon \{5, 10, 20, 765\}$, $s_3\colon \{3, 3, 3, 791\}$. The vertical dotted line indicates the .95 quantile of the equal group size distribution.

(the proportion of the distributions beyond the dotted line) will be less than α and we meet the condition for valid p-values.

In the following, we show that under any scenario, the expected value of the F_1 statistic is maximized when the group sizes are equal. Recall that the form of the statistic is

$$F_1(\mathbf{x}) = \frac{SA(\mathbf{x})/(k-1)}{SE(\mathbf{x})/(N-k)}.$$

where

$$SA(\mathbf{x}) = \sum_{j=1}^{k} n_j |\bar{y}_j - \bar{y}|$$

and

$$SE(\mathbf{x}) = \sum_{i=1}^{N} |y_i - \bar{y}_{c_i}|.$$

Only SA is a function of the n_j, so we restrict our attention to that term as $E(F_1)$ for different allocations will scale with $E(SA)$ by a multiplicative constant.

Denote each term in the summation SA_j. By using the approach used in Appendix A to find the distribution of $y_i - y_{c_i}$, it can be shown that $\bar{y}_j - \bar{y}$ is distributed normal mean zero, and variance $\omega_j^2 = \frac{\sigma^2}{n_j} - \frac{\sigma^2}{N}$.

$$\begin{aligned}
\mathrm{E}(SA) &= \mathrm{E}\left(\sum_{j=1}^{k} SA_j\right) \\
&= \sum_{j=1}^{k} \mathrm{E}(SA_j) \\
&= \sum_{j=1}^{k} \mathrm{E}\left(n_j |\bar{y}_j - \bar{y}|\right) \\
&= \sum_{j=1}^{k} \mathrm{E}\left(n_j |\frac{\omega_j(\bar{y}_j - \bar{y})}{\omega_j}|\right) \\
&= \sum_{j=1}^{k} n_j \omega_j \mathrm{E}\left(|\frac{(\bar{y}_j - \bar{y})}{\omega_j}|\right) \qquad \text{since } \omega_j > 0 \\
&= \sum_{j=1}^{k} n_j \omega_j (\sqrt{\frac{2}{\pi}})
\end{aligned}$$

Since each $|\frac{(\bar{y}_j - \bar{y})}{\omega_j}|$ is distributed standard half normal, the expectation of each evaluates to the constant $\sqrt{\frac{2}{\pi}}$. Now we may also factor out σ from ω_j, providing the final answer:

$$\mathrm{E}(SA) = \sigma(\sqrt{\frac{2}{\pi}}) \sum_{j=1}^{k} n_j \sqrt{\frac{1}{n_j} - \frac{1}{N}}.$$

Proposition 1. *The expectation of SA is maximized when group sizes are all equal, i.e., when $n_j = n_{j'}$, for all $j \neq j'$*

Proof. It suffices to show that the sum between any two summands of the expectation is maximized when the other $k-2$ group sizes are fixed, so we begin by fixing each n_j except for n_1 and n_2, without loss of generality. Define $c = n_1 + n_2$ to be the sum between the two unfixed group sizes, so then the sum between the first two summands becomes:

$$\begin{aligned}
&\sigma(\sqrt{\frac{2}{\pi}})(n_1\sqrt{\frac{1}{n_1} - \frac{1}{N}} + (c - n_1)\sqrt{\frac{1}{(c-n_1)} - \frac{1}{N}}) \\
&= \sigma(\sqrt{\frac{2}{\pi}})(\sqrt{n_1 - \frac{n_1^2}{N}} + \sqrt{(c-n_1) - \frac{(c-n_1)^2}{N}})
\end{aligned}$$

since n_1 and $c - n_1 > 0$. Now factoring constants and differentiating with respect to n_1:

$$\frac{1}{\sigma(\sqrt{\frac{2}{\pi}})} \frac{d}{dn_1} \mathrm{E}(SA_1 + SA_2) =$$

$$\frac{\frac{2(c-n_1)}{N} - 1}{2\sqrt{-\frac{(c-n_1)^2}{N} + c - n_1}} + \frac{1 - \frac{2n_1}{N}}{2\sqrt{n_1 - \frac{n_1^2}{N}}},$$

which has only one solution in n_1, when $n_1 = \frac{c}{2} = \frac{n_1+n_2}{2}$, which implies $n_1 = n_2$. This critical point is associated with a maximum expectation on the first two summands, as desired. Applying this result across all $n_j, n_{j'}$ pairs such that $j \neq j'$, will result in maximizing each summand of the expectation when all $n_j = n_{j'}$, hence maximizing $\mathrm{E}(SA)$. □

C Appendix: Sensitivity proofs for *SQA* and *SQE*

Recall that, from Definition 10, we have the following functions *SQA* and *SQE*:

$$SQA(\mathbf{x}) = \sum_{j=1}^{k} n_j \left|\overline{y}_j - \overline{y}\right|^q$$

$$SQE(\mathbf{x}) = \sum_{i=1}^{N} \left|y_i - \overline{y}_{c_i}\right|^q$$

Our goal here is to bound the sensitivities for *SQA* and *SQE*.

Theorem 7 (*SQE* Sensitivity)**.** *The sensitivity of SQE is bounded above by*

$$2\left(\frac{N}{2}\right)^{(1-q)} + 1$$

when $q \in (0,1)$ and

$$N - N\left(1 - \frac{2}{N}\right)^q + 1$$

when $q \geq 1$. Note that both give an upper bound of 3 when $q = 1$.

Proof. As in the previous proofs of the sensitivity of *SE* and *SA*, suppose neighboring databases $\mathbf{x}$ and $\mathbf{x}'$ differ by some row r, with $c_r = a$ in $\mathbf{x}$ and $c_r = b$ in $\mathbf{x}'$. Rewrite the *SQE* as a sum that indexes over group size and entries within each group.

$$SQE(\mathbf{x}) = \sum_{j=1}^{k} \sum_{i \in C_j} \left|y_i - \overline{y}_{c_i}\right|^q.$$

Let $t_i = \left|y_i - \overline{y}_{c_i}\right|^q$ for any entry i. Note that if $c_i \neq a, b$, then t_i will not change between databases $\mathbf{x}$ and $\mathbf{x}'$, as the group means of the other groups are not altered. Thus, unless i is in group a or b, t_i will contribute nothing to the overall sensitivity of the *SQE*. For notational ease, let $z = y_i - \overline{y}_{c_i}$ and parameterize t_i as a function of z. The maximum z can change by is $1/n_{c_i}$. We now bound the sensitivity by individually bounding the sensitivity Δt_i of each term with $c_i = a, b$.

Case 1: When q is less than 1, $t_i(z)$ is a concave function with minimum at $z = 0$ that is symmetric about the y-axis and which monotonically increases for positive z. Because the slope of $t_i(z)$ is highest near $z = 0$, the worst case sensitivity is between $z = 0$ and $z = 1/n_{c_i}$, and hence

$$\begin{aligned}\Delta t_i &\leq |t_i(0) - t_i(1/n_{c_i})| \\ &= (1/n_{c_i})^q.\end{aligned}$$

Note that when $c_i = a$ and $i \neq r$, $\Delta t_i \leq (1/n_a)^q$. The analogous statement holds for group b. By multiplying these bounds by the number of terms in each group, we get

$$\Delta SQE \leq n_a^{1-q} + n_b^{1-q} + 1,$$

where the first and second terms are the total change possible to terms in groups a and b respectively, and the final term is for the contribution of row r itself, which we cannot bound other than by noting that its value both in $\mathbf{x}$ and $\mathbf{x}'$ falls inside of $[0, 1]$.

We must now take the worst-case value of this bound over all possible database sizes n_a and n_b. Since n_a is a positive integer and $q < 1$, n_a^{1-q} increases for a given q as n_a increases (and the same is true for n_b). So, the worst-case sensitivity will occur when as much of the total database is in groups a and b as possible. Write $n_b = N - n_a$. Then

$$n_a^{1-q} + (N - n_a)^{1-q}$$

is a downward facing parabola-like function with maximum value when $n_a = N/2$. So, the sensitivity of the *SQE* is bounded above by

$$\Delta SQE \leq 2\left(\frac{N}{2}\right)^{1-q} + 1.$$

Case 2: When q is greater than 1, t_i is a convex function with maximum at $z = 1$, symmetric about the y-axis, and monotonically increasing for positive z. So, the worst case sensitivity is between $z = 1$ and $z = 1 - 1/n_{c_i}$, and hence

$$\begin{aligned}\Delta t_i &\leq |t_i(1) - t_i(1 - 1/n_{c_i})| \\ &= 1 - (1 - 1/n_{c_i})^q.\end{aligned}$$

Summing these bounds over all terms, we have

$$\Delta SQE \leq n_a(1 - (1 - 1/n_a)^q) + n_b(1 - (1 - 1/n_b)^q) + 1.$$

Again, the sensitivity will be maximized when as much of the database is distributed between n_a and n_b as possible. To determine what the worst case allocation is, let

$$f = n_a(1-(1-1/n_a)^q)+(N-n_a)(1-(1-1/(N-n_a))^q)+1,$$

i.e., f is an expression for the upper bound of ΔSQE with n_b replaced by $N - n_a$ to maximize sensitivity. Then, we can maximize f in terms of n_a:

$$\frac{\partial \Delta f}{\partial n_a} = -\frac{1-q}{(N-n_a)^q} + \frac{1-q}{n_a^q}$$

has a critical point at $n_a = N/2$, and

$$\frac{\partial \Delta^2 f}{\partial n_a^2} = -\frac{(1-q)q}{(N-n_a)^{-1-q}} - \frac{(1-q)q}{n_a^{-1-q}}$$

is always negative. So, f is concave down and $n_a = N/2$ is a global maximum. Hence, the worst case sensitivity occurs when the database is distributed equally between groups a and b, i.e.,

$$\Delta SQE \leq N\left(1-\left(1-\frac{2}{N}\right)^q\right)+1.$$

□

Theorem 8 (SQA Sensitivity)**.** *The sensitivity of SQA is bounded above by*

$$N\left(\frac{3}{N}\right)^q + 1$$

when $q \in (0,1)$ *and*

$$N - N\left(1-\frac{3}{N}\right)^q + 1$$

when $q \geq 1$*. Note that both give an upper bound of 4 when* $q = 1$*.*

Proof. Let $s_j = |\bar{y}_j - \bar{y}|^q$ for any group j. I.e., s_j is the (unweighted) term in the calculation of the SQA that corresponds to group j. Note that as the grand mean changes between databases $\mathbf{x}$ and $\mathbf{x}'$ in addition to the group means, all terms, not just those for groups a and b, will contribute to the sensitivity of the SQA. Recall that the sensitivity of the grand mean is $1/N$, while the sensitivity of the group mean for groups a and b are $1/n_a$ and $1/n_b$ respectively.

Case 1: When $q < 1$, $s_j(z)$ is a concave function with minimum at $z = 0$, symmetric about the y-axis, and monotonically increasing for positive x. So, the worst case sensitivity of Δs_j for $j \neq a, b$ is between $x = 0$ and $x = 1/N$, and the worst case sensitivity of Δs_j for $j = a, b$ is between $z = 0$ and $z = 1/n_j + 1/N$. Then, the total sensitivity of the SQA for $q < 1$ is

$$\Delta SQA \leq (N - n_a - n_b - 1)(1/N)^q + n_a(1/N + 1/n_a)^q + n_b(1/N + 1/n_b)^q + 1.$$

The addition of the 1 comes from the fact that our data point r that switches between groups contributes $|\bar{y}_a - \bar{y}|^q$ to the calculation of the SQA in database $\mathbf{x}$, and contributes $|\bar{y}_b - \bar{y}|^q$ to the calculation of the SQA in database $\mathbf{x'}$; the difference between these two terms is bounded above by 1. Note that since $q < 1$, $(1/z)^q > 1/z$. Hence, $(1/N + 1/n_a)^q > (1/N)^q$ and thus the worst-case sensitivity occurs when all of N is allocated to groups a and b. I.e.,

$$\Delta SQA \leq n_a(1/N+1/n_a)^q+(N-n_a)(1/N+1/(N-n_a))^q.$$

Then, as in the proof of the SQE's sensitivity, to determine the worst-case sensitivity in terms of N, let

$$g = n_a(1/N + 1/n_a)^q + (N - n_a)(1/N + 1/(N - n_a))^q$$

and maximize this expression in terms of n_a.

$$\frac{\partial g}{\partial n_a} = -\left(\frac{1}{N} + \frac{1}{N-n_a}\right)^q + \left(\frac{1}{N} + \frac{1}{n_a}\right)^q + \frac{q\left(\frac{1}{N} + \frac{1}{N-n_a}\right)^{q-1}}{N-n_a} - \frac{q\left(\frac{1}{N} + \frac{1}{n_a}\right)^{q-1}}{n_a}$$

This is a symmetric expression between N and $N - n_a$. So, there must be a critical point at $n_a = N/2$. Note also that

$$\frac{\partial^2 g}{\partial n_a^2} = N^2(q-1)q\left(\frac{(\frac{1}{N} + \frac{1}{N-n_a})^q}{(N-n_a)(-2N+n_a)^2} + \frac{\left(\frac{1}{N} + \frac{1}{n_a}\right)^q}{n_a(N+n_a)^2}\right) \leq 0,$$

since $N > n_a$ and $q < 1$. So, $n_a = N/2$ is a global maximum, and hence

$$\Delta SQA \leq N\left(\frac{3}{N}\right)^q.$$

Case 2: When q is greater than 1, s_j is a convex function with minimum at $z = 0$, symmetric about $z = 0$, and monotonically increasing for positive z. So, the worst case sensitivity of Δs_j for $j \neq a, b$ is between $z = 1$ and $z = 1 - 1/N$, and the worst-case sensitivity of Δs_j for $i = a, b$ is between $z = 1$ and $z = 1 - 1/N - 1/n_i$.

Then, the total sensitivity of the SPA for $q < 1$ is

$$\begin{aligned}\Delta SQA \leq (N - n_a - n_b)(1 - (1 - 1/N)^q) \\ + n_a(1 - (1 - 1/N - 1/n_a)^q) \\ + n_b(1 - (1 - 1/N - 1/n_b)^q) + 1.\end{aligned}$$

Note that

$$\begin{aligned}1 - 1/N - 1/n_b < 1 - 1/N \\ \Rightarrow (1 - 1/N - 1/n_b)^q < (1 - 1/N)^q \\ \Rightarrow 1 - (1 - 1/N - 1/n_b)^q > 1 - (1 - 1/N)^q.\end{aligned}$$

Thus, ΔSQA is maximized when N is maximally allocated to groups a and b. As in the proof for $q < 1$, this occurs when $n_a = N/2 = n_b$. Then,

$$\Delta SQA < N\left(1 - \left(1 - \frac{3}{N}\right)^q\right).$$

□

D Appendix: Direct calculation of σ

In our work, we used $\widehat{SE}$ to form an estimator for σ when calculating the null distribution. We developed another version of a differentially-private ANOVA that calculates σ directly using a portion of the privacy budget. We first define a generalized variance and prove the sensitivity bounds on this quantity.

Definition 11 (VAR_q). *Given a database* $\mathbf{x}$ *with* k *groups and* n_j *entries in the* j*-th group, the* VAR_q *calculation is defined as*

$$VAR_q = \sum_{i=1}^{N} |y_i - \bar{y}|^q$$

where q *is a positive real number.*

Theorem 9 (VAR_q-Sensitivity). *The sensitivity of* VAR_q *is bounded above by*

$$\frac{N-1}{N^q} + 1$$

when $q < 1$, *and is bounded above by*

$$(N - 1)(1 - (1 - 1/N)^q) + 1$$

when $q > 1$. *Note that these both give a bound of* $2 - 1/N$ *when* $q = 1$.

Proof. Consider databases $\mathbf{x}$ and $\mathbf{x}'$ which differ in entry r. Recall that the sensitivity of the grand mean is $1/N$. Let $t_i = |y_i - \bar{y}|^q$. When $q < 1, t_i$ is a concave function with positive range, so the worst case sensitivity is between $y_i - \bar{y} = 0$ and $y_i - \bar{y} = 1/N$. That is,

$$\begin{aligned}\Delta t_i &= \left|t_i(0) - t_i\left(\frac{1}{N}\right)\right| \\ &= \left(\frac{1}{N}\right)^q.\end{aligned}$$

Every single term can be affected by at most $\left(\frac{1}{N}\right)^q$, except the term r, which can change by 1. So,

$$\begin{aligned}\Delta VAR_q &\leq \left|(N-1)(\frac{1}{N})^q + 1\right| \\ &= \frac{N-1}{N^q} + 1.\end{aligned}$$

When $q > 1, t_i$ is convex. The worst case sensitivity is between $y_i - \bar{y} = 1$ and $y_i - \bar{y} = 1 - 1/N$ Then,

$$\Delta VAR_q \leq \left|(N-1)(1 - (1 - \frac{1}{N})^q) + 1\right|.$$

□

We again use the Laplace mechanism, and the sensitivity of a database's variance follows directly from Thm. 9:

Corollary 1. *The sensitivity of the variance of a database is bounded above by*

$$3 + 1/N^2 - 3/N.$$

Algorithm 4 divides the privacy budget into ρ_1 for the SA, ρ_2 for the SE, and ρ_3 for the VAR calculations respectively. The values of ρ are provided as input along with the database $\mathbf{x}$ and the ε value. When $\widehat{SE}$ is used as the σ estimate, we found a 70-30 split of the privacy budget between SA and SE was optimal (Figure 3). In the power analysis of Algorithm 4, we vary the proportion of the privacy budget dedicated to the VAR calculation. We fixed the proportion of the privacy budget for ρ_1 and ρ_2 to be a 70-30 split budget not used by ρ_3 (the VAR calcuation).

Unsurprisingly, allocating part of the budget to calculating the standard deviation negatively impacts the statistical power of the test (Fig. 8). However, it is surprising that this allocation does not impact the power of the test more strongly, since VAR has a larger sensitivity and requires adding noise to smaller values than the SA and SE.

Algorithm 4 Differentially private F_1-statistic with direct calculation of variance

```
Input: Database $\mathbf{x}$, $\varepsilon$ value, $\rho_1, \rho_2, \rho_3$
Compute $\widehat{SA} = SA + Z_1$ where $Z_1 \sim \mathsf{Lap}\left(\frac{4}{\varepsilon\rho_1}\right)$
Compute $\widehat{SE} = SE + Z_2$ where $Z_2 \sim \mathsf{Lap}\left(\frac{3}{\varepsilon\rho_2}\right)$
Compute $\widehat{VAR} = VAR + Z_3$ where $Z_3 \sim \mathsf{Lap}\left(\frac{3+1/N^2-3/N}{\varepsilon\rho_3}\right)$
Compute $\widehat{F_1} = \frac{\widehat{SA}/(k-1)}{\widehat{SE}/(N-k)}$
Output: $\widehat{F_1}, \widehat{SA}, \widehat{SE}, \widehat{VAR}$
```

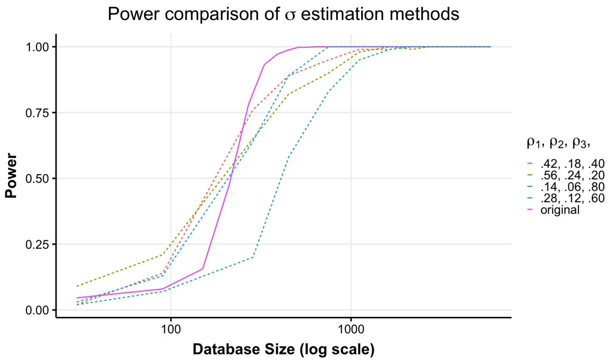

Fig. 8. The power of F_1 for a fixed ϵ value budgeted across *SA*, *SE*, and *VAR* calculations, with the portion of ϵ not allocated to ρ_3 allocated between ρ_1 and ρ_2 with a 70-30 split.

E Appendix: Parameter selection

Empirical results guiding parameter selection are presented on the following pages in Figure 9 ($\epsilon = .1$) and Figure 10 ($\epsilon = 1$).

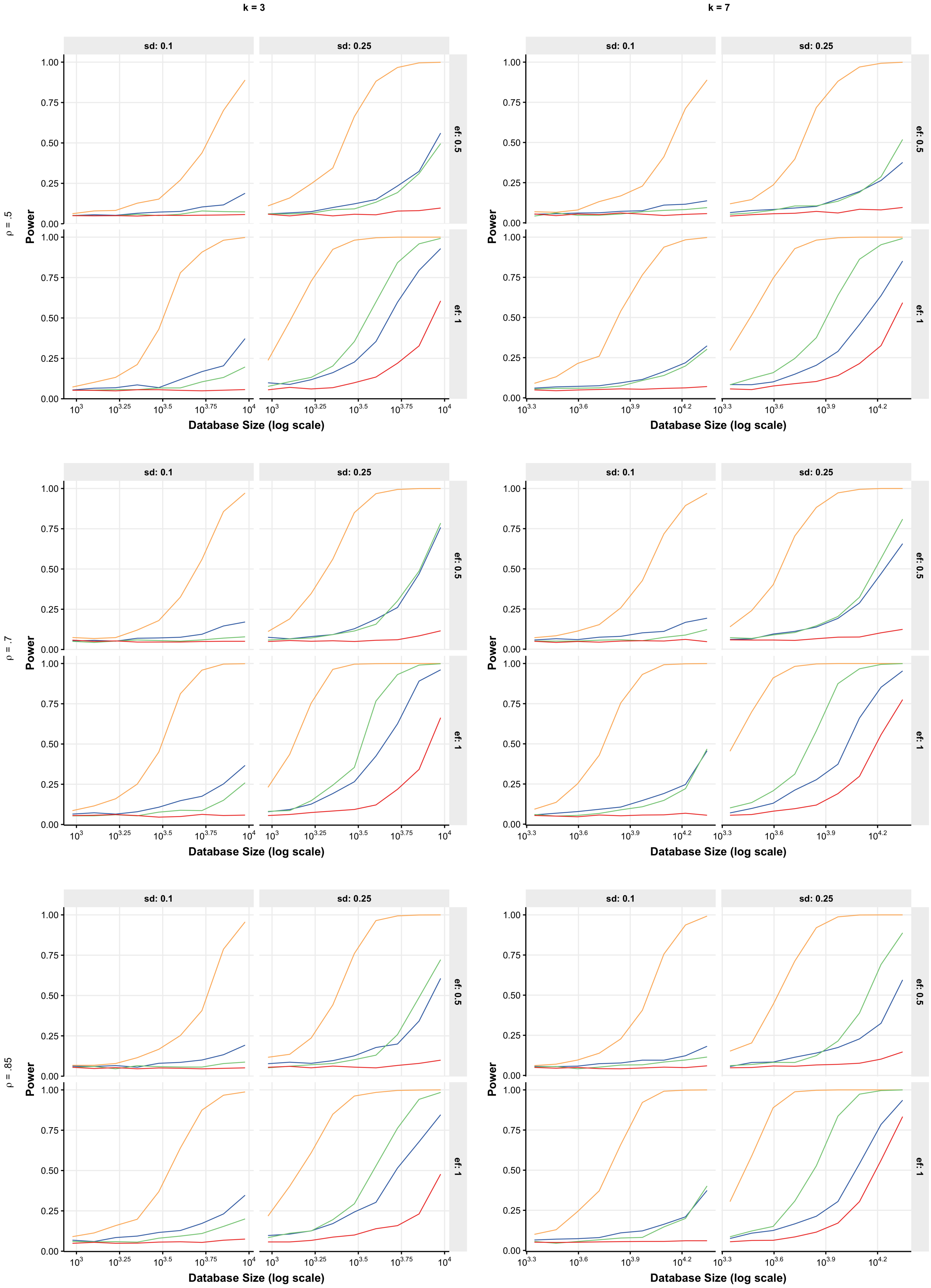

Fig. 9. Power comparison in settings where $\epsilon = .1$. Within each subplot, each color curve corresponds to a different value of q (blue: 0.75, gold: 1, green: 1.5, red: 2). The two main columns of plots correspond to different values of k and the three main rows correspond to difference values of ρ. Note that the scale on the x-axis differs with k ($k = 7$ requires more data).

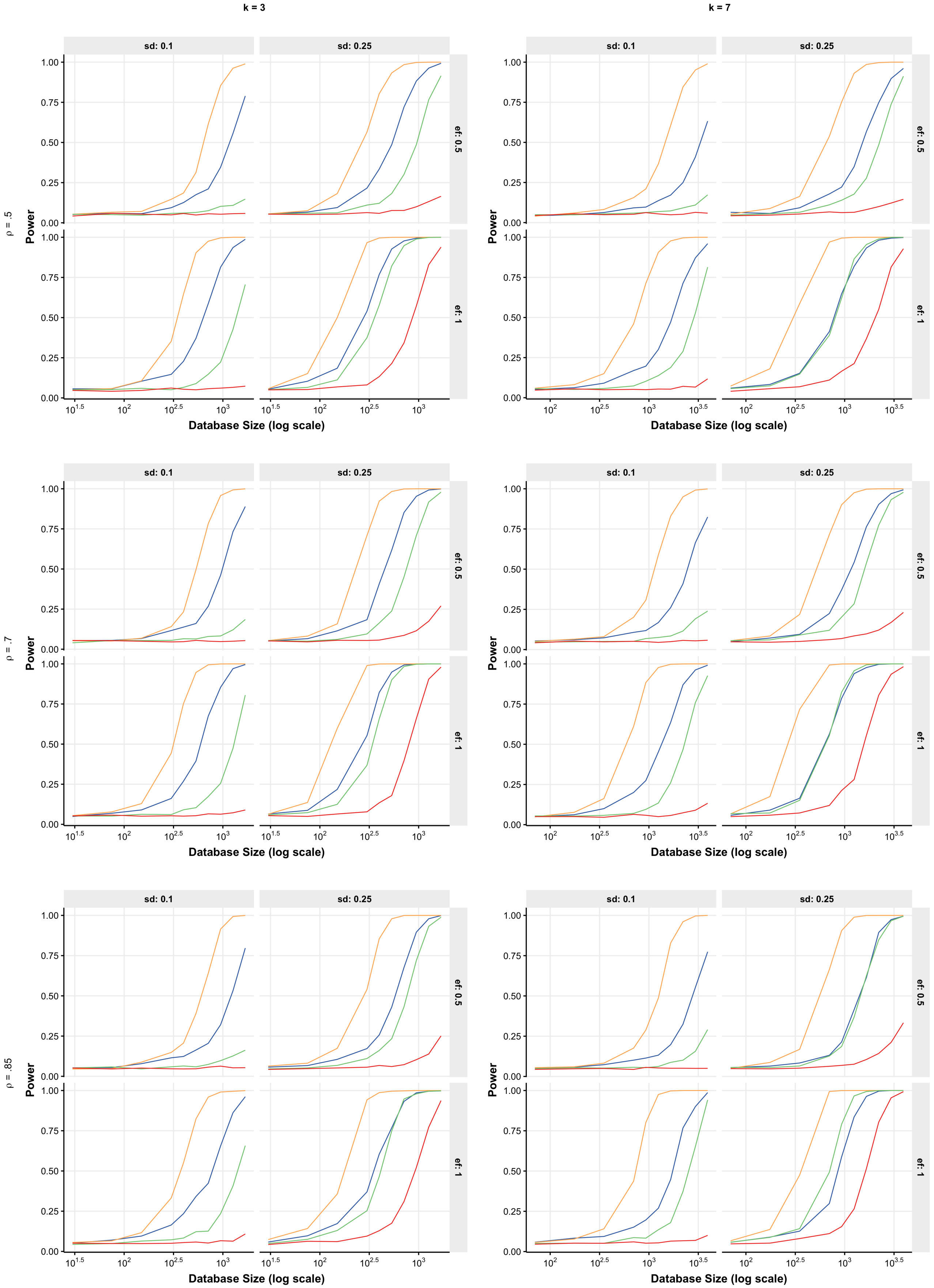

Fig. 10. Power comparison in settings where $\epsilon = 1$. Within each subplot, each color curve corresponds to a different value of q (blue: 0.75, gold: 1, green: 1.5, red: 2). The two main columns of plots correspond to different values of k and the three main rows correspond to difference values of ρ. Note that the scale on the x-axis differs with k ($k = 7$ requires more data).