Statistica Sinica Preprint No: SS-2022-0279	
Title	Differentially Private Hypothesis Testing With the Subsampled and Aggregated Randomized Response Mechanism
Manuscript ID	SS-2022-0279
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0279
Complete List of Authors	Victor Pena and Andres Barrientos
Corresponding Authors	Victor Pena
E-mails	victor.pena.pizarro@upc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Statistica Sinica

DIFFERENTIALLY PRIVATE HYPOTHESIS TESTING WITH THE SUBSAMPLED AND AGGREGATED RANDOMIZED RESPONSE MECHANISM

Víctor Peña, Andrés F. Barrientos

Universitat Politècnica de Catalunya, Florida State University

Abstract:

Randomized response is one of the oldest and most well-known methods used to analyze confidential data. However, its utility for differentially private hypothesis testing is limited because it cannot simulaneously achieve high privacy levels and low type-I error rates. We overcome this problem using the subsample and aggregate technique. The result is a general-purpose method that can be used for both frequentist and Bayesian testing. We demonstrate the performance of the proposed method in three scenarios: goodness-of-fit testing for linear regression models, nonparametric testing of a location parameter using the Wilcoxon test, and the nonparametric Kruskal–Wallis test.

Key words and phrases: Differential privacy, randomized response, hypothesis testing, Bayesian hypothesis testing.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

1. Introduction

In this paper, we propose a method for testing hypotheses based on confidential data. It is conceptually simple, widely applicable, and can simultaneously attain high privacy levels and low type-I error rates.

We work within the *differential privacy* framework (Dwork et al., 2006). From a data privacy perspective, differentially private algorithms are appealing because they are robust to deanonymization attacks (Dwork et al., 2014). From a statistical perspective, differentially private algorithms are useful because they facilitate inferences from private data.

There is a growing body of literature on differentially private hypothesis testing. For example, Gaboardi et al. (2016) and Rogers and Kifer (2017) provide differentially private chi-squared tests, Couch et al. (2019) develop differentially private versions of nonparametric tests such as the Mann–Whitney and Kruskal–Wallis tests, and Barrientos et al. (2019), Peña and Barrientos (2021), and Alabi and Vadhan (2022) propose methods for testing in linear regression models.

Our proposed method applies the subsample and aggregate technique (Nissim et al., 2007) to randomized response (Warner, 1965). The result is a general-purpose algorithm that can create differentially private versions of existing nonprivate hypothesis test. In our simulation studies and an

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

application, we find that the method is especially useful when the type-I error α of the tests is low. Testing hypotheses using low significance levels (as low as $\alpha = 0.005$) has been proposed as a way of ameliorating what has become known as the *replication crisis*, where published significant results (typically at a significance level $\alpha = 0.05$) fail to replicate in subsequent follow-up experiments (Benjamin et al., 2018).

The subsample and aggregate technique splits the data into subsets, computes the statistics for each subset, and combines the results in a way that ensures that the output is differentially private. From a theoretical perspective, Smith (2011) studies general asymptotic properties of the strategy. From an applied perspective, the technique has been used to build differentially private algorithms for clustering (Mohan et al., 2012; Su et al., 2016), feature selection with the LASSO (Thakurta and Smith, 2013), hypothesis testing for normal linear models (Barrientos et al., 2019; Peña and Barrientos, 2021), and logistic regression (Mohan et al., 2012).

Randomized response was originally motivated as a method for reducing bias in answers to sensitive questions. Since its inception more than 50 years ago, it has been extended and applied to many different contexts; see Blair et al. (2015) or the monograph Chaudhuri and Mukerjee (2020) for further details. Importantly, randomized response is differentially private (Dwork

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

et al., 2014). Its properties within the framework have been studied by Wang et al. (2016) and Ma and Wang (2021), and it is used as a building block for differentially private algorithms by Erlingsson et al. (2014), Bassily and Smith (2015), and Ye et al. (2019).

Unfortunately, randomized response by itself is not useful for differentially private hypothesis testing. As we argue in Section 2, it cannot simultaneously achieve acceptable privacy levels and type-I error rates. Fortunately, this problem can be resolved with the subsample and aggregate technique.

The output of our method is a binary decision that indicates whether or not we reject a null hypothesis. The decision can be used for both frequentist and Bayesian hypothesis testing. For the latter, one has to specify prior probabilities on the hypotheses and a prior distribution on the power of the nonprivate test used to build the private test.

Previous works on differentially private hypothesis testing focus on releasing differentially private p-values or, from a Bayesian perspective, Bayes factors. In contrast, our output is a binary decision. Although our output is, in some sense, less informative, it need not be less useful from a practical standpoint. If we perform a hypothesis test, the type-I error α must be set in advance. If we are using a p-value to make that decision, we should reject

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

the null hypothesis if it is less than α. Otherwise, we may be tempted to p-hack (Gelman and Loken, 2013) or misinterpret the p-value (Schervish, 1996). In our application and simulation studies in Section 4, we show that approaches based on binarized outcomes are often more powerful than approaches based on p-values. This makes intuitive sense, because a bit (a decision to reject or not reject a null hypothesis) is less informative than a p-value.

In Section 2, we define differential privacy and randomized response. In Section 3, we define the subsampled and aggregated randomized response mechanism, study its properties, and devise simple strategies to implement it in practice. In Section 4, we demonstrate the performance of the method in differentially private implementations of the goodness-of-fit tests proposed in Peña and Slate (2006), the one-sample Wilcoxon test, and the Kruskal–Wallis test. Section 5 closes with a brief discussion and ideas for future work. All proofs are relegated to the Supplementary Material, which also includes an additional simulation study comparing our method with the differentially private test for regression coefficients proposed in Barrientos et al. (2019).

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

2. Preliminaries

In this section, we give a brief introduction to differential privacy and randomized response. We state a simplified version of the general definition of differential privacy that is sufficient for our purposes.

Before we define differential privacy, we first need to define neighboring data sets.

Definition 1. Let $D = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ and $D' = (x'_1, x'_2, \ldots, x'_n) \in \{0,1\}^n$. Then, D and D' are neighbors if they differ in only one component: $x_i = x'_i$ for all $i \in \{1, 2, \ldots, n\}$, except for one $j \in \{1, 2, \ldots, n\}$ for which $x_j \neq x'_j$.

Differential privacy bounds the extent to which the output of randomized algorithms can vary for neighboring data sets. In the differential privacy literature, privacy-ensuring randomized algorithms are referred to as *mechanisms*. We formally define differential privacy below.

Definition 2. A mechanism $M : \{0,1\}^n \to \{0,1\}$ is ε-differentially private if there exists $\varepsilon > 0$ such that, for all neighboring $D, D' \in \{0,1\}^n$,

$$\max\left\{\frac{\mathbb{P}[M(D)=1]}{\mathbb{P}[M(D')=1]}, \frac{\mathbb{P}[M(D')=1]}{\mathbb{P}[M(D)=1]}, \frac{\mathbb{P}[M(D)=0]}{\mathbb{P}[M(D')=0]}, \frac{\mathbb{P}[M(D')=0]}{\mathbb{P}[M(D)=0]}\right\} \leq e^{\varepsilon}.$$

The mechanism is exactly ε-differentially private if the upper bound is tight.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

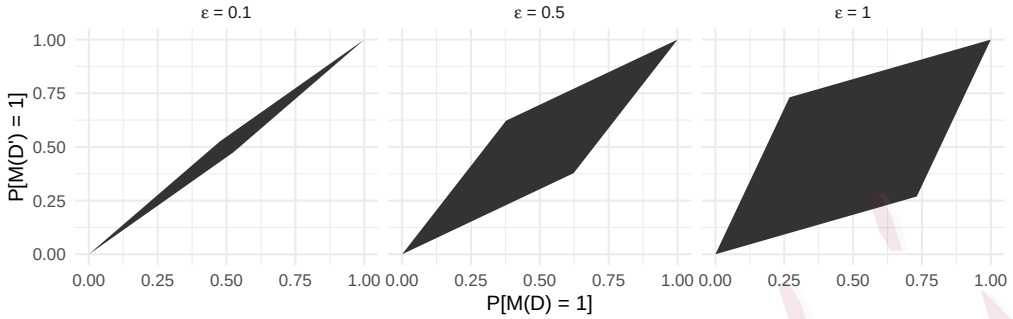


Figure 1: Shaded regions show the choices of $\mathbb{P}[M(D) = 1]$ and $\mathbb{P}[M(D') = 1]$ that achieve ε-differential privacy for $\varepsilon \in \{0.1, 0.5, 1\}$.

Low values of ε are associated with high privacy levels, whereas high values of ε are associated with low privacy. Figure 1 illustrates how ε restricts $\mathbb{P}[M(D) = 1]$ and $\mathbb{P}[M(D') = 1]$ for $\varepsilon \in \{0.1, 0.5, 1\}$. As ε goes to zero, $\mathbb{P}[M(D) = 1]$ and $\mathbb{P}[M(D') = 1]$ are forced to be equal; as ε goes to infinity, any values of $\mathbb{P}[M(D) = 1]$ and $\mathbb{P}[M(D') = 1]$ satisfy ε-differential privacy.

A key building block of our method is the randomized response mechanism. It takes a binary input $x \in \{0, 1\}$ and outputs

$$r(x) = \begin{cases} x, \text{ with probability } p, \\ 1 - x, \text{ with probability } 1 - p, \qquad 1/2 < p < 1. \end{cases}$$

We consider that x is the outcome of a hypothesis test, and $x = 1$ implies rejection of a null hypothesis H_0.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

The proposition below is well known in the differential privacy literature (e.g., example, Dwork et al. (2014)), and states that $r(x)$ is ε-differentially private.

Proposition 1. *Let* $\varepsilon > 0$ *and* $p = \exp(\varepsilon)/[1 + \exp(\varepsilon)]$. *Then,* $r(x)$ *is exactly* ε*-differentially private.*

We would like $r(x)$ to be ε-differentially private and have a type-I error rate of at most α. Unfortunately, $r(x)$ cannot achieve low values of ε and α simultaneously. If x is conducted at significance level $0 < \alpha_0 < 1$, the type-I error of $r(x)$ is $p\alpha_0 + (1-p)(1-\alpha_0) \geq 1 - p$, which is very limiting; for example, if $\varepsilon = 1$, the type-I error of $r(x)$ is at least 0.268.

We can control the type-I error of $r(x)$ by randomizing it further. That is, we can report $Br(x)$ for $B \sim \text{Bernoulli}(\varrho)$, where ϱ is set so that $Br(x)$ has type-I error α. However, the introduction of B causes a substantial loss in power. In particular, the power of $Br(x)$ is bounded above by ϱ: for instance, if $\varepsilon = 1$, $\alpha_0 = 0.05$, and ϱ is such that the type-I error of $r(x)$ is $\alpha = 0.05$, the power of $Br(x)$ is bounded above by $\varrho \approx 0.17$. In Sections 3 and 4, we show that subsampling and aggregating provides a more powerful solution.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

3. Subsampled and aggregated randomized response

3.1 General properties

In this section, we define the subsampled and aggregated randomized response mechanism and study its properties.

First, we split the data uniformly at random into $2k+1$ disjoint subsets, indexed by $i \in \{1, 2, \ldots, 2k+1\}$, where k is a nonnegative integer. Within the subsets, we run the nonprivate test of interest at significance level α_0. The outcomes of the tests are denoted as x_i, where $x_i = 1$ indicates rejection of the null hypothesis H_0 in the ith subset. Then, we apply independent randomized response mechanisms to x_i, obtaining $r(x_i)$. Finally, we combine the results in $T = \sum_{i=1}^{2k+1} r(x_i)$, and report $d_c = \mathbb{1}(T > c)$.

The proposition below shows that the privacy level ε of d_c has a closed-form expression. We derived it using facts about stochastically ordered random variables found in Shaked and Shanthikumar (2007). We use the notation Binomial(i, p) + Binomial(j, q) for the distribution of the sum of independent Binomial(i, p) and Binomial(j, q) random variables, with the understanding that if the number of trials is zero, the random variable is zero with probability one.

Proposition 2. *The statistic $d_c = \mathbb{1}(T > c)$ is exactly ε-differentially*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

private, with

$$\varepsilon = \log\left(\frac{\mathbb{P}(B_1 > c_*)}{\mathbb{P}(B_0 > c_*)}\right),$$

where $c_* = \max(c, 2k - c)$ *and* $B_i \sim \text{Binomial}(i, p) + \text{Binomial}(2k + 1 - i, 1 - p)$, *for* $i \in \{0, 1\}$.

Proposition 2 shows that ε depends on c, k, and p. We study how these parameters affect ε by fixing two of them at a time and letting the other one vary.

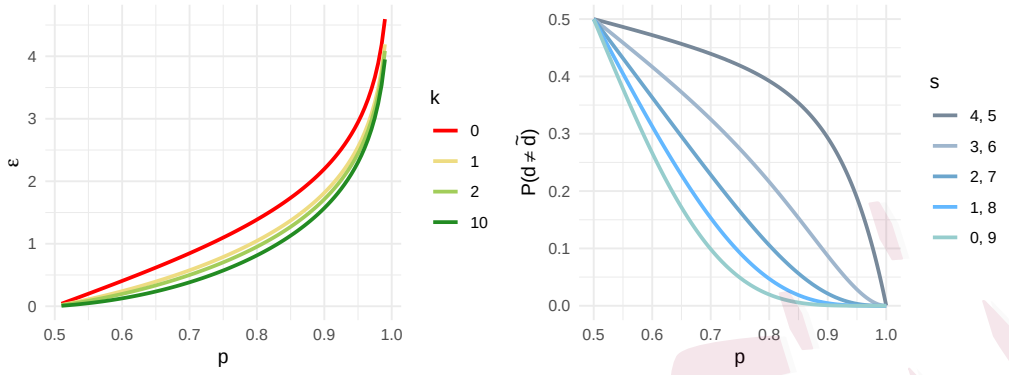
Proposition 3. *The statistic* $d_c = \mathbb{1}(T > c)$ *has the following properties:*

1. *For any fixed* k *and* c, ε *is increasing in* p.
2. *For any fixed* p *and* $c \geq k$, ε *is decreasing in* k.
3. *For any fixed* k *and* p, ε *is minimized at* $c = k$.

Proposition 3 establishes that subsampling and aggregating lowers ε whenever $c \geq k$. It also suggests the majority vote $d = \mathbb{1}(T > k)$ as a default choice of d_c because this minimizes ε for any fixed k and p. For this reason and its intuitive appeal, we restrict our attention to d from this point onward.

Figure 2a shows ε given k and p. Splitting the data into additional subsets reduces ε, but the gains are limited as k increases. The proposition

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279



(a) (b)

Figure 2: (a) Privacy parameter ε as a function of p and k. (b) $\mathbb{P}(d \neq \tilde{d})$ as a function of p and $T = s$ for $k = 4$.

below confirms this intuition: the limit of ε as k goes to infinity is a positive constant that is increasing in p. Combined with Proposition 3, the limit of ε as k goes to infinity establishes a nontrivial necessary condition on p for achieving ε-differential privacy.

Proposition 4. *The statistic $d = \mathbb{1}(T > k)$ has the following properties:*

1. *For any fixed p,*

$$\lim_{k\to\infty} \varepsilon = \log\left(1 + \frac{(2p-1)^2}{2p(1-p)}\right) > 0.$$

2. *A necessary condition on p for achieving ε-differential privacy is*

$$p \leq \frac{1}{2}\left(1 + \frac{\sqrt{\exp(2\varepsilon) - 1}}{1 + \exp(\varepsilon)}\right).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

A sufficient condition on p for achieving ε-differential privacy is

$$p \leq \frac{\exp(\varepsilon)}{1+\exp(\varepsilon)}.$$

There are two sources of uncertainty in d: the uncertainty in x_i and the uncertainty introduced by the randomized response mechanisms. We focus on the latter now, comparing $d = \mathbb{1}(T > k)$ with $\tilde{d} = \mathbb{1}(\sum_{i=1}^{2k+1} x_i > k)$, treating $\sum_{i=1}^{2k+1} x_i$ as fixed.

Given $\sum_{i=1}^{2k+1} x_i = s$, the probability that d is not equal to $\tilde{d}$ is

$$\mathbb{P}(d \neq \tilde{d} \mid \textstyle\sum_{i=1}^{2k+1} x_i = s) = \begin{cases} \mathbb{P}(B_s > k), \text{ if } s \leq k, \\ \mathbb{P}(B_s \leq k), \text{ if } s > k, \end{cases}$$

where $B_s \sim \text{Binomial}(s, p) + \text{Binomial}(2k+1-s, 1-p)$. Figure 2b displays this probability as a function of p and s for $k = 4$. There is an interesting symmetry in s that holds in general.

Proposition 5. *For any given $s \in \{0, 1, \ldots, k\}$,*

$$\mathbb{P}(d \neq \tilde{d} \mid \textstyle\sum_{i=1}^{2k+1} x_i = s) = \mathbb{P}(d \neq \tilde{d} \mid \textstyle\sum_{i=1}^{2k+1} x_i = 2k+1-s).$$

The probability that d and $\tilde{d}$ disagree depends on s, k, and p. Below, we describe this dependence.

Proposition 6. *The probability $\mathbb{P}(d \neq \tilde{d} \mid \sum_{i=1}^{2k+1} x_i = s)$ has the following properties:*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

1. *For any fixed k and p, $\mathbb{P}(d \neq \tilde{d} \mid \sum_{i=1}^{2k+1} x_i = s)$ is decreasing in s if $s > k$, and increasing in s if $s \leq k$.*

2. *For any fixed p and s, $\mathbb{P}(d \neq \tilde{d} \mid \sum_{i=1}^{2k+1} x_i = s)$ is decreasing in k if $s \leq k$, and increasing in k if $s > k$.*

3. *For any fixed k and s, $\mathbb{P}(d \neq \tilde{d} \mid \sum_{i=1}^{2k+1} x_i = s) = 1/2$ if $p = 1/2$, and $\mathbb{P}(d \neq \tilde{d} \mid \sum_{i=1}^{2k+1} x_i = s) = 0$ if $p = 1$.*

A direct consequence of Propositions 5 and 6 is that the probability that d and $\tilde{d}$ disagree is minimized when $\sum_{i=1}^{2k+1} x_i \in \{0, 2k+1\}$ and maximized when $\sum_{i=1}^{2k+1} x_i \in \{k, k+1\}$.

3.2 Hypothesis testing

In this section, we focus on properties related to hypothesis testing. For simplicity, we assume that the subsets are balanced (i.e., they have the same sample size), but the proposed method can still be used, provided the subsets are not heavily unbalanced. Throughout, we assume that the tests behind x_i are all conducted at a fixed significance level α_0.

The type-I error and power of d depend on the probability that the nonprivate test x_i rejects H_0, which we denote as γ_0. Under H_0, γ_0 is the type-I error $\alpha_0 = \mathbb{P}(x_i = 1 \mid H_0)$; under H_1, it is the power $\mathbb{P}(x_i = 1 \mid H_1)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Let $T = \sum_{i=1}^{2k+1} r(x_i) \sim \text{Binomial}(2k+1, p\gamma_0 + (1-p)(1-\gamma_0))$ be the number of subsets in which H_0 is rejected. Because $d = \mathbb{1}(T > k)$, the distribution of T lets us quantify how the probability of rejecting H_0 depends on k, p, and γ_0.

Proposition 7. *The probability that d rejects H_0 has the following properties:*

1. *For any fixed k and p, the probability that d rejects H_0 is increasing in γ_0.*

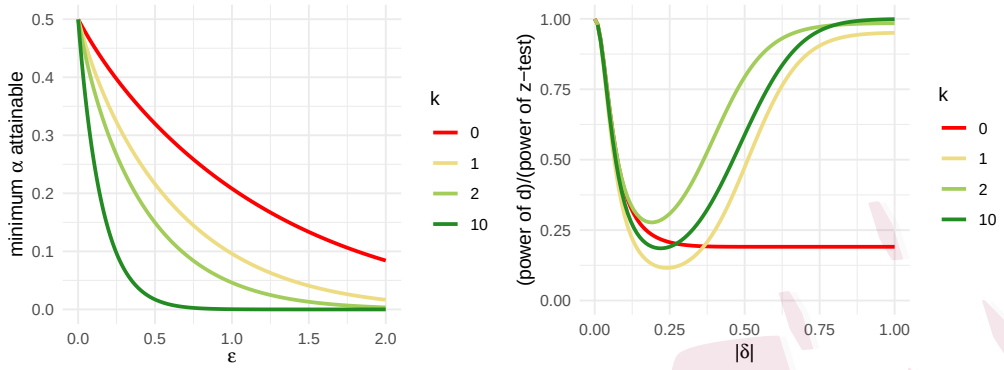
2. *For any fixed γ_0 and k, the probability that d rejects H_0 is decreasing in p if $\gamma_0 < 1/2$, and increasing in p if $\gamma_0 > 1/2$.*

3. *Let $p > 1/2$ be fixed. If $\gamma_0 > 1/2$, then the probability that d rejects H_0 goes to one as $k \to \infty$. Alternatively, if $\gamma_0 < 1/2$, then the probability that d rejects H_0 goes to zero as $k \to \infty$.*

Part 3 of Proposition 7 establishes that d is consistent under H_1, as long as the power of the tests within the subsets is greater than $1/2$ as k goes to infinity.

The type-I error α of d depends on k, p, and α_0. By Proposition 2, ε does not depend on α_0, so decreasing α_0 decreases α without sacrificing ε. In Proposition 4, we saw that ε is decreasing in k, but that the gains are

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279



(a) (b)

Figure 3: (a) Minimum α attainable as a function of ε and k. (b) Difference in power between d and the z-test, as a function of the effect size $|\delta|$ and k, for $n = 105$.

limited. This is not the case for α, because the minimum α attainable as k grows to infinity is zero. However, recall that reducing α_0 decreases the power of the test.

Proposition 8. *For any $\varepsilon > 0$, the minimum type-I error α attainable by d goes to zero as k goes to infinity.*

3.3 Tuning parameters of the mechanism

We propose two strategies for choosing k, p, and α_0. In both cases, we set p given k, so that d is exactly ε-differentially private. Once k and p are fixed, we find α_0 such that d has type-I error α.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

The first strategy is to inspect the power curves for values of k on a grid. For each k, we determine if there are p and α_0 such that d is ε-differentially private and has type-I error α. If such p and α_0 exist, we can find a power curve. In some cases, power curves have closed-form expressions; in others, they can be simulated. After finding power curves for all values of k on the grid, we can choose a value visually.

The second strategy is a heuristic that can be used when the tests have low power, in which case increasing the number of subsets typically decreases the power of d. With that in mind, we propose setting k to the minimum value k^* for which there exist p and α_0 such that d is ε-differentially private and has type-I error α. To avoid α_0 that are too small, we add the restriction $\alpha_0 \geq \alpha_{0,\min}$ for a user-defined $\alpha_{0,\min} > 0$.

Table 1: Minimum k needed for different combinations of α and ε.

min k	$\varepsilon = 0.5$	$\varepsilon = 0.75$	$\varepsilon = 1$	$\varepsilon = 1.25$	$\varepsilon = 1.5$
$\alpha = 0.005$	13	8	6	4	3
$\alpha = 0.01$	11	7	5	4	3
$\alpha = 0.05$	6	4	3	2	1
$\alpha = 0.1$	4	2	2	1	1

Proposition 8 guarantees that, for sufficiently large k, we can achieve

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

any privacy level ε and type-I error α simultaneously. However, if we want both ε and α to be small and we have a small sample size, we may not be able to split the data into enough subsets to satisfy both requirements. Table 1 shows the minimum k needed to simultaneously achieve a type-I error level α and ε-differential privacy for $\alpha \in \{0.005, 0.01, 0.05, 0.1\}$ and $\varepsilon \in \{0.5, 0.75, 1, 1.25, 1.5\}$. The lower α and ε are, the larger k needs to be to simultaneously achieve α type-I error level and ε-differential privacy.

Below, we apply the two strategies to the one sample z-test. The example shows that we need to set $\alpha_{0,\min} > 0$: without a nonzero minimum, $\alpha_{0,\min}$ can be too small and the mechanism can be underpowered.

Example 1 (One-sample z-test)**.** Let the data be 105 independent and identically distributed observations distributed as Normal(μ, σ^2), with σ^2 known. We set $\varepsilon = 1.5$ and test $H_0 : \mu = \mu_0$ against $H_1 \neq \mu_0$ with the one-sample z-test at significance level $\alpha = 0.05$. In this example, the classical randomized response mechanism ($k = 0$) cannot achieve ε-differential privacy and a type-I error α at the same time. To solve this problem, we let $p = \exp(\varepsilon)/[1 + \exp(\varepsilon)]$ and define $Br(x)$, where $B \sim$ Bernoulli(ϱ). The probability ϱ is set so that $Br(x)$ has a type-I error α. Figure 3b shows the ratio of the power of d to the power of the usual nonprivate z-test for $k \in \{0, 1, 2, 10\}$ and effect sizes $|\delta| = |\mu - \mu_0|/\sigma$ ranging from zero to

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

one. The classical randomized response mechanism ($k = 0$) and $k = 1$ do not perform well. In the latter case, the method has low power because $\alpha_0 \approx 0.0025$. The performance of $k = 2$ and $k = 10$ is similar for small effect sizes. For moderate effect sizes, $k = 2$ is preferable. For large effect sizes, $k = 10$ outperforms $k = 2$, but at that point both approaches are essentially as powerful as the nonprivate z-test. The values of α_0 are 0.089 for $k = 2$ and 0.281 for $k = 10$. If we use the automatic strategy to select the parameters of the mechanism with $\alpha_{0,\min} = 0$, it selects $k^* = 1$; if we set a minimum $\alpha_{0,\min} > 0.0025$, it chooses $k^* = 2$ instead.

3.4 Extensions: Multiple hypotheses and Bayesian testing

The subsampled and aggregated randomized response mechanism can be used to test multiple hypotheses. Indeed, it is straightforward to apply a Bonferroni correction to multiple independent runs of the mechanism. If each test is ε-differentially private, the vector $(d_1, d_2, \ldots, d_m)$ is $m\varepsilon$-differentially private by the sequential composition property of differential privacy (McSherry, 2009). Therefore, to test m null hypotheses $H_0^1, H_0^2, \ldots, H_0^m$ at a familywise error rate α, we can run m independent subsampled and aggregated randomized response mechanisms at significance level α/m. We pursue this idea in Section 4.1.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

The binary decision d can also be used for Bayesian hypothesis testing. If d is calibrated to have type-I error α, then the posterior probability of H_1 given that $d = 0$ is

$$\begin{aligned}\mathbb{P}(H_1 \mid d = 0) &= \frac{\mathbb{P}(H_1)\mathbb{P}(d = 0 \mid H_1)}{\mathbb{P}(H_0)\mathbb{P}(d = 0 \mid H_0) + \mathbb{P}(H_1)\mathbb{P}(d = 0 \mid H_1)} \\ &= \frac{\mathbb{P}(H_1)\mathbb{P}(d = 0 \mid H_1)}{\mathbb{P}(H_0)(1 - \alpha) + \mathbb{P}(H_1)\mathbb{P}(d = 0 \mid H_1)}.\end{aligned}$$

If $\mathbb{P}(d = 0 \mid H_1)$ goes to zero as the sample size increases (i.e., d is consistent under H_1), then $\mathbb{P}(H_1 \mid d = 0)$ goes to zero as the sample size increases for all α and $\mathbb{P}(H_0)$. Therefore, the Bayesian test can give decisive evidence in favor of H_0 asymptotically.

Analogously, the posterior probability of H_1 given that $d = 1$ is

$$\mathbb{P}(H_1 \mid d = 1) = \frac{\mathbb{P}(H_1)\mathbb{P}(d = 1 \mid H_1)}{\mathbb{P}(H_0)\alpha + \mathbb{P}(H_1)\mathbb{P}(d = 1 \mid H_1)}.$$

If d is consistent under H_1, $\mathbb{P}(H_1 \mid d = 1)$ converges to $\mathbb{P}(H_1)/[\mathbb{P}(H_0)\alpha + \mathbb{P}(H_1)]$. If the null and alternative hypotheses are equally likely *a priori*, the limit simplifies to $1/(\alpha+1)$, which is greater than $1-\alpha$. In this case, the Bayesian test cannot give decisive evidence in favor of H_1 asymptotically, but it can give fairly strong evidence in its favor.

For finite sample sizes, we can evaluate $\mathbb{P}(H_1 \mid d)$ given $\mathbb{P}(H_1)$, d, and a prior distribution on the power $\pi(\gamma_0 \mid H_1) = \mathbb{P}(x_i = 1 \mid H_1)$. Once this

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

prior is set,

$$\mathbb{P}(d = 1 \mid H_1) = \int_0^1 \mathbb{P}(T > k \mid \gamma_0, H_1)\, \pi(\gamma_0 \mid H_1)\, \mathrm{d}\gamma_0,$$

where $T \mid \gamma_0, H_1 \sim \text{Binomial}(2k+1, p\gamma_0 + (1-p)(1-\gamma_0))$.

In the absence of strong prior information about $\gamma_0 \mid H_1$, we recommend running a sensitivity analysis with different priors. When the power can be expressed as a function of an effect size δ, we can use it to induce a prior distribution. We illustrate this point using the z-test.

Example 2 (Bayesian one-sample z-test)**.** Let the data be independent and identically distributed as Normal(μ, σ^2). We test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ using a Bayesian test. We set $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$ and put a unit information prior Normal(μ_0, σ^2) on $\mu \mid H_1$, which induces $\delta \mid H_1 \sim$ Normal$(0, 1)$ on the effect size $\delta = (\mu - \mu_0)/\sigma$. The prior on $\delta \mid H_1$, in turn, induces a prior $\pi(\gamma_0 \mid H_1) = \mathbb{P}(x_i = 1 \mid H_1)$. The unit information prior is a common default choice for this problem, and contains roughly as much information as one observation in the sample (Kass and Wasserman, 1995). We set $\varepsilon = 1.5$ and consider $k \in \{1, 2, 10\}$ and subgroup sample sizes $b \in \{2, 3, \ldots, 50\}$. For any given k and b, the total sample size is $n = (2k+1)b$. The results are shown in Figure 4. As b increases, the posterior probability is more decisive against or in favor of H_1, depending on whether $d = 0$ or $d = 1$, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

A peculiarity of our Bayesian analysis is that d has a fixed type-I error rate α. From a strictly Bayesian perspective, one could output a binary decision d that does not have a fixed type-I error rate. However, we appreciate that d can be interpreted by both frequentists and Bayesians, which is in line with ongoing efforts to reconcile frequentist and Bayesian answers (e.g., Bayarri and Berger (2004) and Bayarri et al. (2016)).

The Bayesian approach described here is based on conditioning on a binary outcome. Other proposals in the differential privacy literature approach the problem differently.

For example, an alternative approach is to condition on perturbed sufficient statistics rather than binary outcomes, as proposed in Amitai and Reiter (2018) and Peña and Barrientos (2021).

Another option is to draw directly from posterior distributions in a way that ensures differential privacy (e.g., Dimitrakakis et al. (2017), Heikkilä et al. (2019), Geumlek et al. (2017), and Hu et al. (2022)). However, these strategies often assume an upper bound on the likelihood, which may require users to modify their models to meet the assumption. A major drawback of these methods is that they typically require a privacy budget proportional to the number of posterior samples desired. This, in turn, may lead to unreliable Monte Carlo approximations if ε is small. In spite of the potential

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

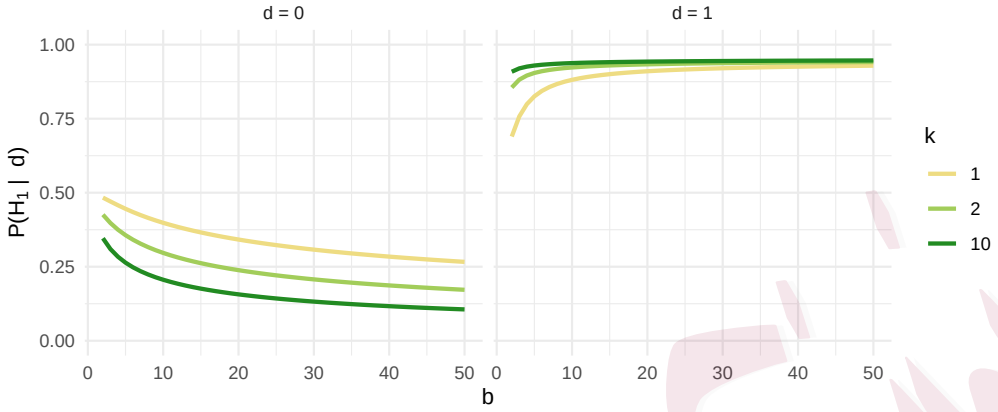


Figure 4: Posterior probability $\mathbb{P}(H_1 \mid d)$ as a function of k and the sample size within the subgroups b if $\mathbb{P}(H_0) = 1/2$ and the effect size is $\delta \sim \text{Normal}(0, 1)$.

drawbacks, applying these approaches to hypothesis testing is worthy of further exploration and development.

Another promising approach is to use the data augmentation Markov chain Monte Carlo scheme proposed in Ju et al. (2022), which has been applied to estimation problems, but not to hypothesis testing.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

4. Simulation Studies and Applications

In this section, we evaluate the performance of our methods in simulation studies and applications. In Section 4.1, we use the housing data set in Lei et al. (2018) to implement differentially private versions of the goodness-of-fit tests developed in Peña and Slate (2006). In Section 4.2, we present a simulation study that applies the Bayesian extension devised in Section 3.4 to the Wilcoxon test. Finally, in Section 4.3, we compare our general-purpose method with the differentially private Kruskal–Walis test proposed in Couch et al. (2019). In the Supplementary Material, we provide an additional simulation study, in which we compare our method to the differentially private t-test proposed in Barrientos et al. (2019).

In our this section, we include the subsampled and aggregated Laplace mechanism (Nissim et al., 2007; Smith, 2011) as a competitor. For this method, we split the data into $2k+1$ subsets and run the corresponding nonprivate tests within them. The output is made differentially private after adding a Laplace perturbation term. We consider two variants of this approach.

In the first, we find $2k+1$ p-values, one for each hypothesis test, and find the average p-value. The result is made differentially private after adding a perturbation term $\eta \sim \text{Laplace}(0, 1/[\varepsilon(2k+1)])$ to the average p-value. The

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

distribution of the differentially private statistic can be simulated under H_0, so it is straightforward to find a critical value that ensures a fixed type-I error rate α.

The second approach is based on the sum of the binary outcomes $\sum_{i=1}^{2k+1} x_i$, rather than on the average p-value. Here, the output is made differentially private after adding a perturbation term $\eta \sim \text{Laplace}(0, 1/\varepsilon)$ to the sum. We can easily find a critical value that ensures a type-I error rate α by simulating the distribution of the statistic under H_0.

4.1 Goodness-of-fit tests for regression

In this section, we study the performance of the subsampled and aggregated randomized response mechanism in a differentially private implementation of four goodness-of-fit tests for regression proposed in Peña and Slate (2006).

We perform a simulation study based on the housing data set used in Lei et al. (2018). The data set contains information on houses sold in the San Francisco Bay area between 2003 and 2006. In our analysis, we consider houses with prices within the \$105000–905000 range and sizes smaller than 3000 ft^2. After preprocessing, the data set contains 235760 rows and the following variables: price (used as the response Y), base square footage, time of transaction, lot square footage, latitude, longitude, age, number of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

bedrooms, and a binary variable indicating whether the house is located in a small county.

Peña and Slate (2006) develop tests for checking model assumptions in the normal linear model. They provide a global test of goodness-of-fit and individual tests for detecting specific violations of assumptions. Here, we consider four tests: (1) a test whose null hypothesis is that the kurtosis of the errors is equal to three, which is satisfied when the errors are normal, (2) a test in which null is that the errors are symmetric, (3) a test in which null is that the errors are homoscedastic, and (4) and a test in which the null is that the expected value of the response is linear in the predictors. For each of our simulations, we perform the four tests at significance level $\alpha/4$ and privacy level $\varepsilon/4$. This ensures that our answers have a familywise error rate α and a global privacy level ε.

To simulate data, we first fit a normal linear model using price as the response Y and the remaining variables as predictors X, obtaining maximum likelihood estimates of the regression coefficients $\widehat{\beta}$ and the residual standard deviation $\widehat{\sigma}$. Then, we use $\widehat{\beta}$ and $\widehat{\sigma}$, along with the observed X, to simulate new values of the response. More precisely, we simulate data from the model $Y^* = X\widehat{\beta} + \widehat{\sigma}W^*$, where W^* is a vector with independent and identically distributed skew-normal components with location and scale

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

parameters equal to zero and one, respectively, and skew parameter equal to θ. If $\theta = 0$, all the assumptions of the normal linear model hold, but if $\theta \neq 0$, the null hypotheses related to the kurtosis and skewness of the errors are false.

We set the significance level to $\alpha \in \{0.005, 0.01, 0.05, 0.1\}$ and consider privacy parameters $\varepsilon \in \{0.5, 0.75, 1, 1.25, 1.5\}$. The skew parameter θ ranges from 0 to 1.5. For each value of α, ε, and θ, we perform 10^4 simulations. The number of subgroups $2k + 1$ is determined using the automatic strategy outlined in Section 3.3 with $\alpha_{0,\min} = \alpha$.

The results for the test of skewness are shown in Figure 5. The results for the test of kurtosis are provided in the Supplementary Material, and are similar to those for skewness. The results for the tests of linearity and homoscedasticity are uninteresting: because the null hypothesis holds in these cases, the probability of rejecting the null is fixed to $\alpha/4$ for all ε and θ. In Figure 5, we also include the "truth", defined as the result of running the nonprivate test without splitting the data or running any mechanisms. The subsampled and aggregated sum and randomized response (labeled as SARR in the figure) outperform the average p-value in most scenarios. The average p-value is best for small α and ε, especially for low θ. The performance of the sum and randomized response, which are both based on

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

binarized outcomes, is quite similar. Randomized response performs best when $\varepsilon \in \{1, 1.25, 1.5\}$ and $\alpha = 0.005$.

4.2 Bayesian answers from one-sample Wilcoxon test

In this section, we report the results of a simulation study that compares the posterior probabilities of hypotheses based on one-sample Wilcoxon tests, using the approach proposed in Section 3.4.

We test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ for a location parameter $\theta \in \mathbb{R}$. The tests behind x_i are one-sample Wilcoxon tests, and the parameters of the mechanism are tuned so that d has type-I error rate $\alpha = 0.05$.

We repeatedly simulate data sets of sample size $n = 200$ comprising independent and identically distributed observations $y_i = \theta + \tau_i$ for $\theta \in \{0, 0.25, \ldots, 2\}$, where τ_i has a Student-t distribution with 1.5 degrees of freedom. We consider $\alpha \in \{0.005, 0.01, 0.05, 0.1\}$ and $\varepsilon \in \{0.5, 0.75, 1, 1.25, 1.5\}$, and run 10^4 simulations for each combination of θ, α, and ε. We select k using the automatic strategy described in Section 3.3 with $\alpha_{0,\min} = \alpha$.

The prior probabilities on the hypotheses are $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$. As mentioned in Section 3.4, we need a prior on $\mathbb{P}(d = 1 \mid H_1)$ to perform a Bayesian test based on d. Table 2 lists the definitions of $\mathbb{P}(d = 1 \mid H_1)$ for the methods included in the simulation study. As in Example 2, we induce

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

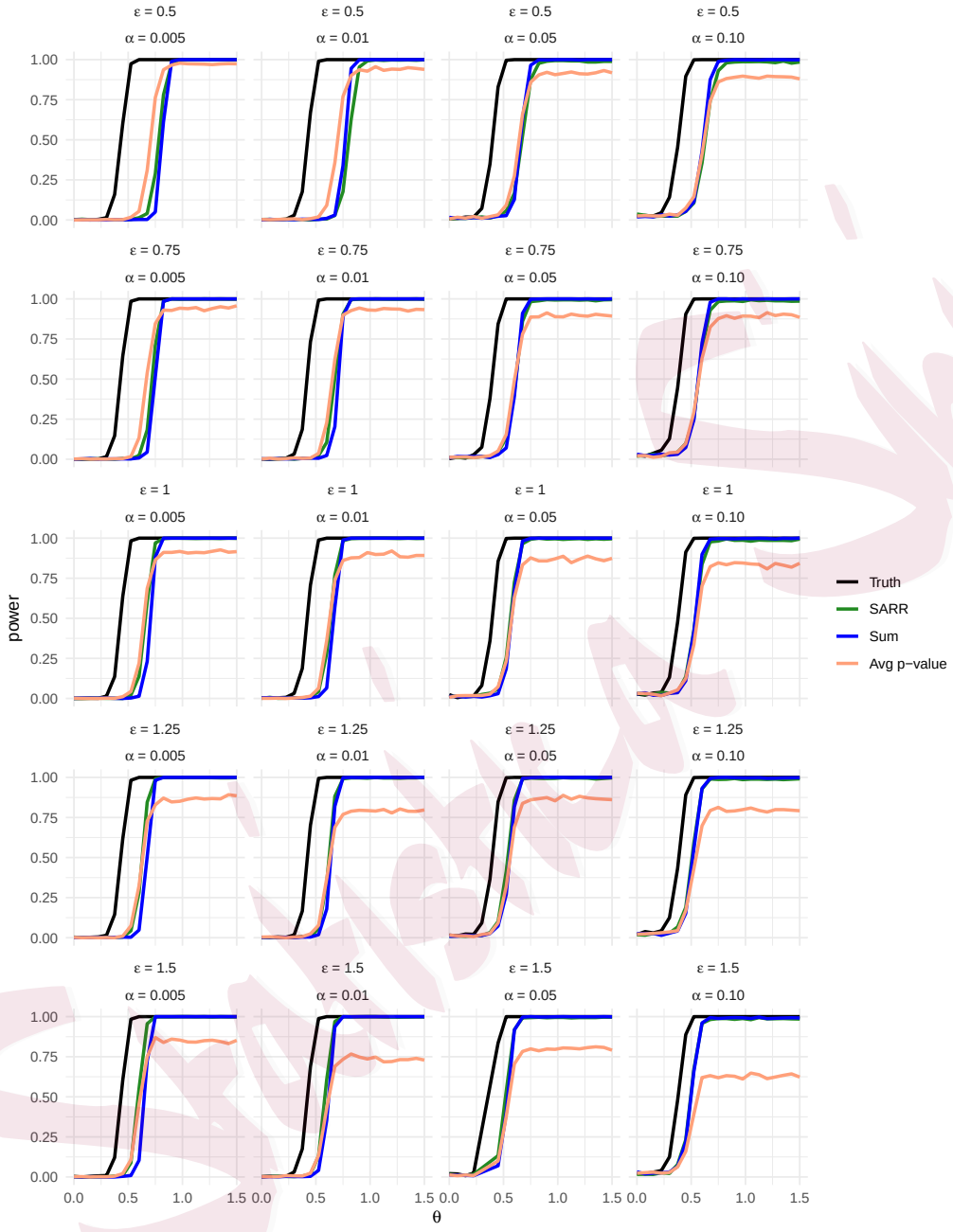


Figure 5: Goodness-of-fit tests: Average power of tests for skewness for different combinations of α and ε.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Table 2: Prior distributions for the one-sample Wilcoxon test and their values of $\mathbb{P}(d = 1 \mid H_1)$. "SA" stands for "subsample and aggregate."

Method	$\mathbb{P}(d = 1 \mid H_1)$
Truth	$0.8\overline{\gamma}_0^z$
SA + Randomized Response	$\int_0^1 \mathbb{P}(T > k \mid \gamma_0, H_1)\, \pi(\gamma_0 \mid H_1)\, \mathrm{d}\gamma_0$
SA + Average p-value	$\mathbb{P}(\overline{p} + \eta < c_\alpha \mid H_1)$
SA + Sum	$\mathbb{P}(\sum_{i=1}^{2k+1} x_i + \eta > \tilde{c}_\alpha \mid H_1)$

a prior on $\mathbb{P}(d = 1 \mid H_1)$ through simpler quantities for which we can define a prior more comfortably.

For the subsampled and aggregated randomized response mechanism, we define the prior $\pi(\gamma_0 \mid H_1) = \mathbb{P}(x_i = 1 \mid H_1)$ as follows. Let $\overline{\gamma}_0^z$ be the average power of the z-test induced by the unit information prior used in Example 2. Then, our prior $\pi(\gamma_0 \mid H_1)$ is a beta distribution parametrized in terms of its expected value μ and effective sample size κ with $\mu = 0.8\overline{\gamma}_0^z$ and $\kappa = 2k + 1$ (see Chapter 6 of Kruschke (2014) for a discussion on the convenience of this parametrization in Bayesian analysis).

For the subsampled and aggregated average p-value, we put a prior on the average p-value $\overline{p}$ under H_1. First, we find the expected p-value for the z-test $\overline{p}_z$ induced by the unit information prior. Then, we define our

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

prior on $\overline{p} \mid H_1$ as a beta distribution centered at $\mu = 0.8\overline{p}_z$ and with an effective sample size $\kappa = 2k + 1$. Given the prior $\overline{p} \mid H_1$ and $\eta \sim \text{Laplace}(0, 1/[\varepsilon(2k+1)])$, $\mathbb{P}(d = 1 \mid H_1)$ is the probability that $\overline{p} + \eta$ is below a critical value c_α so that $\mathbb{P}(\overline{p} + \eta < c_\alpha \mid H_0) = \alpha$.

For the subsampled and aggregated sum, we follow a similar strategy. We put a prior on the sum $\sum_{i=1}^{2k+1} x_i \sim \text{Binomial}(2k+1, \gamma_0)$ through $\pi(\gamma_0 \mid H_1) = \mathbb{P}(x_i = 1 \mid H_1)$, which is a beta distribution with $\mu = 0.8\overline{\gamma}_0^z$ and $\kappa = 2k + 1$ (as it was for the subsampled and aggregated randomized response mechanism). The probability $\mathbb{P}(d = 1 \mid H_1)$ is the probability that $\sum_{i=1}^{2k+1} x_i + \eta$ is above a critical value $\tilde{c}_\alpha$ so that $\mathbb{P}(\sum_{i=1}^{2k+1} x_i + \eta > \tilde{c}_\alpha \mid H_0) = \alpha$, where $\eta \sim \text{Laplace}(0, 1/\varepsilon)$.

Lastly, we include the "truth", defined as the result of running the usual nonprivate Wilcoxon test, without splitting the data or running any mechanisms. For that case, we take $\mathbb{P}(d = 1 \mid H_1) = 0.8\overline{\gamma}_0^z$.

Figure 6 displays the results of the simulation study. The methods that are based on binarized outcomes, namely the subsampled and aggregated randomized response mechanism and the subsampled and aggregated perturbed sum, tend to outperform the subsampled and aggregated p-value. The exception is the case $\varepsilon = 0.5$ and $\alpha = 0.005$. The performance of randomized response and the perturbed sum is similar. Randomized response

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

seems to be especially helpful when α is low and ε is greater than or equal to one.

4.3 Nonparametric ANOVA: Kruskal–Walis test

In this section, we report the results of a simulation study involving the Kruskal–Walis test. As a competitor, we include the test proposed in Section 3.3 of Couch et al. (2019), which is a differentially private method built specifically for the Kruskal–Walis test.

We simulate data independently from three groups: one where the data are distributed as Normal$(1, 1)$, another where the data are Normal$(2, 1)$, and another one where the data are Normal$(3, 1)$. We consider sample sizes ranging from 15 to 500 in increments of three so that the groups are balanced. The power is approximated after performing 10^4 simulations. This simulation study is similar that in Section 3.4 in Couch et al. (2019). We consider $\alpha \in \{0.005, 0.01, 0.05, 0.1\}$ and $\varepsilon \in \{0.5, 0.75, 1, 1.25, 1.5\}$.

As we did previously, we set the number of subgroups using the heuristic recommended in Section 3.3. When n, ε, and α are all small, the sample size is not large enough to simultaneously guarantee ε-differential privacy and a type-I error α if we use the randomized response mechanism (see Table 1 and the discussion in Section 3.3). For those cases, the method in

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

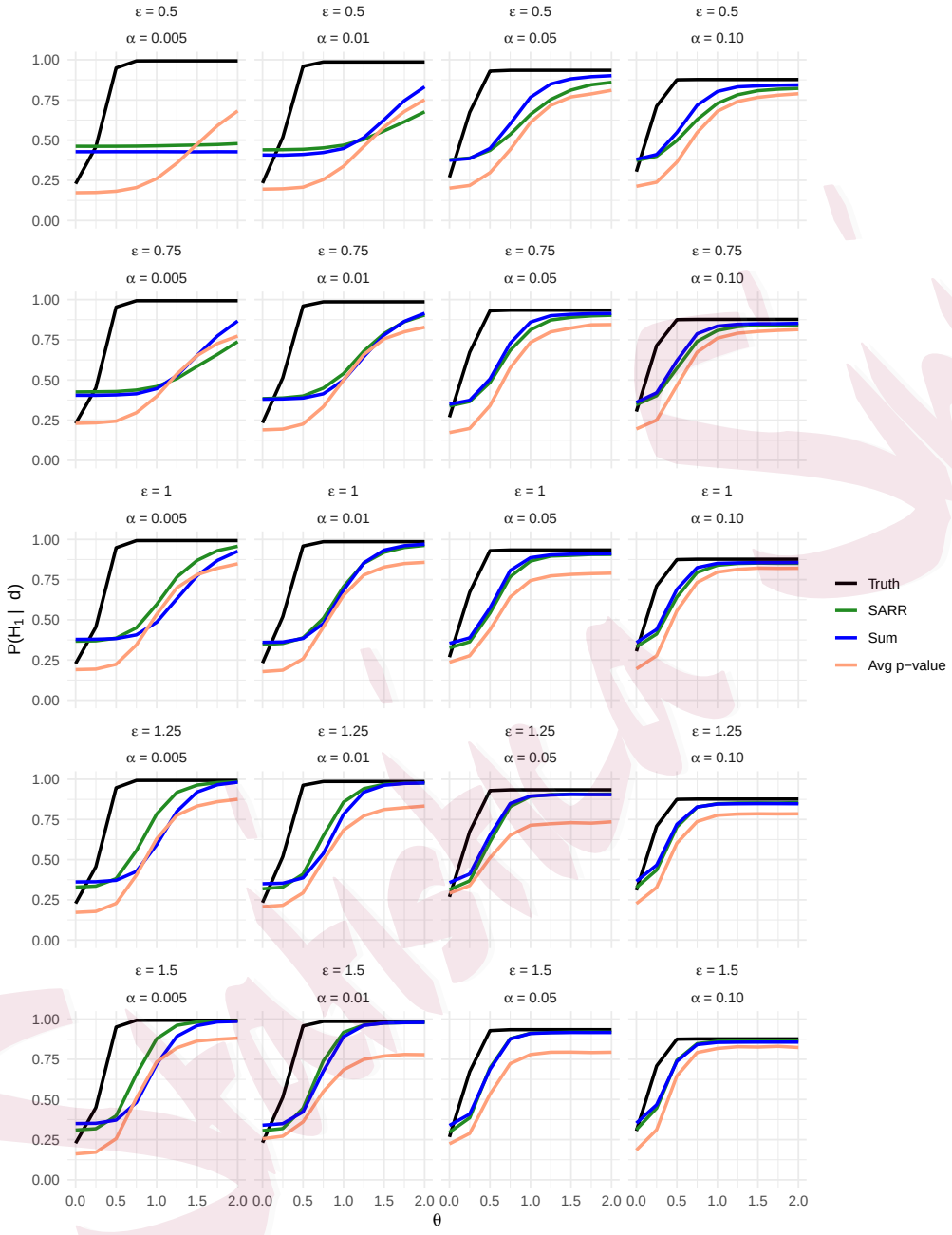


Figure 6: One-sample Wilcoxon test: Average posterior probability $\mathbb{P}(H_1 \mid d)$ for different values of α, ε, and location parameter θ.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Couch et al. (2019) can be run, but its power is low.

The results of the simulation study are displayed in Figure 7. Overall, the method in Couch et al. (2019) (labeled KWabs), which is tailored to this task, outperforms the general-purpose algorithms. The loss in power is most noticeable for smaller n, α, and ε. Randomized response is preferable over the sum when $\alpha \times \varepsilon \in \{0.005, 0.01\} \times \{1, 1.25, 1.5\}$. The sum outperforms randomized response when $\alpha \times \varepsilon \in \{0.05, 0.10\} \times \{0.5, 0.75\}$. In the remaining cases, the performance of the two approaches is relatively similar. The average p-value performs best (out of all general-purpose algorithms) for $\alpha \times \varepsilon \in \{0.005, 0.01\} \times \{0.5, 0.75\}$.

5. Discussion and future work

The subsampled and aggregated randomized response mechanism is a simple and effective tool for constructing differentially private tests from non-private tests. In our our examples, we have shown that the method is especially useful when the type-I error rate α is small and ε is greater than or equal to one.

We have focused on hypothesis testing, but the subsampled and aggregated randomized response mechanism can be useful in other contexts, especially when the data are naturally split into groups. One such example is federated learning (Konečnỳ et al., 2016), where the data are assumed to

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

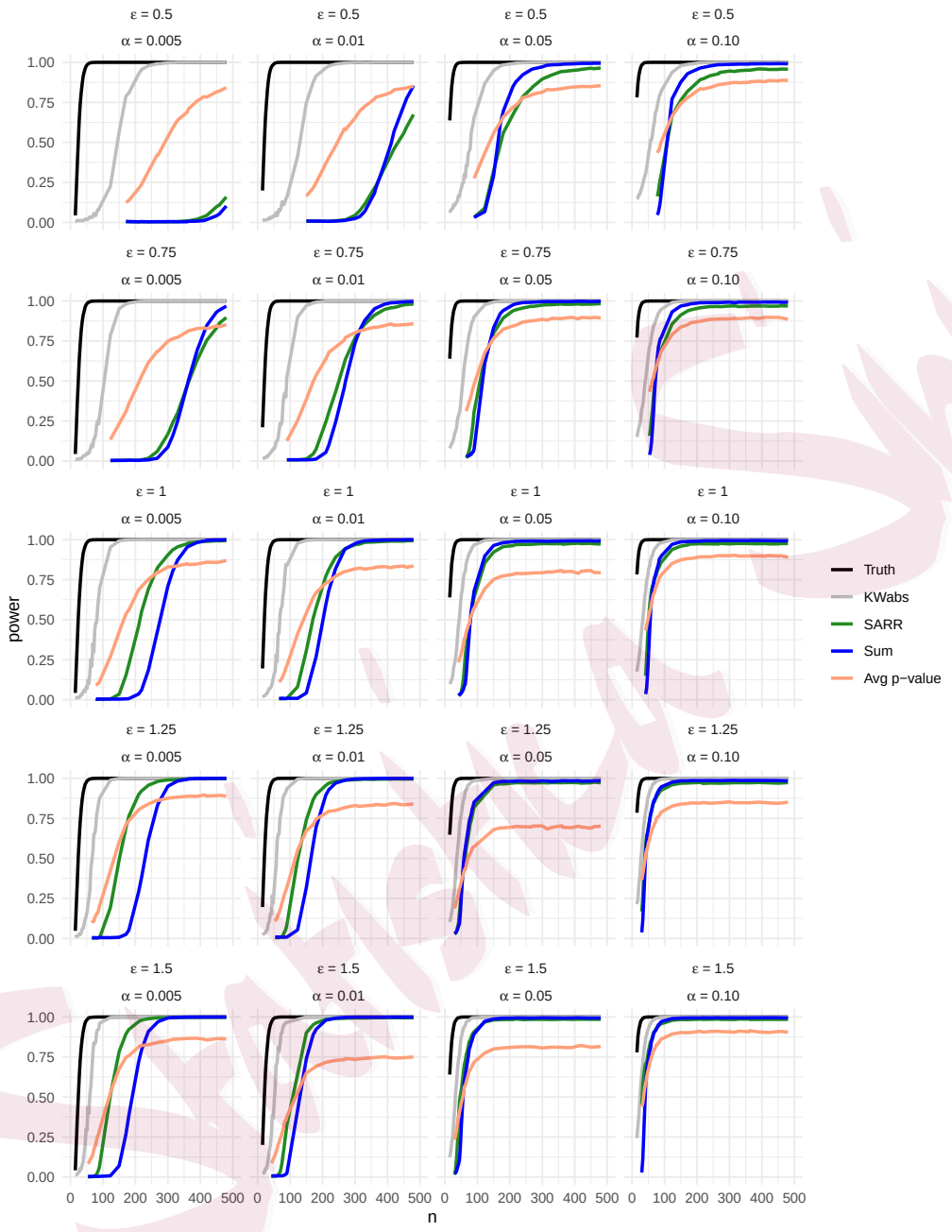


Figure 7: Kruskal–Walis test: Power for different values of α and ε for a range of total sample sizes n.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

be stored in different clients. Another application is multi-agent decision problems with confidentiality constraints, where the goal is to make a collaborative decision ensuring that the individual recommendations are kept private.

A drawback of our approach is that it cannot be used if the sample size, type-I error α, and privacy parameter ε are all small (see Table 1). However, in such instances, the differentially private tests that can be implemented are low-powered.

In Section 4.3, we compared general-purpose algorithms for differentially private testing with a test proposed in Couch et al. (2019) that was developed specifically for the Kruskal–Walis test. The test proposed in Couch et al. (2019) is considerably more powerful than the general-purpose algorithms for small α and ε, but its performance is comparable with that of the subsampled and aggregated randomized response mechanism when $\alpha \geq 0.05$ and $\varepsilon \geq 1$.

The subsampled and aggregated randomized response mechanism can be extended in a number of ways. For instance, the mechanism could output a categorical variable with multiple categories, instead of a binary decision. Another extension could accommodate multi-step multiple hypothesis testing methods, such as the Benjamini–Hochberg procedure (Benjamini and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Hochberg, 1995). In that case, the privacy level of the algorithm should be computed with care, because the outputs of the tests become dependent.

Supplementary Material

The Supplementary Material to this article contains proofs of the propositions stated in the main text and additional results from simulation studies.

References

Alabi, D. and S. Vadhan (2022). Hypothesis testing for differentially private linear regression. *arXiv preprint arXiv:2206.14449*.

Amitai, G. and J. Reiter (2018). Differentially private posterior summaries for linear regression coefficients. *Journal of Privacy and Confidentiality 8*(1).

Barrientos, A. F., J. P. Reiter, A. Machanavajjhala, and Y. Chen (2019). Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics 28*(2), 440–453.

Bassily, R. and A. Smith (2015). Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 127–135.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Bayarri, M., D. J. Benjamin, J. O. Berger, and T. M. Sellke (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology 72*, 90–103.

Bayarri, M. J. and J. O. Berger (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science 19*(1), 58–80.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. (2018). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 289–300.

Blair, G., K. Imai, and Y.-Y. Zhou (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association 110*(511), 1304–1319.

Chaudhuri, A. and R. Mukerjee (2020). *Randomized response: Theory and techniques*. Routledge.

Couch, S., Z. Kazan, K. Shi, A. Bray, and A. Groce (2019). Differentially

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

private nonparametric hypothesis testing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 737–751.

Dimitrakakis, C., B. Nelson, Z. Zhang, A. Mitrokotsa, and B. I. Rubinstein (2017). Differential privacy for bayesian inference through posterior sampling. *Journal of Machine Learning Research 18*(11), 1–39.

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

Dwork, C., A. Roth, et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science 9*(3-4), 211–407.

Erlingsson, Ú., V. Pihur, and A. Korolova (2014). RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067.

Gaboardi, M., H. Lim, R. Rogers, and S. Vadhan (2016). Differentially private chi-squared hypothesis testing: Goodness of fit and independence

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

testing. In *International Conference on Machine Learning*, pp. 2111–2120. PMLR.

Gelman, A. and E. Loken (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University, Technical Report*.

Geumlek, J., S. Song, and K. Chaudhuri (2017). Renyi differential privacy mechanisms for posterior sampling. *Advances in Neural Information Processing Systems 30*.

Heikkilä, M., J. Jälkö, O. Dikmen, and A. Honkela (2019). Differentially private Markov chain Monte Carlo. *Advances in Neural Information Processing Systems 32*.

Hu, J., T. D. Savitsky, and M. R. Williams (2022). Private tabular survey data products through synthetic microdata generation. *Journal of Survey Statistics and Methodology 10*(3), 720–752.

Ju, N., J. A. Awan, R. Gong, and V. A. Rao (2022). Data augmentation MCMC for Bayesian inference from privatized data. *arXiv preprint arXiv:2206.00710*.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*(431), 928–934.

Konečnỳ, J., H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lei, J., A.-S. Charest, A. Slavkovic, A. Smith, and S. Fienberg (2018). Differentially private model selection with penalized and constrained likelihood. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Ma, F. and P. Wang (2021). Randomized response mechanisms for differential privacy data analysis: Bounds and applications. *arXiv preprint arXiv:2112.07397*.

McSherry, F. D. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pp. 19–30.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Mohan, P., A. Thakurta, E. Shi, D. Song, and D. Culler (2012). GUPT: privacy preserving data analysis made easy. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 349–360.

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

Peña, E. A. and E. H. Slate (2006). Global validation of linear model assumptions. *Journal of the American Statistical Association 101*(473), 341–354.

Peña, V. and A. F. Barrientos (2021). Differentially private methods for managing model uncertainty in linear regression models. *arXiv preprint arXiv:2109.03949*.

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

Schervish, M. J. (1996). p-values: what they are and what they are not. *The American Statistician 50*(3), 203–206.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Shaked, M. and J. G. Shanthikumar (2007). *Stochastic orders*. Springer Science & Business Media.

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

Su, D., J. Cao, N. Li, E. Bertino, and H. Jin (2016). Differentially private k-means clustering. In *Proceedings of the sixth ACM conference on data and application security and privacy*, pp. 26–37.

Thakurta, A. G. and A. Smith (2013). Differentially private feature selection via stability arguments, and the robustness of the LASSO. In *Conference on Learning Theory*, pp. 819–850. PMLR.

Wang, Y., X. Wu, and D. Hu (2016). Using randomized response for differential privacy preserving data collection. In *EDBT/ICDT Workshops*, Volume 1558, pp. 0090–6778.

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

Ye, Q., H. Hu, X. Meng, and H. Zheng (2019). PrivKV: Key-value data

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

collection with local differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 317–331. IEEE.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0279

Víctor Peña, Universitat Politècnica de Catalunya

E-mail: victor.pena.pizarro@upc.edu

Andrés F. Barrientos, Florida State University

E-mail: abarrientos@fsu.edu