Differentially Private Nonparametric Hypothesis Testing

CCS 2019

Andrew Bray, Simon Couch, Adam Groce, Zeki Kazan, Kaiyan Shi

Mathematics Department, Reed College (email: agroce@reed.edu, abray@reed.edu)

Objectives

We aim to develop and assess DP analogs to three rank-based statistical tests. For each we:

I Construct a mechanism for the release of a private statistic and bound its sensitivity.

II Assess the relative effectiveness of methods by comparing power curves.

Rank-based Statistics

Rank-based tests were devised as an alternative to tests with distributional assumptions. Instead of using the raw data where each obs. has a value (y_i), and group membership (g_i), statistics are based on the ranks (r_i) and signs (s_i) of the y_i.

i	y_i	g_i		r_i	s_i
1	3	1		5	1
2	2	1		4	1
3	-2	2		1	-1
4	-1	2	$\rightarrow$	2.5	-1
5	-1	3		2.5	-1
6	4	3		6	1

Raw data — Rank and sign data

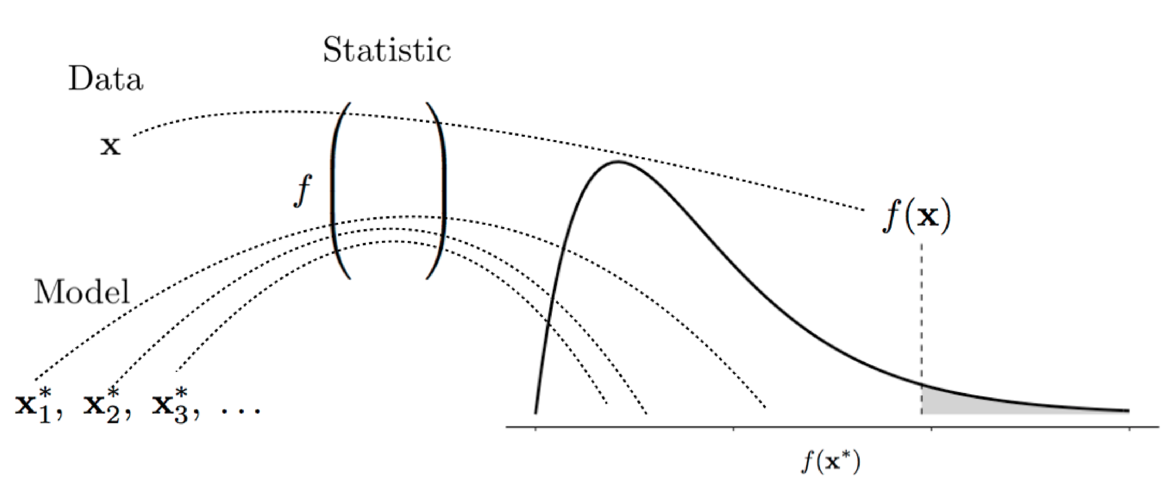
Hypothesis Testing

Goal: measure whether a particular data set is consistent with a given theory (H_0).

Steps:

1. Select and compute meaningful test statistic t.
2. Determine distribution of $T = f(\mathbf{X})$ when database $\mathbf{X}$ is drawn according to H_0.
3. Compute the *p-value*:

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



Wilcoxon Signed-Rank

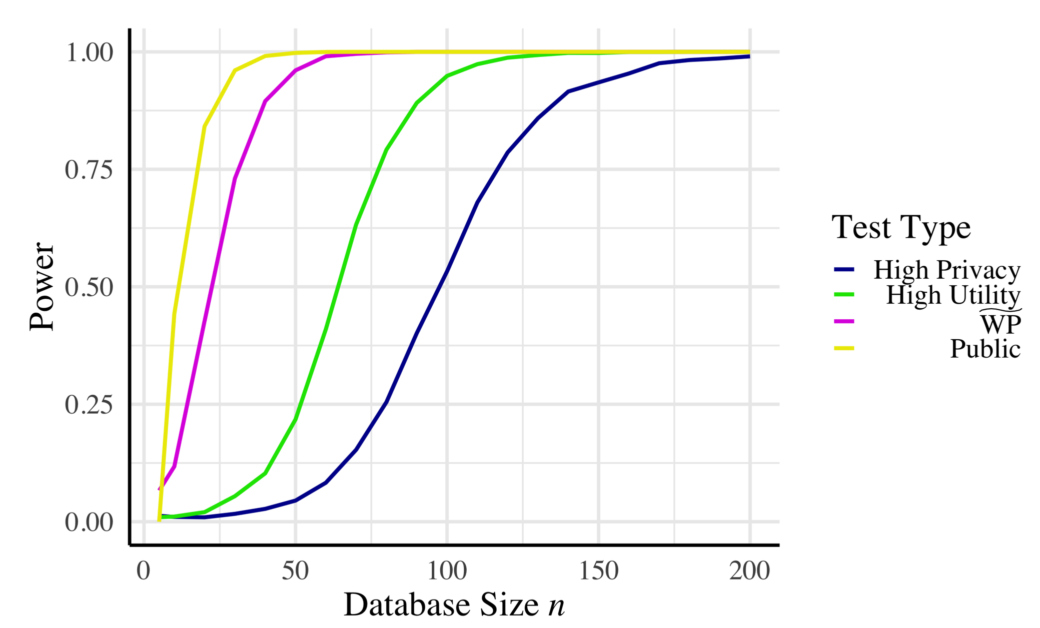
Setting Each observation has two paired values and their difference d_i. We evaluate whether these two values come from the same distribution.

Public statistic Rows with $d_i = 0$ are removed, then the remaining rows are assigned ranks and signs. The test statistic $\mathcal{W}$ is:

$$\mathcal{W} = \sum_i s_i r_i$$

Our contribution Prior work [2] adds Laplace noise to $\mathcal{W}$ and analytically bounds the reference distribution. We instead use an alternate statistic that does not drop $d_i = 0$ rows, and simulate the exact reference distribution.

Results We require between 8% and 40% as much data as prior work to achieve the same power.



Simulation and Power

After computing the private test statistic, two forms of simulation are used to find the reference distribution.

- Take many draws from $\mathbf{X} \leftarrow H_0$ directly, calculate many $f(\mathbf{x}^*)$, and add i.i.d. Laplacian to each.
- If distribution of T is known, draw many t, and add i.i.d. Laplacian to each.

Hypothesis tests are judged by their *statistical power*: the probability to detect an effect if it exists.

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

The empirical power curves above show power as a function of database size, with an effect size of 1σ.

Mann-Whitney Rank Sum

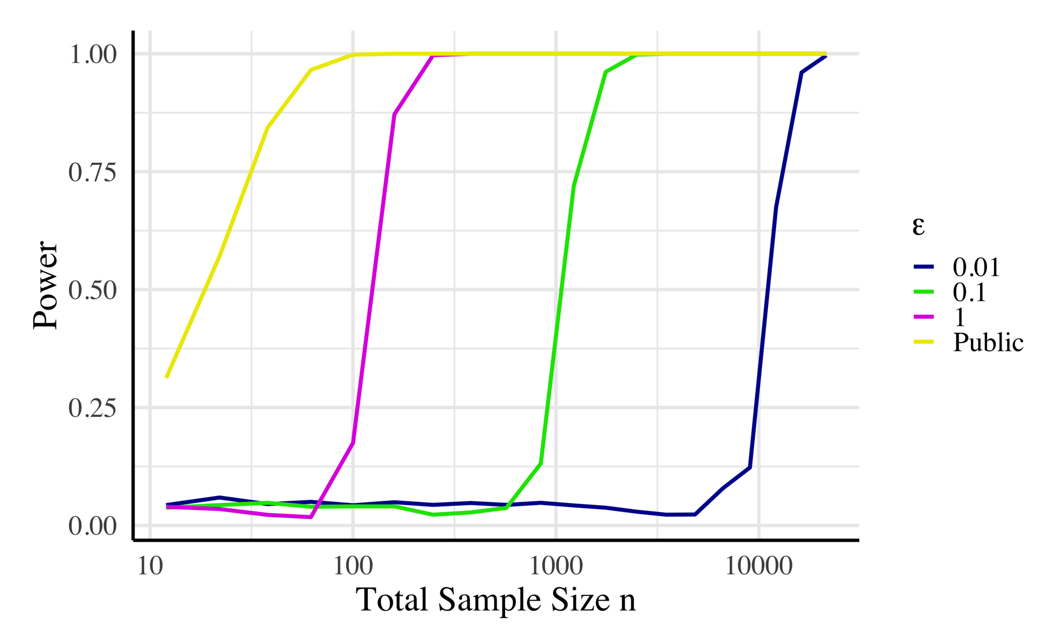
Setting Testing if two independent sets of data share the same pop. distribution.

Public statistic For each group $j \in \{1, 2\}$, define the rank sum, $R_j = \sum_{g_i = j} r_i$. The statistic $\mathcal{U}$ is:

$$\mathcal{U} = \min\{R_1 - \frac{n_1(n_1+1)}{2}, R_2 - \frac{n_2(n_2+1)}{2}\}$$

Our contribution The sensitivity of $\mathcal{U}$ depends on the group sizes, so we develop a two-stage private algorithm to first release the group sizes and then release $\mathcal{U}$ with Laplace noise. We use the normal approximation to simulate the reference distribution.

Results Our approach sets a benchmark for power in this class of tests, requiring $n \approx 10^{2.2}$ to achieve high power at $\epsilon = 1$.



Take Aways

- Customizing the traditional public statistics for the private setting can lead to dramatic improvements in power.
- Power curves are a useful metric by which to compare multiple statistics.
- In the private setting, rank-based statistics can out-perform Gaussian-based statistics, even when the assumptions of the normal methods are met.

This material is based upon work supported by the National Science Foundation under Grant No. SaTC-1817245 and the Richter Funds.

Kruskal-Wallis

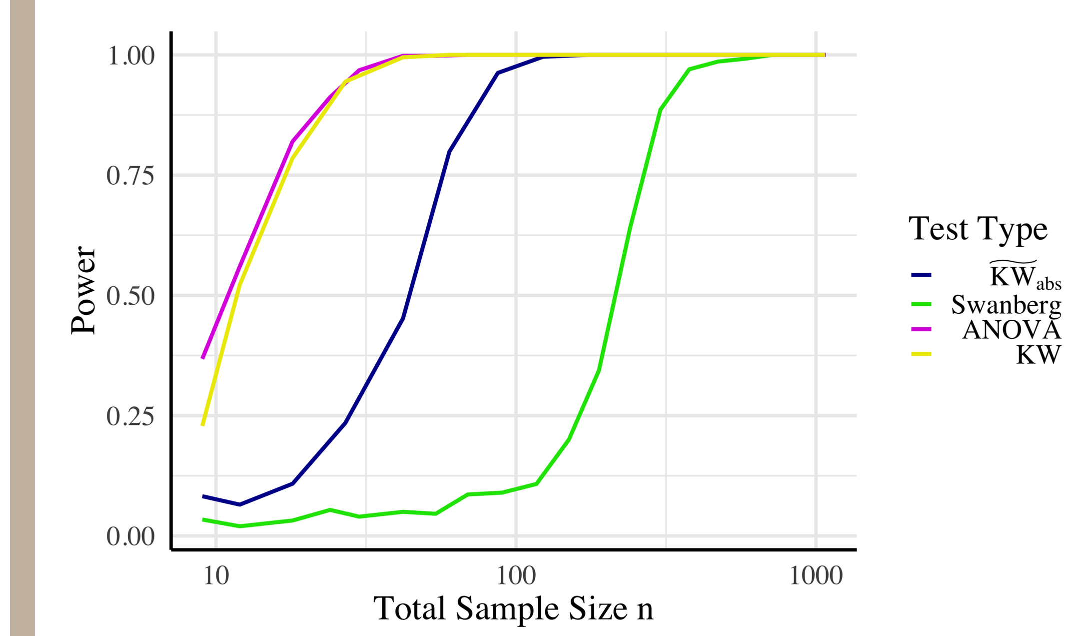
Setting Testing if ≥ 2 independent sets of data share the same pop. distribution.

Public statistic Let the size of each group $j \in \{1, \ldots, k\}$ be n_j, its mean rank be $\bar{r}_j$, and the mean of all ranks $\bar{r} = \frac{n+1}{2}$. The statistic $\mathcal{H}$ is:

$$\mathcal{H} = (n-1)\frac{\sum_{j=1}^{k} n_j(\bar{r}_j - \bar{r})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(r_{ij} - \bar{r})^2}$$

Our contribution We adapt $\mathcal{H}$ to use the L^1 instead of L^2 norm and privatize it with the Laplace mechanism. We simulate the exact reference distribution.

Results We find our test requires 20% as much data as the best existing method [1] to achieve the same power.



References

[1] M. Swanberg, I. Globus-Harris, I. Griffith, A. Groce, and A. Bray, *Improved differentially private analysis of variance*, preprint, (2018).

[2] C. Task and C. Clifton, *Differentially private significance testing on paired-sample data*, in Proceedings of the 2016 SIAM International Conference on Data Mining, SIAM, 2016, pp. 153–161.

REED COLLEGE