Statistica Sinica: Supplement

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

Víctor Peña, Andrés Felipe Barrientos

Universitat Politècnica de Catalunya, Florida State University

Supplementary Material

This document contains supplementary material to the main text of the article. Section S1 includes auxiliary results needed to prove the results in the main text, which are proved in Section S2. In Section S3, we present an additional simulation study where we compare our method to the differentially private t-test proposed in Barrientos et al. (2019). Lastly, Section S4 presents the results of the test for the kurtosis of errors proposed in Peña and Slate (2006), in the context of the simulation study reported in Section 4.1 of the main text.

S1 Auxiliary results

First, we prove auxiliary propositions that are ancillary for proving the results in the main text. All of them use theorems in Shaked and Shanthikumar (2007).

We use the notation Binomial(i, p) + Binomial(j, p) for the distribution of the sum of independent Binomial(i, p) and Binomial(j, p) random variables, with the understanding that if the number of trials is zero, the random variable is zero with probability one.

Definition 1. Let X and Y be discrete random variables with common support S, which is a subset of the integers. We say that X stochastically dominates Y with respect to the likelihood ratio order if $\mathbb{P}(X = t)/\mathbb{P}(Y = t)$ is increasing in t for $t \in S$.

Proposition S1.1. Let $B_i \sim$ Binomial$(n - i, 1 - p)$ + Binomial(i, p) with $1/2 < p < 1$ for $i \in \{0, 1, ..., n\}$. For any $\{i, j\} \subset \{0, 1, \dots, n\}$ such that $j > i$, B_j stochastically dominates B_i with respect to the likelihood ratio ordering.

Proof. The result follows by an application of Theorem 1.C.9. in Shaked and Shanthikumar (2007). To see this, let $\{i, j\} \subset \{0, 1, \dots, n\}$ such that $j > i$. Let $C_i, C_j \sim$ Binomial(i, p) + Binomial$(n - j, 1 - p)$, $N_i \sim$ Binomial$(j -$

$i, 1-p)$, and $N_j \sim \text{Binomial}(j-i, p)$. We can write $B_i = C_i + N_i$ and $B_j = C_j + N_j$. Binomial random variables have log-concave probability mass functions, C_j stochastically dominates C_i with respect to the likelihood ratio ordering (they are equal in distribution), and N_j dominates N_i with respect to the likelihood ratio ordering because $p > 1/2$ by assumption. We can apply Theorem 1.C.9. in Shaked and Shanthikumar (2007) and the result follows. □

Proposition S1.2. Let $1/2 < p < 1$ and $B_i \sim \text{Binomial}(i, p)+\text{Binomial}(n-i, 1-p)$ for $i \in \{0, 1, ..., n\}$. For any $\{i, j\} \subset \{0, 1, \dots, n\}$ such that $j > i$ and $x \in \{0, 1, \dots., n\}$,

$$\frac{\mathbb{P}(B_i = x)}{\mathbb{P}(B_i > x)} \geq \frac{\mathbb{P}(B_j = x)}{\mathbb{P}(B_j > x)}.$$

Proof. This is a direct consequence of Proposition S1.1 and the fact that stochastic domination according to the likelihood ratio order implies stochastic domination according to the hazard ratio order (see e.g. Theorem 1.C.1. Shaked and Shanthikumar (2007)). □

Proposition S1.3. Let $1/2 < p < 1$ and $B_i \sim \text{Binomial}(n-i, 1-p) + \text{Binomial}(i, p)$ and $x \in \{0, 1, \dots, n\}$. Then, the ratio

$$r_j = \frac{\mathbb{P}(B_i > x)}{\mathbb{P}(B_{i-1} > x)}$$

is decreasing in i.

Proof. It suffices to show that for $i \in \{1, 2, \ldots, n-1\}$ and $x \in \{0, 1, \ldots, n\}$,

$$\frac{\mathbb{P}(B_i > x)}{\mathbb{P}(B_{i-1} > x)} \geq \frac{\mathbb{P}(B_{i+1} > x)}{\mathbb{P}(B_i > x)}.$$

Let $C_{i+1} \sim \text{Binomial}(i, p) + \text{Binomial}(n - i - 1, 1 - p)$. Then,

$$\mathbb{P}(B_{i+1} > x) - \mathbb{P}(B_i > x) = (2p - 1)\mathbb{P}(C_{i+1} = x)$$

$$\mathbb{P}(B_i > x) = (1 - p)\mathbb{P}(C_{i+1} > x - 1) + p\mathbb{P}(C_{i+1} > x).$$

Similarly, letting $C_i \sim \text{Binomial}(i - 1, p) + \text{Binomial}(n - i, 1 - p)$,

$$\mathbb{P}(B_i > x) - \mathbb{P}(B_{i-1} > x) = (2p - 1)\mathbb{P}(C_i = x)$$

$$\mathbb{P}(B_{i-1} > x) = (1 - p)\mathbb{P}(C_i > x - 1) + p\mathbb{P}(C_i > x).$$

Now,

$$\frac{\mathbb{P}(B_i > x)}{\mathbb{P}(B_{i-1} > x)} \geq \frac{\mathbb{P}(B_{i+1} > x)}{\mathbb{P}(B_i > x)} \Leftrightarrow \frac{\mathbb{P}(B_{i-1} > x)}{\mathbb{P}(B_i > x) - \mathbb{P}(B_{i-1} > x)} \leq \frac{\mathbb{P}(B_i > x)}{\mathbb{P}(B_{i+1} > x) - \mathbb{P}(B_i > x)}.$$

The inequality on the right-hand side is equivalent to

$$(1 - p)\frac{\mathbb{P}(C_i > x - 1)}{\mathbb{P}(C_i = x)} + p\frac{\mathbb{P}(C_i > x)}{\mathbb{P}(C_i = x)} \leq (1 - p)\frac{\mathbb{P}(C_{i+1} > x - 1)}{\mathbb{P}(C_{i+1} = x)} + p\frac{\mathbb{P}(C_{i+1} > x)}{\mathbb{P}(C_{i+1} = x)}$$

Since C_{i+1} stochastically dominates C_i according to the likelihood ratio ordering, Proposition S1.2 implies that $\mathbb{P}(C_i > x)/\mathbb{P}(C_i = x) \leq \mathbb{P}(C_{i+1} > x)/\mathbb{P}(C_{i+1} = x)$. It remains to show that

$$\frac{\mathbb{P}(C_i > x - 1)}{\mathbb{P}(C_i = x)} \leq \frac{\mathbb{P}(C_{i+1} > x - 1)}{\mathbb{P}(C_{i+1} = x)}.$$

The inequality is true after substituting $\mathbb{P}(C_j > x - 1) = \mathbb{P}(C_j > x) + \mathbb{P}(C_j = x)$ in both numerators and then applying Proposition S1.2. $\square$

Proposition S1.4. Let $1/2 < p < 1$ and $B_i \sim \text{Binomial}(n - i, 1 - p) + \text{Binomial}(i, p)$. Then, the ratio

$$r_x = \frac{\mathbb{P}(B_i > x)}{\mathbb{P}(B_{i-1} > x)}$$

is increasing in x for $x \in \{0, 1, \ldots, n\}$.

Proof. Let $i \in \{0, 1, ..., n\}$ and $x \in \{0, 1, \ldots n - 1\}$. It is enough to show that $r_x \leq r_{x+1}$. Now,

$$\begin{aligned} r_x \leq r_{x+1} &\Leftrightarrow \frac{\mathbb{P}(B_i > x)}{\mathbb{P}(B_i > x + 1)} \leq \frac{\mathbb{P}(B_{i-1} > x)}{\mathbb{P}(B_{i-1} > x + 1)} \\ &\Leftrightarrow \frac{\mathbb{P}(B_i = x + 1) + \mathbb{P}(B_i > x + 1)}{\mathbb{P}(B_i > x + 1)} \leq \frac{\mathbb{P}(B_{i-1} = x + 1) + \mathbb{P}(B_{i-1} > x + 1)}{\mathbb{P}(B_{i-1} > x + 1)}, \end{aligned}$$

which is equivalent to

$$\frac{\mathbb{P}(B_i = x + 1)}{\mathbb{P}(B_i > x + 1)} \leq \frac{\mathbb{P}(B_{i-1} = x + 1)}{\mathbb{P}(B_{i-1} > x + 1)}.$$

The inequality above is shown to be true in Proposition S1.2. $\square$

S2 Proofs of Propositions in main text

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

Proof. This is a standard result. It follows directly from the definition of differential privacy. □

Proposition 2. The statistic $d_c = \mathbb{1}(T > c)$ is exactly ε-differentially 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 $i \in \{0, 1\}$.

Proof. Let $1/2 < p < 1$ and $B_i \sim \text{Binomial}(i,p)+\text{Binomial}(2k+1-i, 1-p)$ for $i \in \{0, 1, ..., 2k+1\}$. Let $r_i = \mathbb{P}(B_i > c)$ be the probability of rejecting the null hypothesis given that $\sum_{j=1}^{2k+1} x_j = i$. By the definition of differential privacy,

$$\exp(\varepsilon) = \max_{i\in\{1,\ldots,2k+1\}} \max\left\{\frac{r_i}{r_{i-1}}, \frac{r_{i-1}}{r_i}, \frac{1-r_i}{1-r_{i-1}}, \frac{1-r_{i-1}}{1-r_i}\right\}.$$

From Proposition S1.1, we know that $r_i > r_{i-1}$ (first order or "usual" stochastic domination is implied by likelihood ratio domination, see e.g. Theorem 1.C.1 in Shaked and Shanthikumar (2007)). Therefore,

$$\max\left\{\frac{r_i}{r_{i-1}}, \frac{r_{i-1}}{r_i}, \frac{1-r_i}{1-r_{i-1}}, \frac{1-r_{i-1}}{1-r_i}\right\} = \max\left\{\frac{r_i}{r_{i-1}}, \frac{1-r_{i-1}}{1-r_i}\right\}.$$

From Proposition S1.3, we know that r_i/r_{i-1} is decreasing in i. This narrows down our candidates for the maximum to

$$\exp(\varepsilon) = \max\{r_1/r_0, (1-r_{2k})/(1-r_{2k+1})\}.$$

Note that

$$1 - r_{2k+1} = 1 - \mathbb{P}(B_{2k+1} > t) = 1 - \mathbb{P}(B_0 < 2k + 1 - t) = \mathbb{P}(B_0 > 2k - t)$$

$$1 - r_{2k} = 1 - \mathbb{P}(B_{2k} > t) = 1 - \mathbb{P}(B_1 < 2k + 1 - t) = \mathbb{P}(B_1 > 2k - t).$$

We can rewrite

$$\exp(\varepsilon) = \max\left\{\frac{\mathbb{P}(B_1 > c)}{\mathbb{P}(B_0 > c)}, \frac{\mathbb{P}(B_1 > 2k - c)}{\mathbb{P}(B_0 > 2k - c)}\right\}.$$

The result follows because Proposition S1.4 shows that the ratio $\mathbb{P}(B_1 > x)/\mathbb{P}(B_0 > x)$ is increasing in x.

□

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 $c = k$.

Proof. First, we show that for any fixed k and c, ε is increasing in p. We do so by showing that $1/(\exp(\varepsilon) - 1)$ is decreasing in p. We can write

$$\frac{1}{\exp(\varepsilon) - 1} = \frac{\mathbb{P}(B_0 > c_*)}{\mathbb{P}(B_1 > c_*) - \mathbb{P}(B_0 > c_*)}.$$

Let $B \sim \text{Binomial}(2k, 1 - p)$. Then,

$$\mathbb{P}(B_1 > c_*) - \mathbb{P}(B_0 > c_*) = (2p - 1)\mathbb{P}(B = c_*).$$

Plugging in our new expression for $\mathbb{P}(B_1 > c_*) - \mathbb{P}(B_0 > c_*)$, we obtain

$$\frac{1}{\exp(\varepsilon)-1} = \frac{\mathbb{P}(B_0 > c_*)}{(2p-1)\mathbb{P}(B = c_*)} = \frac{1-p}{2p-1}\frac{P(B > c_* - 1)}{P(B = c_*)} + \frac{p}{2p-1}\frac{P(B > c_*)}{P(B = c_*)}.$$

Rearranging terms,

$$\frac{1}{\exp(\varepsilon)-1} = \frac{1-p}{2p-1} + \frac{1}{2p-1}\frac{P(B > c_*)}{P(B = c_*)}.$$

The result follows because $(1-p)/(2p-1)$ and $1/(2p-1)$ are decreasing in p for $1/2 < p < 1$, and $P(B > c_*)/P(B = c_*)$ is also decreasing in p. The latter is true because if p increases, B is stochastically decreasing according to the likelihood ratio and hazard ratio order (see e.g. Example 1.C.51. in Shaked and Shanthikumar (2007)), which in turn implies that $P(B > c_*)/P(B = c_*)$ is decreasing, as required.

Now, we show that for any fixed p and $c \geq k$, ε is decreasing in k. Since $c \geq k$, we know by Proposition 1 that $c_* = c$ and $\exp(\varepsilon) = \mathbb{P}(B_1 > c)/\mathbb{P}(B_0 > c)$. Let $B \sim \text{Binomial}(2k, 1-p)$. Then, the ratio can be rewritten as

$$\frac{\mathbb{P}(B_1 > c)}{\mathbb{P}(B_0 > c)} = 1 + \frac{2p-1}{1-p+\frac{\mathbb{P}(B>c)}{\mathbb{P}(B=c)}}$$

Therefore, ε is decreasing in k if and only if $\mathbb{P}(B = c)/\mathbb{P}(B > c)$ is decreasing in k. The result follows because B is stochastically increasing in k with respect to the likelihood ratio order, so the ratio $\mathbb{P}(B = c)/\mathbb{P}(B > c)$ is decreasing in k (again, this is implied by the fact that stochastic domination with respect to the likelihood ratio order implies domination with respect to the hazard ratio order).

Finally, we show that for any fixed k and p, ε is minimized at $c = k$. From Proposition S1.4, we know that ε is increasing in c_*. Since $c_* = \max(c, 2k - c)$, it follows that for fixed k, and p, ε attains its minimum at $c = k$, as required.

□

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

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

A sufficient condition on p for achieving ε differential privacy is

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

Proof. For proving the results in this proposition, it will be useful to rewrite $\exp(\varepsilon)$ in terms of $B \sim \text{Binomial}(2k, 1-p)$.

Let $B_0 \sim \text{Binomial}(2k+1, 1-p)$, and $B_1 \sim \text{Binomial}(1, p) + \text{Binomial}(2k, 1-$

p). By Proposition 1, we know that

$$\exp(\varepsilon) = \frac{\mathbb{P}(B_1 > k)}{\mathbb{P}(B_0 > k)}.$$

Let $B \sim \text{Binomial}(2k, 1-p)$. The ratio can be rewritten as

$$\frac{\mathbb{P}(B_1 > k)}{\mathbb{P}(B_0 > k)} = 1 + \frac{2p-1}{1-p+\frac{\mathbb{P}(B>k)}{\mathbb{P}(B=k)}}. \tag{S2.1}$$

This shows that $\exp(\varepsilon)$ depends on k only through $\mathbb{P}(B > k)/\mathbb{P}(B = k)$, which can be expressed as

$$\frac{\mathbb{P}(B > k)}{\mathbb{P}(B = k)} = \sum_{i=1}^{k} \frac{P(B = k+i)}{P(B = k)} = \sum_{i=1}^{k} \left\{ \prod_{j=1}^{i} \frac{k-j+1}{k+j} \right\} \left(\frac{1-p}{p} \right)^i. \tag{S2.2}$$

First, we find the the limit of ε as k grows for fixed $1/2 < p < 1$.

The product term in Equation (S2.2) can be bounded as follows:

$$\left(\frac{k-i+1}{k+i} \right)^i \leq \prod_{j=1}^{i} \frac{k-j+1}{k+j} \leq 1.$$

Then,

$$\lim_{k\to\infty} \frac{\mathbb{P}(B > k)}{\mathbb{P}(B = k)} \leq \sum_{i=1}^{\infty} \left(\frac{p}{1-p} \right)^i = \frac{1-p}{2p-1}.$$

Consider the series

$$\lim_{k\to\infty} \sum_{i=1}^{k} \left(\frac{k-i+1}{k+i} \right)^i \left(\frac{1-p}{p} \right)^i.$$

All the terms are positive and the summand is increasing in k, so we can apply the monotone convergence theorem for series:

$$\lim_{k\to\infty} \sum_{i=1}^{k} \left(\frac{k-i+1}{k+i} \right)^i \left(\frac{1-p}{p} \right)^i = \lim_{k\to\infty} \sum_{i=1}^{k} \left(\frac{1-p}{p} \right)^i = \frac{1-p}{2p-1}.$$

Therefore, we conclude that

$$\lim_{k\to\infty} \frac{\mathbb{P}(B > k)}{\mathbb{P}(B = k)} = \frac{1-p}{2p-1}.$$

Plugging the limit into Equation (S2.1), we obtain:

$$\frac{\mathbb{P}(B_1 > k)}{\mathbb{P}(B_0 > k)} = 1 + \frac{(2p-1)^2}{2p(1-p)},$$

as required.

The second part of the proposition is a direct consequence of previous results. The sufficient condition corresponds to the case $k = 0$, and it is sufficient because ε is decreasing in k, all else being equal. The necessary condition can be found by solving for p in the limiting expression of ε when k grows to infinity.

□

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

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

Proof. The result can be proved by letting $s \in \{0, 1, \dots, k\}$ and noting that $\mathbb{P}(B_s > k) = P(B_{2k+1-s} \leq k)$ for $B_s \sim \text{Binomial}(s, p) + \text{Binomial}(2k+1-s, 1-p)$ and $B_{2k+1-s} \sim \text{Binomial}(2k+1-s, p) + \text{Binomial}(s, 1-p)$.

□

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

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 $p = 1$.

Proof. We prove the three statements separately.

1. If $s > k$, the probability of disagreement is $\mathbb{P}(B_s \leq k)$, whereas if $s \leq k$, it is $\mathbb{P}(B_s > k)$. The result follows because, by Proposition S1.1, B_s is stochastically increasing in s.

2. If $s \leq k$, then $\mathbb{P}(d \neq \tilde{d}) = \mathbb{P}(B_s > k)$, which is decreasing in k. To see this, let k be fixed and increase it by one, defining $k_* = k + 1$. Then, we can define $B_s^* \sim B_s + \text{Bernoulli}(1 - p)$.

$$\mathbb{P}(B_s^* > k_*) = (1 - p)\mathbb{P}(B_s > k) + p\mathbb{P}(B_s > k + 1),$$

which is smaller than $\mathbb{P}(B_s > k)$ because $p \geq 1/2$ and $\mathbb{P}(B_s > k + 1) < \mathbb{P}(B_s > k)$. If $s > k$, then $\mathbb{P}(d \neq d) = \mathbb{P}(B_s \leq k)$, and a similar argument to the one we just used shows that it is increasing in k.

3. This proof of this part is direct given the expression of $\mathbb{P}(d \neq \tilde{d} \mid \sum_{i=1}^{2k+1} x_i = s)$.

□

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 decreasing 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 $\gamma_0 \geq 1/2$.

3. Let $p > 1/2$ be fixed. If $\gamma_0 > 1/2$, then the probability that d rejects H_0 goes to 1 as $k \to \infty$. Alternatively, if $\gamma_0 < 1/2$, then the probability that d rejects H_0 goes to 0 as $k \to \infty$.

Proof. The probability that d rejects H_0 is $\mathbb{P}(T > k)$ for $T \sim \text{Binomial}(2k+1, p\gamma_0 + (1-p)(1-\gamma_0))$. We prove the three statements separately.

1. Since $p \geq 1/2$, $\mathbb{P}(T > k)$ is increasing in γ_0.

2. This is similar to part 1. If $\gamma_0 < 1/2$, then $p\gamma_0 + (1-p)(1-\gamma_0)$ is decreasing in p. If $\gamma_0 \geq 1/2$, then it is increasing in p.

3. If $p > 1/2$ and $\gamma_0 > 1/2$, then $p\gamma_0 + (1-p)(1-\gamma_0) > 1/2$ and $\mathbb{P}(T > k)$ goes to one as k goes to infinity. Similarly, if $p > 1/2$ and $\gamma_0 < 1/2$, then $p\gamma_0 + (1-p)(1-\gamma_0) < 1/2$ and $\mathbb{P}(T > k)$ goes to zero as k goes to infinity. One can find these limits using standard tail bounds for

the binomial distribution.

□

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

Proof. The minimum type I error is achieved when $\alpha_0 = 0$. The type I error of d is then $\mathbb{P}(T > k)$ for $T \sim \text{Binomial}(2k+1, 1-p)$, where $p > 1/2$ is set given ε and k. The probability goes to zero as k goes to infinity because $p > 1/2$. □

S3 Differentially private t-test

In this section, we report the result of a simulation study where we compare the performance of the subsampled-and-aggregated randomized response mechanism (labeled "SARR" in the figures) to the differentially private t-test for regression proposed in Barrientos et al. (2019) (labeled "DP t") . This is a specific test that is only applicable to this task. We also compare our method to the test based on the sum of binary variables (labeled "Sum") and the average p-value (labeled "Avg p-value") that we explained in the main text.

The data are simulated from the normal linear model $y = X\beta + \varepsilon$

where $\beta = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]' = [1, 1, 1, 1, 1]'$. We test $H_0 : \beta_1 = 0$ against $H_1 = \beta_1 \neq 0$. For the method in Barrientos et al. (2019), we set the number of subsets to 25 and the truncation parameter to $a = 2$, following what was proposed in Barrientos et al. (2019). The number of subgroups for the other methods based on data-splitting is set using the strategy proposed in Section 3.3 of the main text with $\alpha_{0,\min} = \alpha$. We consider $\alpha \in \{0.005, 0.1, 0.05, 0.1\}$, $\varepsilon \in \{0.5, 0.75, 1, 1.25, 1.5\}$ and a range of sample sizes n that runs up to 10^5. For each scenario, we perform 10^4 simulations.

The test proposed in Barrientos et al. (2019) outperforms the general-purpose algorithms in all cases except when $\varepsilon = 0.5$. The binarized sum and the average p-value are best in this case. The randomized response mechanism performs best when $\alpha = 0.005$ and $\varepsilon \in \{1.25, 1.5\}$.

S4 Goodness-of-fit: Kurtosis

Figure 2 displays the result of the test for kurtosis that is part of Section 4.1 of the main text. Like we observed in the other scenarios, randomized response performs best when $\alpha = 0.005$ and ε is greater or equal to 1. The sum performs best for high α and low ε. The average p-value is best when both α and ε are small.

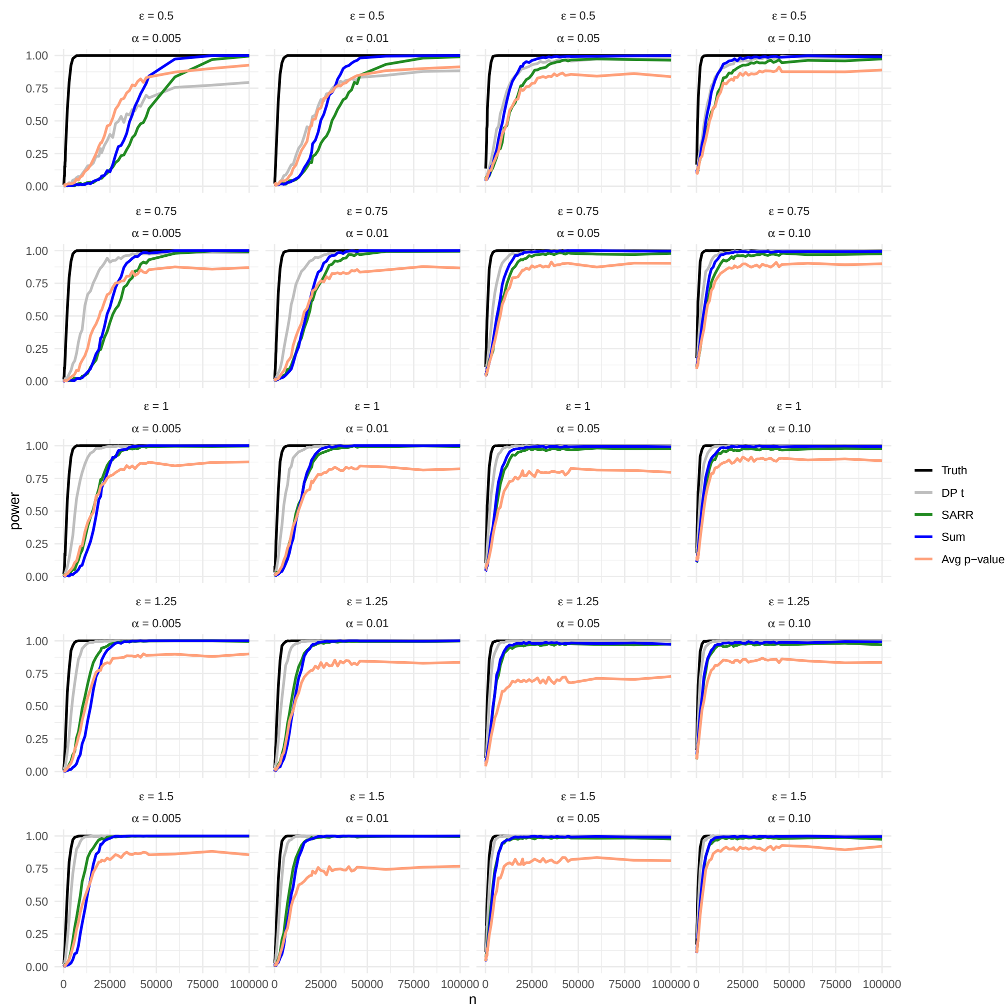

Figure 1: Differentially private t-test: Average power of t-tests for regression for different combinations of α and ε as a function of the total sample size n.

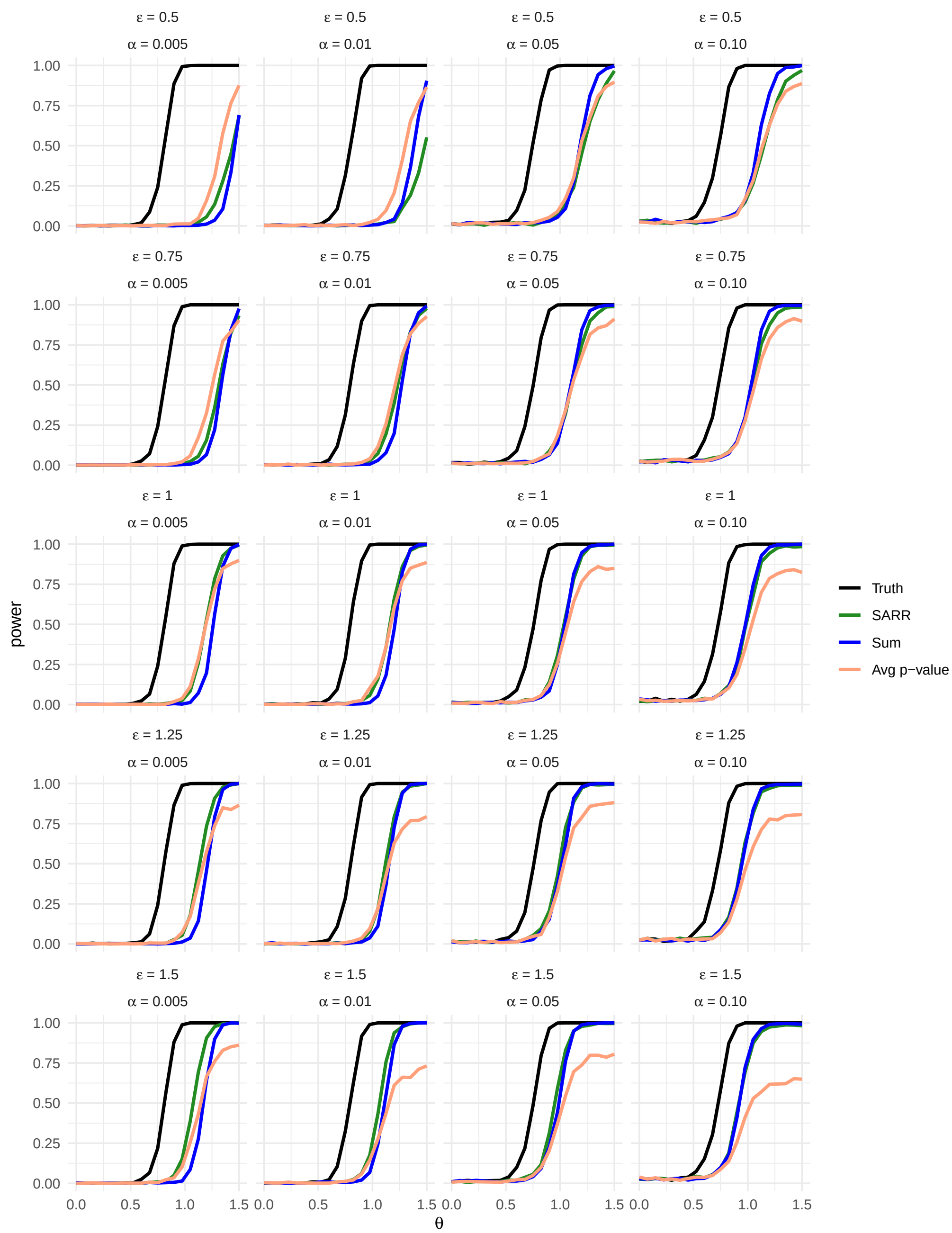

Figure 2: Goodness-of-fit tests: Average power of tests for kurtosis for different combinations of α and ε.

Bibliography

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.

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

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