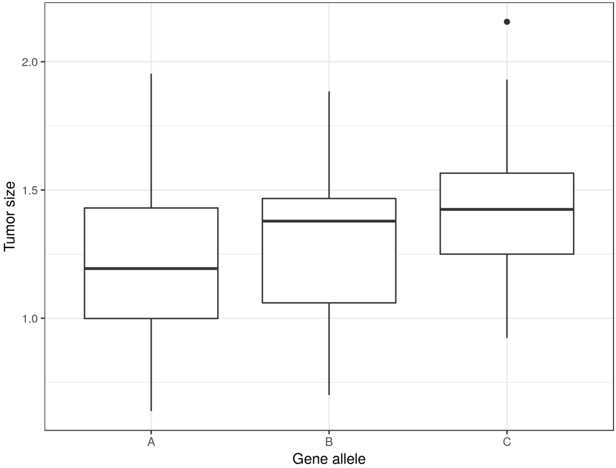
Improved Differentially Private Analysis of Variance

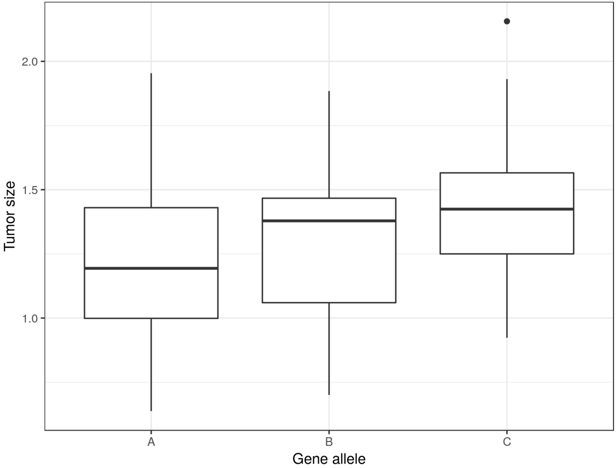
Marika Swanberg   Ira Globus-Harris   Iris Griffith
Anna Ritz   Andrew Bray   Adam Groce

Reed College

Observed Data, $n = 30$



Observed Data, $n = 30$



Are gene allele and tumor size dependent?

Metric for dependency

$$F\text{-test} = \frac{\text{Variation between groups}}{\text{In-group variation}}$$

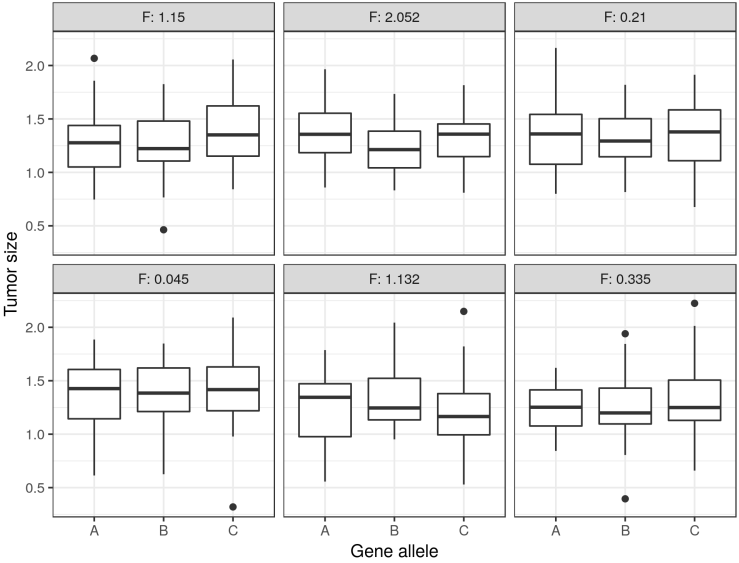
Metric for dependency

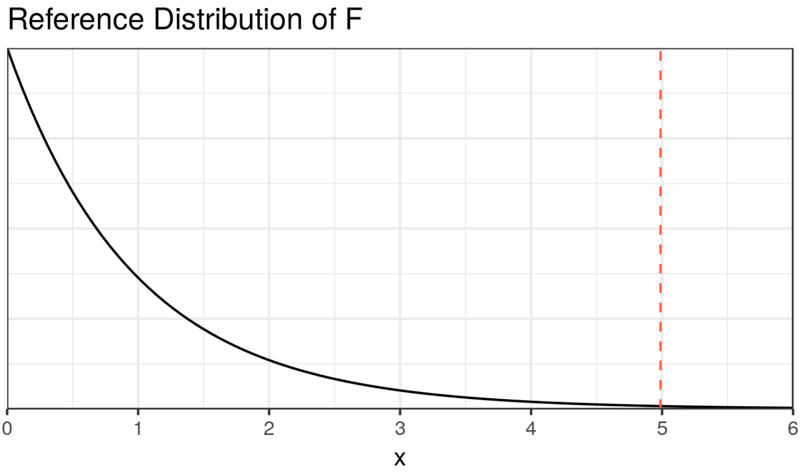
$$F\text{-test} = \frac{\text{Variation between groups}}{\text{In-group variation}}$$

$$SSA(D) = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2/(k-1)$$

$$SSE(D) = \sum_{i=1}^{n} (y_i - \bar{y}_{c_i})^2/(n-k)$$

Simulate Random Data





Now do we think our data supports independence of gene allele and tumor size?

Why is F-test optimal?

Why is F-test optimal?

High probability of indicating dependence when variables are dependent . . .

Why is F-test optimal?

High probability of indicating dependence when variables <u>are</u> dependent . . .

. . . even when dataset is small.

Why is F-test optimal?

High probability of indicating dependence when variables are dependent . . .

. . . even when dataset is small.

Definition (Power)

The **power** of a hypothesis test is the probability it rejects H_0. It depends on the alternate distribution H_A and n.

Why is F-test optimal?

High probability of indicating dependence when variables are dependent ...

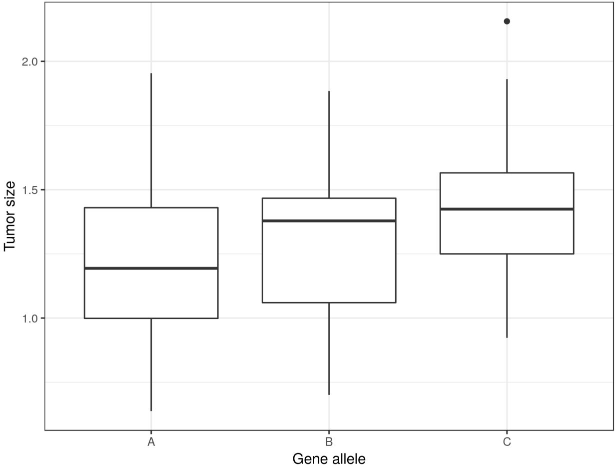
... even when dataset is small.

Definition (Power)

The **power** of a hypothesis test is the probability it rejects H_0. It depends on the alternate distribution H_A and n.

Goal of any test statistic is achieving high power*.

Observed Data, $n = 30$



What if we want to keep this data private?

Differential privacy [DMNS06]

Definition

Two databases are **neighboring** if they differ only in the data of one individual.

Differential privacy [DMNS06]

Definition

Two databases are **neighboring** if they differ only in the data of one individual.

Definition

A query f is ε-**differentially private** if for all neighboring databases D, D' and all output sets S

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

Properties of differential privacy [DMNS06]

Theorem (Post-processing)

If f is ε-differentially private then for any (randomized) function g, then if $h(D) = g(f(D)$, h is also ε-differentially private.

Properties of differential privacy [DMNS06]

Theorem (Post-processing)

If f is ε-differentially private then for any (randomized) function g, then if $h(D) = g(f(D)$, h is also ε-differentially private.

Theorem (Composition)

If f is ε_1-differentially private and g is ε_2-differentially private then if $h(D) = (g(D), f(D))$, h is $(\varepsilon_1 + \varepsilon_2)$-differentially private.

Laplace mechanism

Definition (Sensitivity)

The sensitivity Δf of a deterministic, real-valued function f on databases is the maximum over all pairs of neighboring D, D' of $|f(D) - f(D')|$.

Laplace mechanism

Definition (Sensitivity)

The sensitivity Δf of a deterministic, real-valued function f on databases is the maximum over all pairs of neighboring D, D' of $|f(D) - f(D')|$.

Theorem (Laplace Mechanism)

Given any deterministic, real-valued function f on databases, define $\widehat{f}$ as

$$\widehat{f}(D) = f(D) + Y,$$

where $Y \leftarrow \mathrm{Lap}(\Delta f/\varepsilon)$. The Laplace mechanism is ε-differentially private.

Related Works

Other work on private hypothesis testing:

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]
- Chi-squared test (difference of discrete distributions) [VS09, FSU11, JS13, USF13, WLK15, GLRV16, RK17]

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]
- Chi-squared test (difference of discrete distributions) [VS09, FSU11, JS13, USF13, WLK15, GLRV16, RK17]
- Other tests:

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]
- Chi-squared test (difference of discrete distributions) [VS09, FSU11, JS13, USF13, WLK15, GLRV16, RK17]
- Other tests:
  - Binomial data [AS18] (Proven optimal!)

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]
- Chi-squared test (difference of discrete distributions) [VS09, FSU11, JS13, USF13, WLK15, GLRV16, RK17]
- Other tests:
  - Binomial data [AS18] (Proven optimal!)
  - Difference of two means [OHK15, DNLI18]

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]
- Chi-squared test (difference of discrete distributions) [VS09, FSU11, JS13, USF13, WLK15, GLRV16, RK17]
- Other tests:
  - Binomial data [AS18] (Proven optimal!)
  - Difference of two means [OHK15, DNLI18]
  - Linear regression [BRMC17, Sheffet17]

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]
- Chi-squared test (difference of discrete distributions) [VS09, FSU11, JS13, USF13, WLK15, GLRV16, RK17]
- Other tests:
    - Binomial data [AS18] (Proven optimal!)
    - Difference of two means [OHK15, DNLI18]
    - Linear regression [BRMC17, Sheffet17]

Earlier work is often missing:

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]
- Chi-squared test (difference of discrete distributions) [VS09, FSU11, JS13, USF13, WLK15, GLRV16, RK17]
- Other tests:
  - Binomial data [AS18] (Proven optimal!)
  - Difference of two means [OHK15, DNLI18]
  - Linear regression [BRMC17, Sheffet17]

Earlier work is often missing:

- Rigorous p-value computations

Related Works

Other work on private hypothesis testing:

- Asymptotic analysis [WZ10, Smith11, CKMSU19]
- Chi-squared test (difference of discrete distributions) [VS09, FSU11, JS13, USF13, WLK15, GLRV16, RK17]
- Other tests:
  - Binomial data [AS18] (Proven optimal!)
  - Difference of two means [OHK15, DNLI18]
  - Linear regression [BRMC17, Sheffet17]

Earlier work is often missing:

- Rigorous p-value computations
- Power analysis

Private F-statistic [CBRG18]

Assume data is on the $[0, 1]$ interval.

Private F-statistic [CBRG18]

Assume data is on the $[0, 1]$ interval.

Theorem

SSE has sensitivity bounded by 7.

Private F-statistic [CBRG18]

Assume data is on the $[0, 1]$ interval.

Theorem

SSE has sensitivity bounded by 7.

$$\widehat{SSE}(D) = SSE(D) + \text{Lap}(7/\varepsilon)$$

Private F-statistic [CBRG18]

Assume data is on the $[0,1]$ interval.

Theorem

SSE has sensitivity bounded by 7.

$$\widehat{SSE}(D) = SSE(D) + \mathrm{Lap}(7/\varepsilon)$$

Theorem

SSA has sensitivity bounded by $9 + 5/n$.

Private F-statistic [CBRG18]

Assume data is on the $[0, 1]$ interval.

Theorem

SSE has sensitivity bounded by 7.

$$\widehat{SSE}(D) = SSE(D) + \mathrm{Lap}(7/\varepsilon)$$

Theorem

SSA has sensitivity bounded by $9 + 5/n$.

$$\widehat{SSA}(D) = SSA(D) + \mathrm{Lap}\left(\frac{9 + 5/n}{\varepsilon}\right)$$

Private ANOVA [CBRG18]

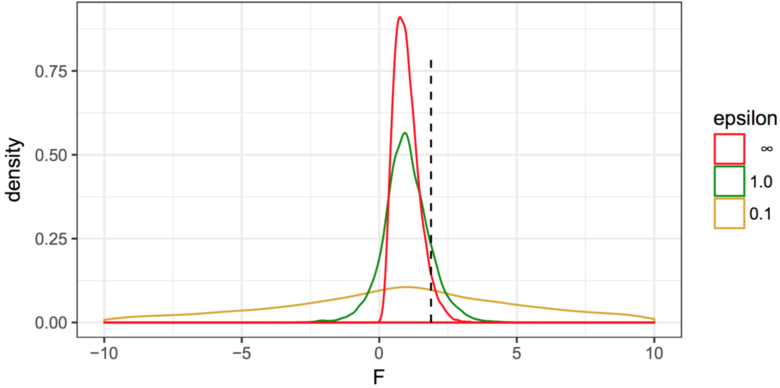
$$\widehat{F}(D) = \frac{\widehat{SSA}(D)/(k-1)}{\widehat{SSE}(D)/(n-k)}$$

Private ANOVA [CBRG18]

$$\widehat{F}(D) = \frac{\widehat{SSA}(D)/(k-1)}{\widehat{SSE}(D)/(n-k)}$$

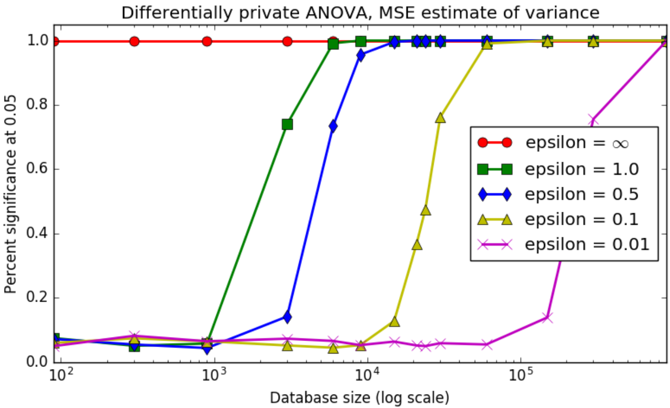
Problem: What is the reference distribution now?

Private ANOVA [CBRG18]



Public reference distribution gives inaccurate p-values.

Private ANOVA [CBRG18]



Definition (Power)

The **power** of a hypothesis test is the probability it rejects H_0. It depends on the alternate distribution H_A and n.

Improving ANOVA [SHGRGB19]

Are there other ways of measuring "dispersion" (analogous to variance)?

Improving ANOVA [SHGRGB19]

Are there other ways of measuring "dispersion" (analogous to variance)?

$$(x_i - \bar{x})^2, \quad |x_i - \bar{x}|, \text{ or maybe } |x_i - \bar{x}|^?$$

Improving ANOVA [SHGRGB19]

Are there other ways of measuring "dispersion" (analogous to variance)?

$$(x_i - \bar{x})^2, \quad |x_i - \bar{x}|, \text{ or maybe } |x_i - \bar{x}|^?$$

$$SSA(D) = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 \Longrightarrow SA(D) = \sum_{j=1}^{k} n_j|\bar{y}_j - \bar{y}|$$

$$SSE(D) = \sum_{i=1}^{n} (y_i - \bar{y}_{c_i})^2 \Longrightarrow SE(D) = \sum_{i=1}^{n} |y_i - \bar{y}_{c_i}|$$

$$F(D) = \frac{SSA(D)/(k-1)}{SSE(D)/(n-k)} \Longrightarrow F_1(D) = \frac{SA(D)/(k-1)}{SE(D)/(n-k)}$$

Improving ANOVA [SHGRGB19]

Are there other ways of measuring "dispersion" (analogous to variance)?

$$(x_i - \bar{x})^2, \quad |x_i - \bar{x}|, \text{ or maybe } |x_i - \bar{x}|^?$$

$$SSA(D) = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 \Longrightarrow SA(D) = \sum_{j=1}^{k} n_j|\bar{y}_j - \bar{y}|$$

$$SSE(D) = \sum_{i=1}^{n}(y_i - \bar{y}_{c_i})^2 \Longrightarrow SE(D) = \sum_{i=1}^{n} |y_i - \bar{y}_{c_i}|$$

$$F(D) = \frac{SSA(D)/(k-1)}{SSE(D)/(n-k)} \Longrightarrow F_1(D) = \frac{SA(D)/(k-1)}{SE(D)/(n-k)}$$

The new F_1 statistic has:

Improving ANOVA [SHGRGB19]

Are there other ways of measuring "dispersion" (analogous to variance)?

$$(x_i - \bar{x})^2, \quad |x_i - \bar{x}|, \text{ or maybe } |x_i - \bar{x}|^?$$

$$SSA(D) = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 \Longrightarrow SA(D) = \sum_{j=1}^{k} n_j|\bar{y}_j - \bar{y}|$$

$$SSE(D) = \sum_{i=1}^{n} (y_i - \bar{y}_{c_i})^2 \Longrightarrow SE(D) = \sum_{i=1}^{n} |y_i - \bar{y}_{c_i}|$$

$$F(D) = \frac{SSA(D)/(k-1)}{SSE(D)/(n-k)} \Longrightarrow F_1(D) = \frac{SA(D)/(k-1)}{SE(D)/(n-k)}$$

The new F_1 statistic has:

- Lower sensitivity (3 for SE, 4 for SA)

Improving ANOVA [SHGRGB19]

Are there other ways of measuring "dispersion" (analogous to variance)?

$$(x_i - \bar{x})^2, \quad |x_i - \bar{x}|, \text{ or maybe } |x_i - \bar{x}|^?$$

$$SSA(D) = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 \Longrightarrow SA(D) = \sum_{j=1}^{k} n_j|\bar{y}_j - \bar{y}|$$

$$SSE(D) = \sum_{i=1}^{n} (y_i - \bar{y}_{c_i})^2 \Longrightarrow SE(D) = \sum_{i=1}^{n} |y_i - \bar{y}_{c_i}|$$

$$F(D) = \frac{SSA(D)/(k-1)}{SSE(D)/(n-k)} \Longrightarrow F_1(D) = \frac{SA(D)/(k-1)}{SE(D)/(n-k)}$$

The new F_1 statistic has:

- Lower sensitivity (3 for SE, 4 for SA)
- Much higher typical value

Improving ANOVA [SHGRGB19]

Computing F_1 privately

Improving ANOVA [SHGRGB19]

Computing F_1 privately

1. $\widehat{SA} = SA + \mathrm{Lap}(4/\rho\varepsilon)$
2. $\widehat{SE} = SA + \mathrm{Lap}(3/(1-\rho)\varepsilon)$

Improving ANOVA [SHGRGB19]

Computing F_1 privately

1. $\widehat{SA} = SA + \text{Lap}(4/\rho\varepsilon)$
2. $\widehat{SE} = SA + \text{Lap}(3/(1-\rho)\varepsilon)$
3. $\widehat{F}_1 = \frac{\widehat{SA}/(k-1)}{\widehat{SE}/(n-k)}$

Improving ANOVA [SHGRGB19]

Computing F_1 privately

1. $\widehat{SA} = SA + \text{Lap}(4/\rho\varepsilon)$
2. $\widehat{SE} = SA + \text{Lap}(3/(1-\rho)\varepsilon)$
3. $\widehat{F}_1 = \frac{\widehat{SA}/(k-1)}{\widehat{SE}/(n-k)}$

Where ρ denotes epsilon allocation

- Empirically checked: optimal $\rho \approx 0.7$

Improving ANOVA [SHGRGB19]

Computing F_1 privately

1. $\widehat{SA} = SA + \text{Lap}(4/\rho\varepsilon)$
2. $\widehat{SE} = SA + \text{Lap}(3/(1-\rho)\varepsilon)$
3. $\widehat{F}_1 = \frac{\widehat{SA}/(k-1)}{\widehat{SE}/(n-k)}$

Where ρ denotes epsilon allocation

- Empirically checked: optimal $\rho \approx 0.7$

Computing accurate p-values

Improving ANOVA [SHGRGB19]

Computing F_1 privately

1. $\widehat{SA} = SA + \text{Lap}(4/\rho\varepsilon)$
2. $\widehat{SE} = SA + \text{Lap}(3/(1-\rho)\varepsilon)$
3. $\widehat{F}_1 = \frac{\widehat{SA}/(k-1)}{\widehat{SE}/(n-k)}$

Where ρ denotes epsilon allocation

- Empirically checked: optimal $\rho \approx 0.7$

Computing accurate p-values

- Simulate reference distribution of $\widehat{F}_1$

Improving ANOVA [SHGRGB19]

Computing F_1 privately

1. $\widehat{SA} = SA + \text{Lap}(4/\rho\varepsilon)$
2. $\widehat{SE} = SA + \text{Lap}(3/(1-\rho)\varepsilon)$
3. $\widehat{F}_1 = \frac{\widehat{SA}/(k-1)}{\widehat{SE}/(n-k)}$

Where ρ denotes epsilon allocation

- Empirically checked: optimal $\rho \approx 0.7$

Computing accurate p-values

- Simulate reference distribution of $\widehat{F}_1$
- Problem: reference distribution depends on σ

Improving ANOVA [SHGRGB19]

Need private estimate of σ

Improving ANOVA [SHGRGB19]

Need private estimate of σ

- Allocate some of epsilon budget?

Improving ANOVA [SHGRGB19]

Need private estimate of σ

- Allocate some of epsilon budget?
- Solution: derive an unbiased estimator for σ

Improving ANOVA [SHGRGB19]

Need private estimate of σ

- Allocate some of epsilon budget?
- Solution: derive an unbiased estimator for σ

$$\hat{\sigma} = \sqrt{\pi/2} \cdot \frac{\widehat{SE}}{(N-k)}$$

Improving ANOVA [SHGRGB19]

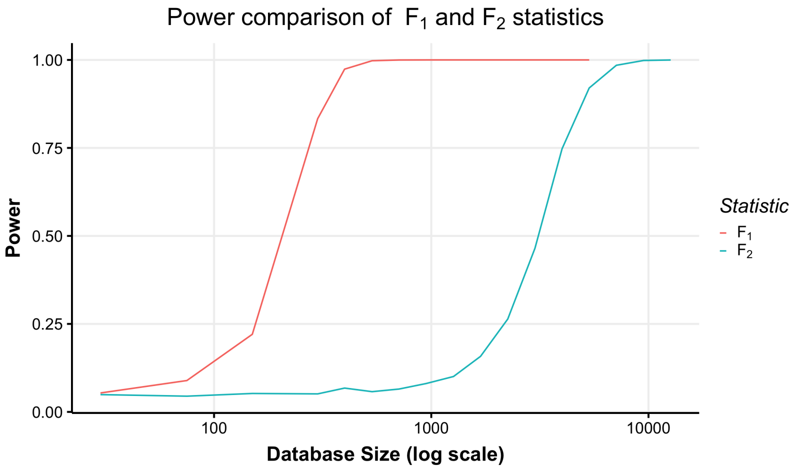
Need private estimate of σ

- Allocate some of epsilon budget?
- Solution: derive an unbiased estimator for σ

$$\hat{\sigma} = \sqrt{\pi/2} \cdot \frac{\widehat{SE}}{(N-k)}$$

Proceed with simulation

Empirically verified to have valid p-values.


Power comparison of F_1 and F_2 statistics
1.00
0.75
0.50
0.25
0.00
Power
100
1000
10000
Database Size (log scale)
Statistic
F_1
F_2

Further Optimization

Are there other ways of measuring "dispersion" (analogous to variance)?

$$(x_i - \bar{x})^2, \quad |x_i - \bar{x}|, \text{ or maybe } |x_i - \bar{x}|^?$$

Further Optimization

Are there other ways of measuring "dispersion" (analogous to variance)?

$$(x_i - \bar{x})^2, \quad |x_i - \bar{x}|, \text{ or maybe } |x_i - \bar{x}|^?$$

Perhaps a different exponent?

Further Optimization

Are there other ways of measuring "dispersion" (analogous to variance)?

$$(x_i - \bar{x})^2, \quad |x_i - \bar{x}|, \text{ or maybe } |x_i - \bar{x}|^?$$

Perhaps a different exponent? Turns out 1 is optimal.

Conclusion

In the private framework . . .

- Not limited to closed-form test statistics
- Room for massive power gains
- Optimal public test $\not\Rightarrow$ optimal private test

. . . all of statistics is fair game.

New developments

[CKSBG19] gained another order of magnitude improvement. See *Differentially Private Nonparametric Hypothesis Testing* at CCS '19.

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.

Thank you