Hypothesis Testing for Differentially Private Linear Regression

Full Paper: https://arxiv.org/abs/2206.14449

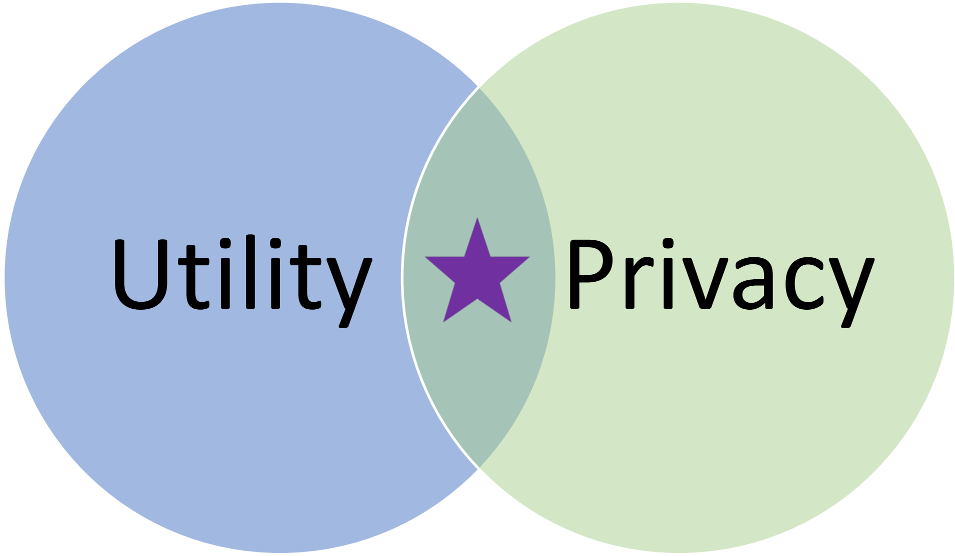
Daniel Alabi

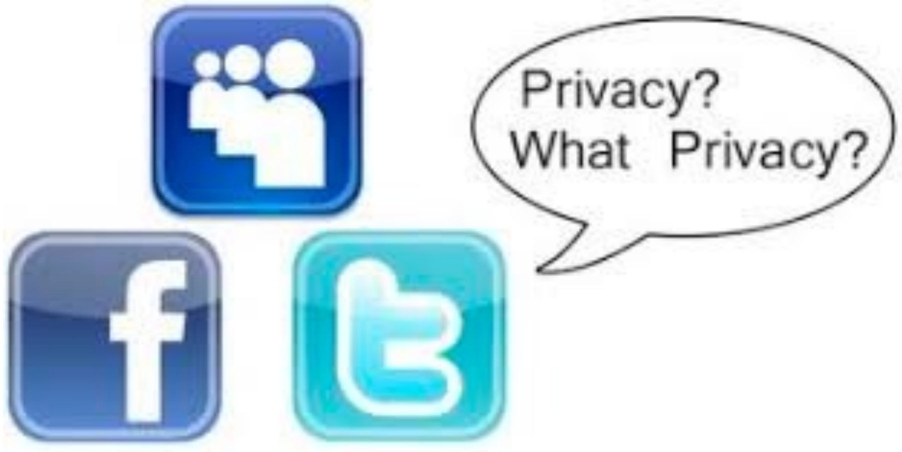
joint work with *Salil Vadhan*

The Privacy Problem

We have a dataset with sensitive information, such as:

1. Health records (e.g., reveals which disease a patient has)
2. Census data (e.g., reveals income range)
3. Social network activity (e.g., which pages you like)





Differential Privacy

Definition: pure and approximate

[Dwork-McSherry-Nissim-Smith ‘06]

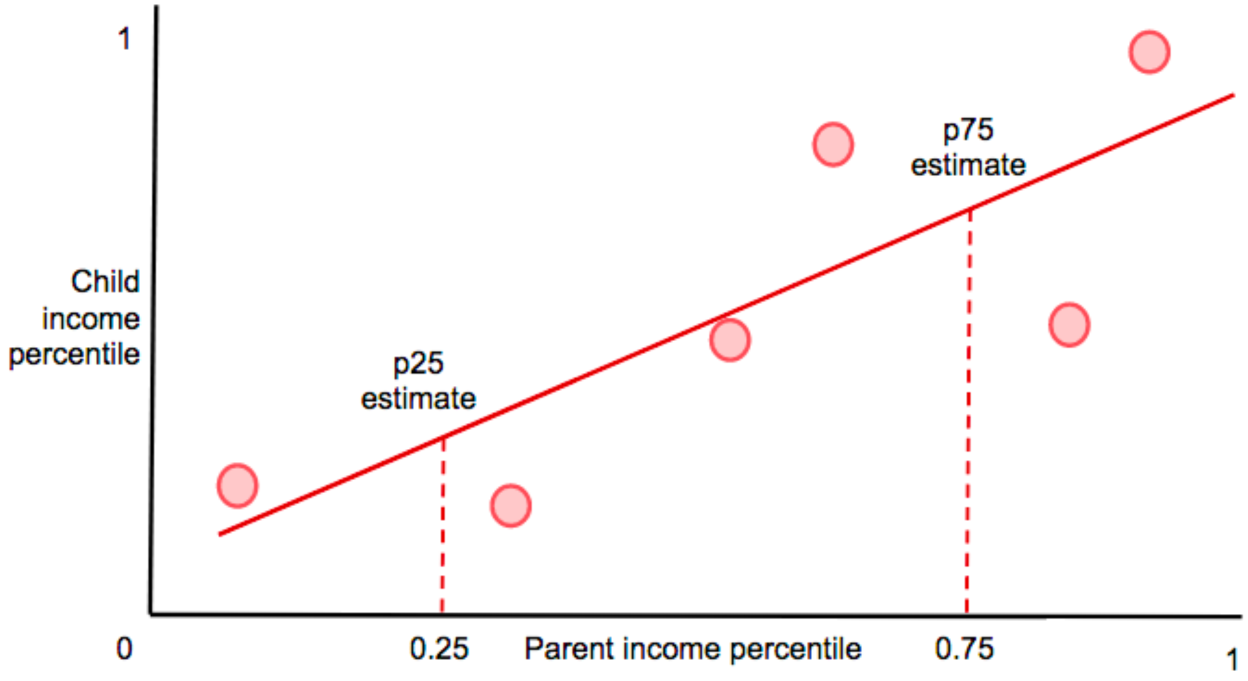
Other references:

Motivated from and based off of work in

[Dinur-Nissim ‘03, Dwork-Nissim ‘04, Blum-Dwork-McSherry-Nissim ‘05]

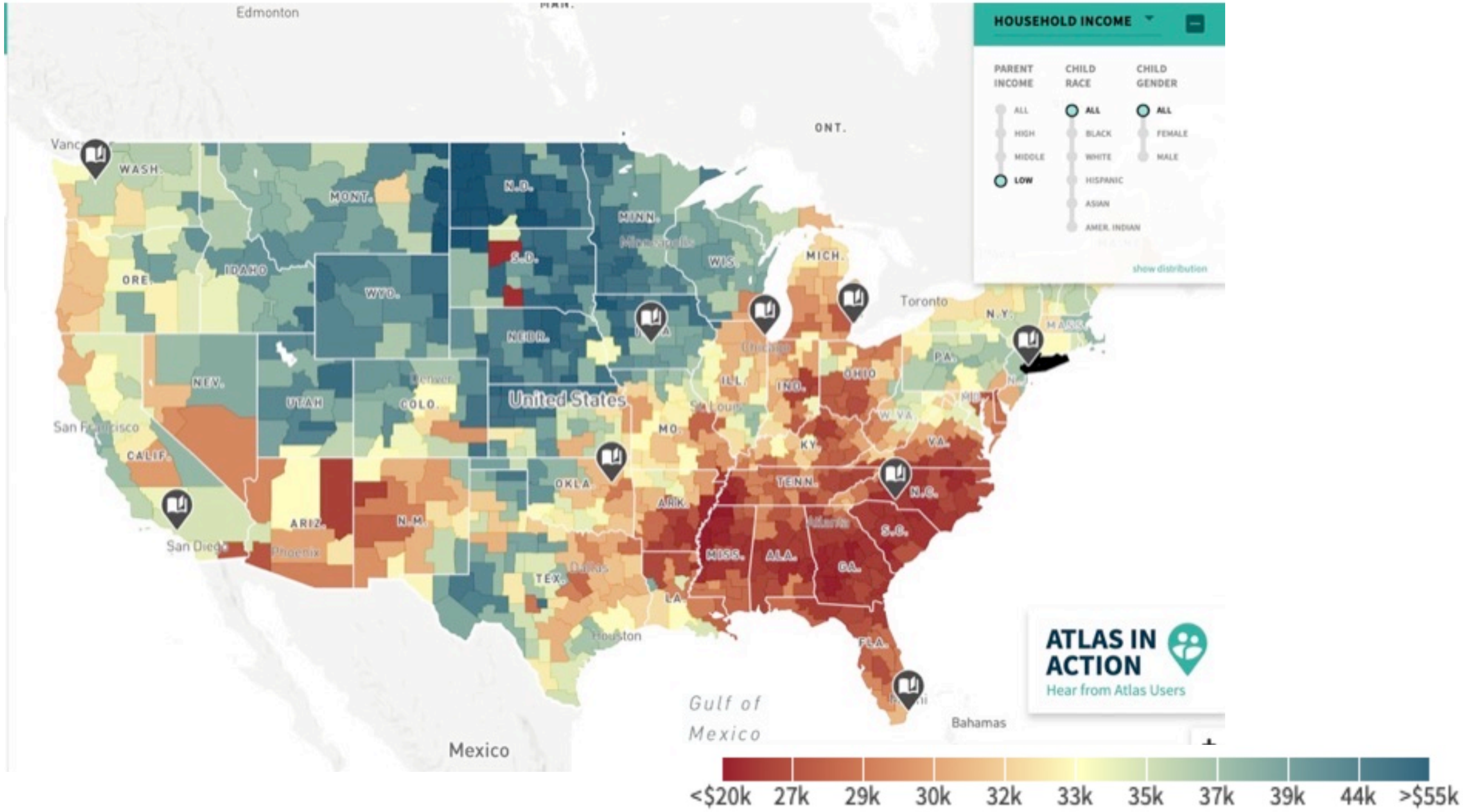
Hypothesis Testing in the General Linear Model

1) **Testing a Linear Relationship**: is the slope of the linear model equal to 0?



Hypothesis Testing in the General Linear Model

2) **Testing for Mixtures**: does the population consist of one or more sub-populations with different regression coefficients?



The General Linear Model

$$Y \sim \mathcal{N}(X\beta, \sigma_e^2 I_{n\times n})$$

1) $X \in \mathbb{R}^{n\times p}$

2) $\beta \in \mathbb{R}^{p}$ (e.g., $p = 2$ for simple linear regression)

For simple linear regression,

$\forall\, i \in [n],\; y_i = \beta_1 \cdot x_i + \beta_2 + e_i,\quad e_i$ are error terms

Hypothesis Testing in the General Linear Model

$$Y \sim \mathcal{N}(X\beta, \sigma_e^2 I_{n\times n})$$

1) H_0: $\beta \in \omega_0$, where ω_0 is a q-dimensional linear subspace of ω
2) H_1: $\beta \in \omega \setminus \omega_0$, where ω is an r-dimensional linear subspace

$$0 \le q < r$$

$$\hat{\beta}^N = \operatorname{argmin}_{z\in\omega_0} \|Xz - Y\|^2$$

$$\hat{\beta} = \operatorname{argmin}_{z\in\omega} \|Xz - Y\|^2$$

Hypothesis Testing in the General Linear Model

$$Y \sim \mathcal{N}(X\beta, \sigma_e^2 I_{n\times n})$$

$$\hat{\beta}^N = \text{ argmin}_{z\in\omega_0}\|Xz - Y\|^2, \qquad \hat{\beta} = \text{ argmin}_{z\in\omega}\|Xz - Y\|^2$$

$\hat{\theta}$: function of statistics of X, Y

$$\hat{E} = \frac{(X^T X)^{1/2}}{n^{1/2}}, \hat{F} = \frac{X^T Y}{n}, \hat{G} = \frac{Y^T Y}{n}$$

Re-write *F*-statistic as the generalized likelihood ratio test statistic:

$$T = T(\hat{\theta}) = \frac{n-r}{r-q} \cdot \frac{\|X\hat{\beta} - X\hat{\beta}^N\|^2}{\|Y - X\hat{\beta}\|^2} = \frac{n-r}{r-q} \cdot \frac{\|\sqrt{n}\hat{E}(\hat{\beta} - \hat{\beta}^N)\|^2}{n(\hat{\beta}^T\hat{E}^2\hat{\beta} - 2\hat{\beta}^T\hat{F} + \hat{G})}.$$

Linear Relationship Tester in the General Linear Model

Re-write *F*-statistic as the generalized likelihood ratio test statistic:

$$T = T(\hat{\theta}) = \frac{n-r}{r-q} \cdot \frac{\left\| X\hat{\beta} - X\hat{\beta}^N \right\|^2}{\left\| Y - X\hat{\beta} \right\|^2} = \frac{n-r}{r-q} \cdot \frac{\left\| \sqrt{n}\hat{E}(\hat{\beta} - \hat{\beta}^N) \right\|^2}{n(\hat{\beta}^T \hat{E}^2 \hat{\beta} - 2\hat{\beta}^T \hat{F} + \hat{G})}.$$

$\hat{\theta}$: function of statistics of X, Y

Add noise to the moments of X, Y :

- Make $\bar{x}, \overline{x^2}$ satisfy $\rho/5$-zCDP.
- Make $\bar{y}, \overline{y^2}$ satisfy $\rho/5$-zCDP.
- Make $\overline{xy}$ satisfy $\rho/5$-zCDP.

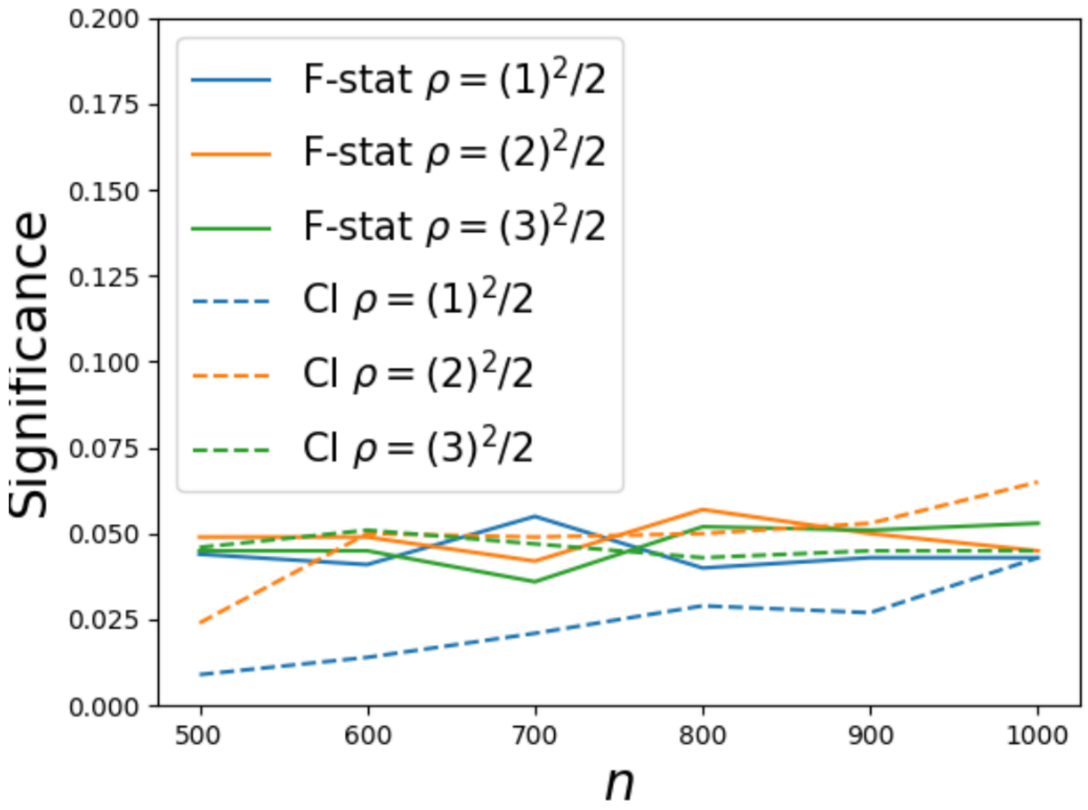
By composition, the entire procedure satisfies ρ-zCDP.

Combine with use of parametric bootstrap.

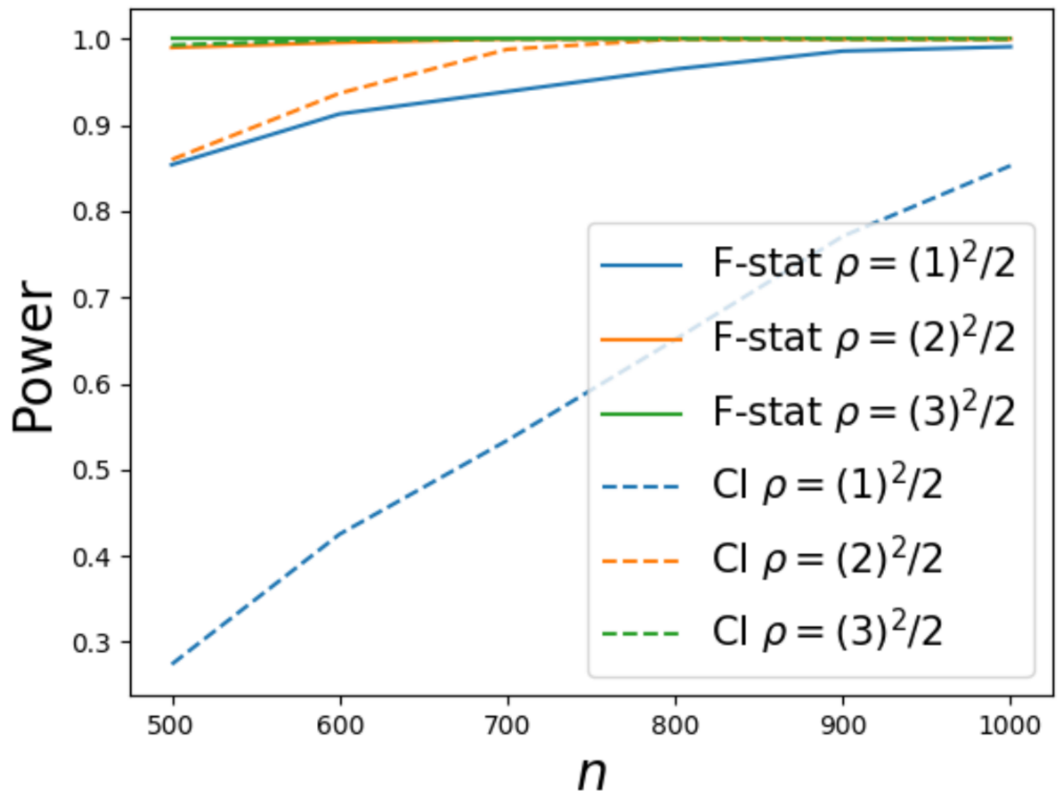
Empirical Performance of Previous Work (e.g., Ferrando, Wang, Sheldon, 2021)

- Computes differentially private confidence intervals
  - Estimates sufficient statistics as subroutine
  - Bootstrap parametric procedure
  - Has good coverage
  - Width of interval could be quite large especially for small privacy parameters
- Can convert to hypothesis test for testing a linear relationship
  - Compute confidence interval for slope: $[a, b]$
  - Reject null if $0 \notin [a, b]$
  - Fail to reject null if $0 \in [a, b]$
  - The larger the widths of the interval produced, the smaller the power of the test

Experimental Results for Linear Model Tester

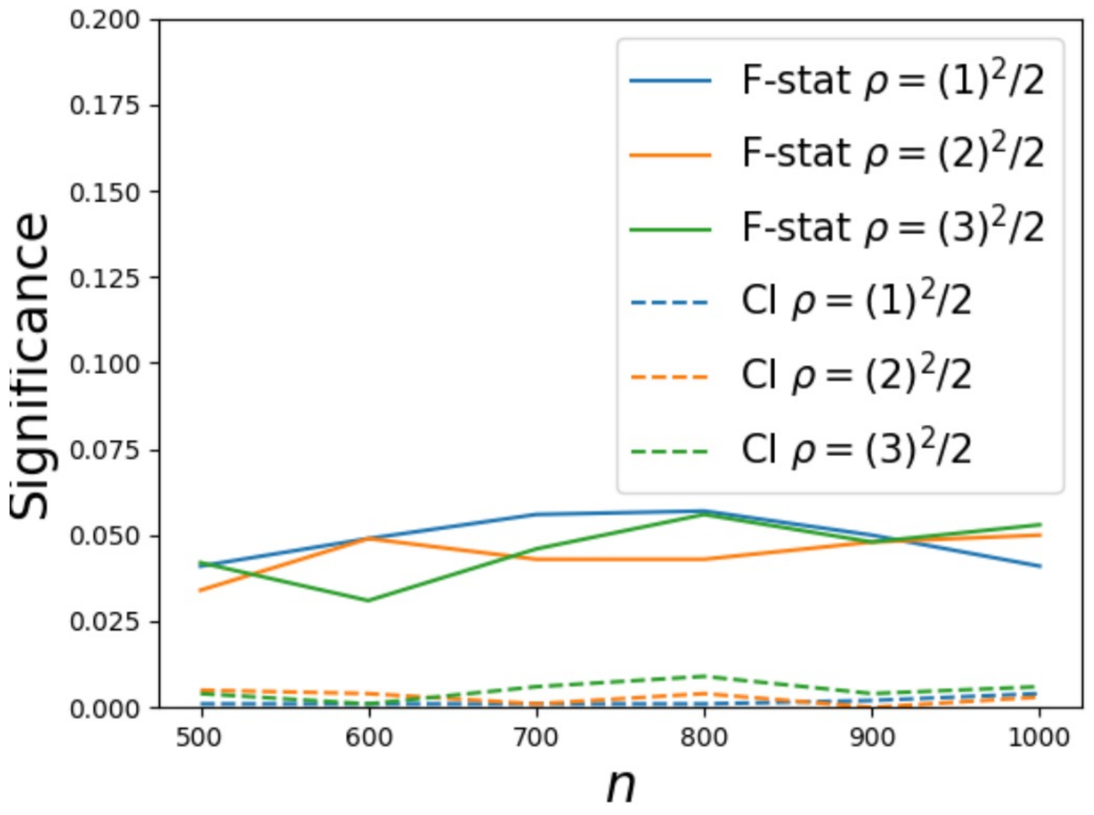


(a) Significance for F-statistic versus confidence interval approach. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 0 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

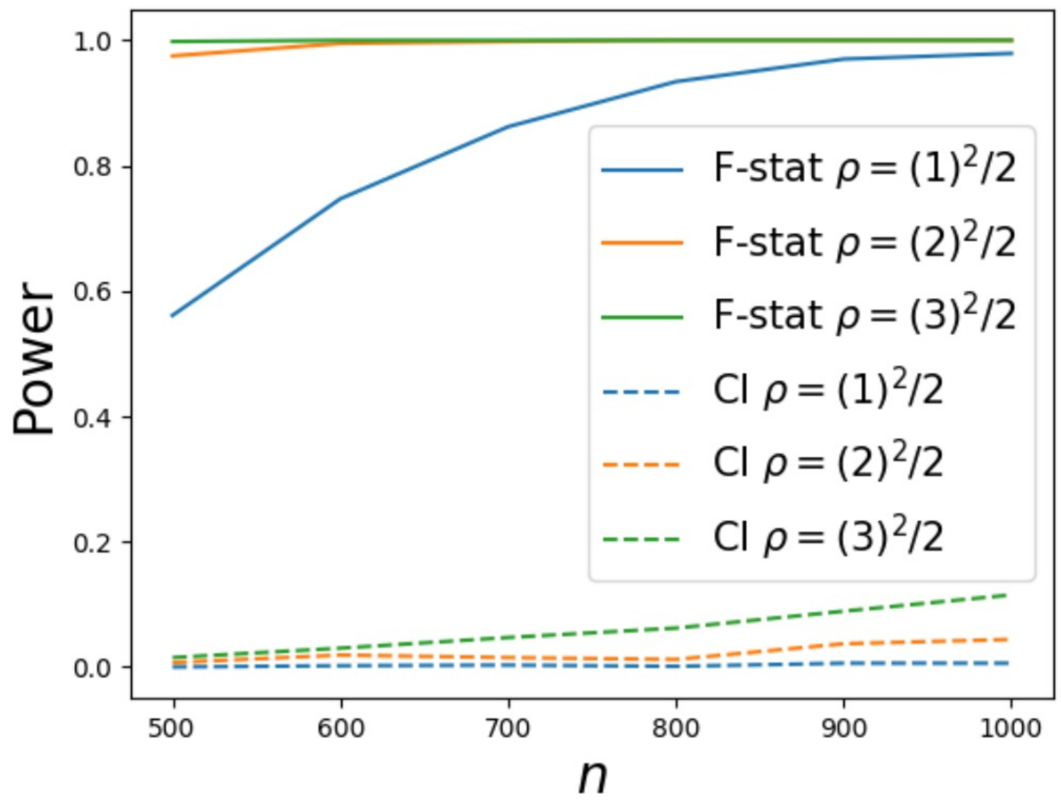


(b) Power for F-statistic versus confidence interval approach. $x_i \sim \mathcal{N}(0.5, 1)$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

Experimental Results for Linear Model Tester



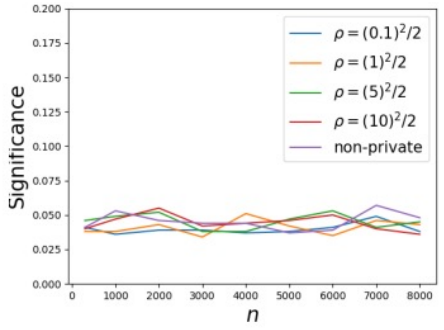
(a) Significance for F-statistic versus confidence interval approach. $x_i \sim \text{Unif}[0,1]$, $y_i \sim 0 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.



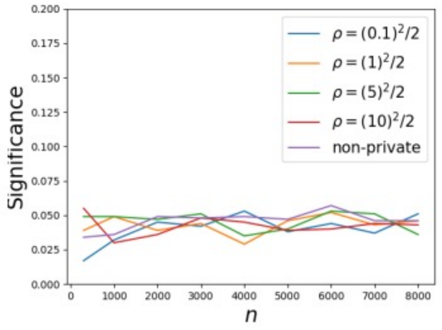
(b) Power for F-statistic versus confidence interval approach. $x_i \sim \text{Unif}[0,1]$, $y_i \sim 1 \cdot x_i + \mathcal{N}(0, 0.35^2)$. $\Delta = 2$.

Other Experimental Results in Full Paper

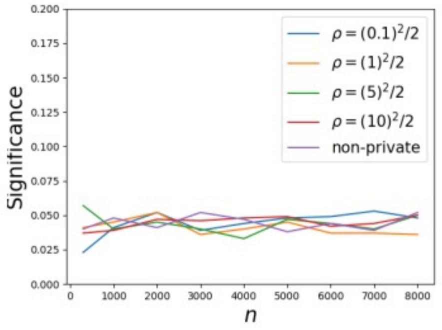
- Linear Model Tester with varying distributions on independent variable
- Mixture Model Tester using *F*-statistic
- Mixture Model Tester via Kruskal-Wallis (non-parametric)
- Results on real-world datasets:
  - UCI bike dataset
  - Opportunity Atlas



(a) Significance for testing a linear relationship. Normal Distribution on X.



(b) Significance for testing a linear relationship. Uniform Distribution on X.



(c) Significance for testing a linear relationship. Exponential Distribution on X.

Figure 4

Thanks! Any questions?
Some References on Differentially Private Uncertainty Quantification

- Differentially Private Linear Regression
  - Sheffet (2017): Tests for linear relationship; only "works" on very large data
  - Alabi, McMillan, Sarathy, Smith, Vadhan (2020): Point estimates for small-area analysis
  - Alabi, Vadhan (2022): hypothesis tests (mostly) based on *F*-statistic on small and large data
- General Differentially Private Hypothesis Testing
  - Gaboardi, Lim, Rogers, Vadhan (2017):
    - Goodness of fit for multinomial data
    - Independence tests for categorical random variables
  - Couch, Kazan, Shi, Bray, Groce (2019):
    - Rank-based nonparametric tests
    - Develop DP analogues of Kruskal-Wallis and Mann-Whitney signed-rank tests
  - Avella-Medina (2020) generalizes the *M*-estimator approach to differentially private statistical inference using an empirical notion of influence functions to calibrate the Gaussian mechanism
- Differentially Private Parametric Confidence Intervals
  - Ferrando, Wang, Sheldon (2021): Bootstrap for parametric inference