Differentially Private Chi-squared Test by Unit Circle Mechanism

Kazuya Kakizaki*[1] Kazuto Fukuchi *[1] Jun Sakuma *[1,2,3]

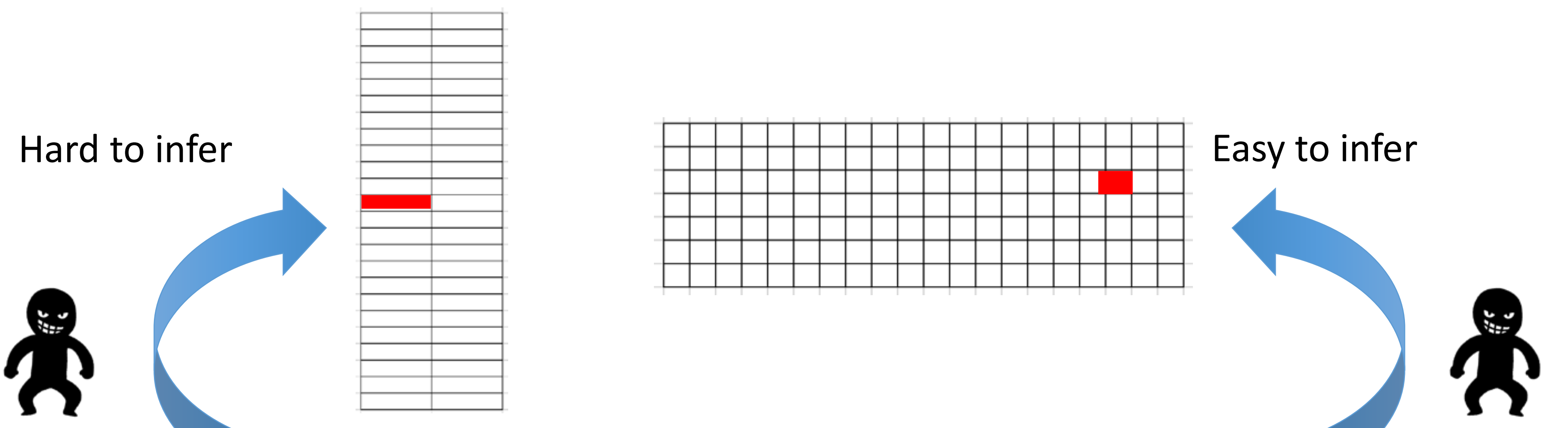
*[1]U. Tsukuba, *[2]JST CREST, *[3]RIKEN AIP

(to appear in ICML 2017)

https://bigdata.nii.ac.jp/eratokansyasai4/

Estimating samples from statistics

- Release of statistics is believed not to reveal information on each sample
- Samples might be inferred if the statistic dimension is high



$$f(X) \qquad f_1(X), f_2(X), \ldots f_p(X)$$

$$n \gg p \qquad n \ll p$$

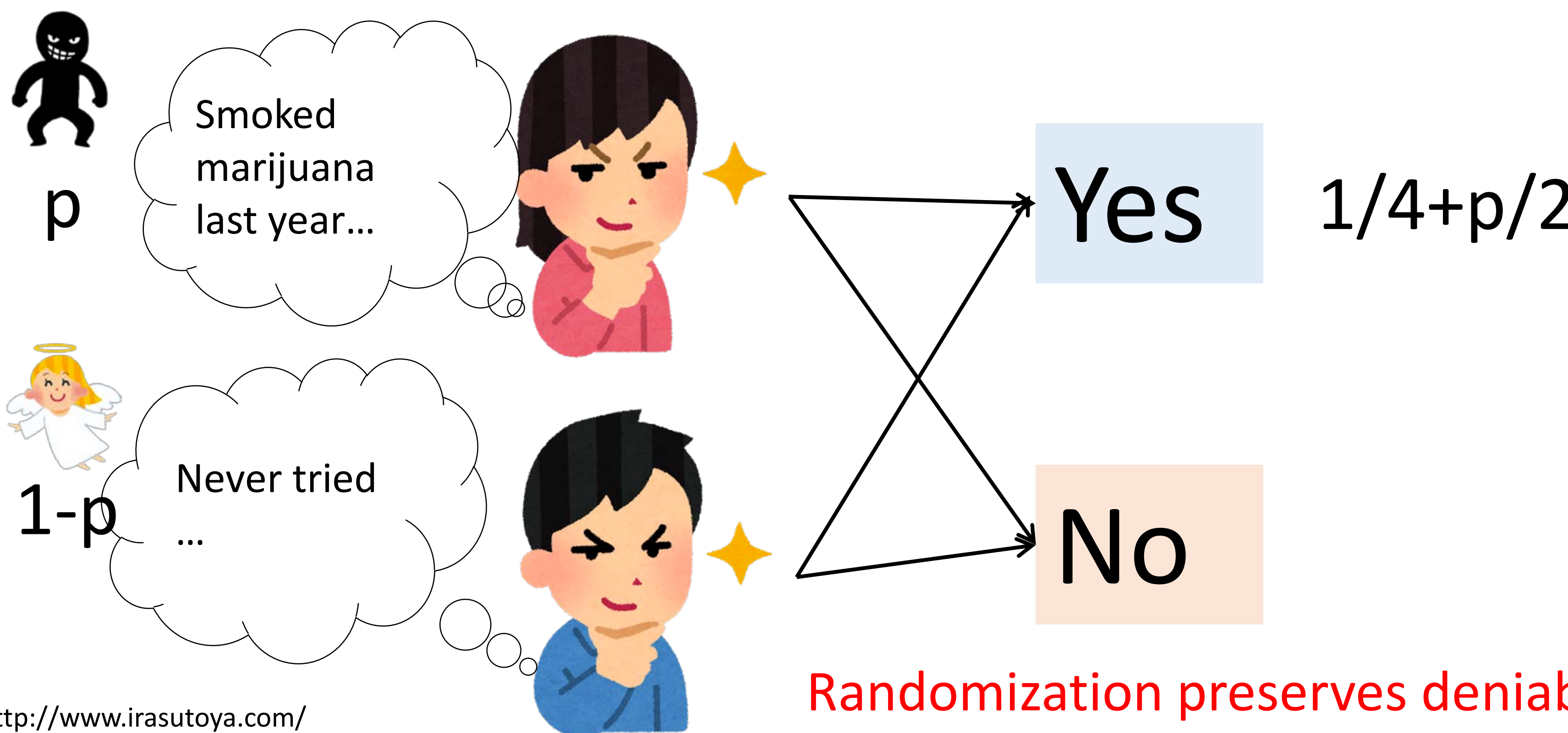
http://www.irasutoya.com/

GWAS case

- Finding disease related SNPs with chi-squared test
  - #samples: $\sim 10^4$
  - #dimension: $\sim 10^6$
- Privacy invasion caused by release of test statistics
  - Patient's disease status could be inferred from aggregate statistics collected for GWAS [Homer+ 2008]
  - NIH decided to stop releasing GWAS-related statistics publicly
- How can we release statistics securely?

Releasing statistics securely

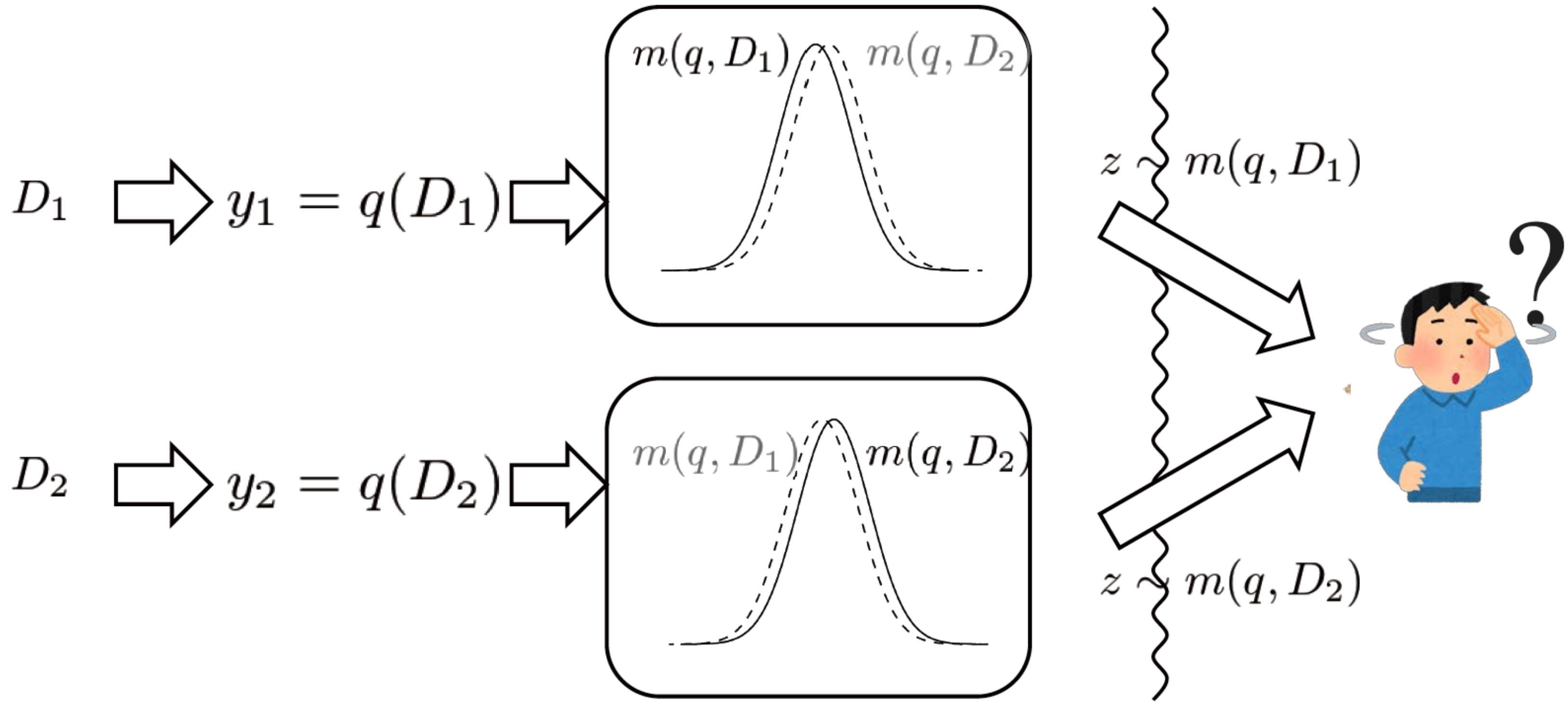
- Plausible deniability and randomized response
  - Question “Have you ever tried marijuana in the past?”
  - Flip a coin. If tail, answer honestly. Else, flip again and answer honestly if tail. Else, answer randomly.



Randomization preserves deniability

http://www.irasutoya.com/

Differential Privacy (DP) [Dwork+ 2006]



Definition 1 (ϵ-DP (Dwork et al., 2006)). *Mechanism* $\mathcal{M} : \mathcal{S} \to \mathcal{Y}$ *provides* ϵ*-DP if, for any* $S \sim S'$ *and* $Y \subseteq \mathcal{Y}$,

$$\Pr[\mathcal{M}(S) \in Y] \leq \exp(\epsilon) \Pr[\mathcal{M}(x') \in Y].$$

http://www.irasutoya.com/

χ^2 test for independence

	Smoker	Non-smoker
Lung cancer	a	b
No lung cancer	c	d

- Chi-squared statistic
  - Is smoking related to lung cancer?
  - If $\chi^2(S) > \tau_\alpha$, associated

Related works

- To achieve DP, the test statistic needs to be randomized in some sense

1. Output perturbation [Fienberg+ 2011, Yu+ 2014, Wang+ 2015]
   - Randomize test statistics
   - Type-I error is controllable
   - High type-I error w.r.t. sample size N
   - Type-II error/FWER is not controllable
2. Input perturbation [Johnson+ 2013]
   - Randomize counts
   - Type-I error is controllable
   - Type-II error/FWER is not controllable

Contributions

1. Investigate the type-II error of DP mechanisms analytically
2. A novel DP mechanism with $O(\exp(-\sqrt{N}))$ type-II error
3. A novel DP mechanism that can control the family-wise error rate (FWER)

Contributions

1. Investigate the type-II error of DP mechanisms analytically
2. A novel DP mechanism with $O(\exp(-\sqrt{N}))$ type-II error
3. A novel DP mechanism that can control the family-wise error rate (FWER)

Type-II error of DP chi-squared test

Theorem (Upper bound of type-II error of DP chi-squared test)

For any $\gamma > 0$, the upper bound of the type-II error of DP chi-squared test mechanism is

$$\Pr[M(S, \hat{\tau}_\alpha) = \mathrm{acc} | H_1 \text{ is true}]$$
$$\leq \sup_{P \in \mathcal{P}} \Big\{ \underbrace{\Pr_{S \sim P}[M(S, \hat{\tau}_\alpha) = \mathrm{acc} | \chi^2(S) > \hat{\tau}_\alpha + \gamma]}_{①} + \underbrace{\beta_{\hat{\tau}_\alpha + \gamma}}_{②} \Big\}$$

β_{τ_α} :type-II error when one use threshold τ_α

$\mathcal{P} = \{P : H_1 \text{is true}\}$:set of distributions of sample sets

$\hat{\tau}_\alpha$: Threshold for mechanism M that is determined so that the type-I error of M becomes α

- ① measures how often the mechanism wrongly accepts H_0 (γ error)
- ② is the type-II error of non-private test with threshold $\hat{\tau}_\alpha + \gamma$

A mechanism M with lower ① and ② achieves greater power

Power analysis

- Output perturbation
  - The γ error is upper-bounded by $\frac{1}{2}\exp\left(\frac{-\gamma\epsilon}{\Delta}\right)$
  - Not decreases w.r.t. N
- Input perturbation
  - The γ error cannot be appropriately derived
- Can we design a randomization mechanism in which the gamma error is decreasing w.r.t. N?

Contributions

1. Investigate the type-II error of DP mechanisms analytically
2. A novel DP mechanism with $O(\exp(-\sqrt{N}))$ type-II error
3. A novel DP mechanism that can control the family-wise error rate (FWER)

Table 1. Contingency table T of two binary varaibles

	$X_1 = 1$	$X_1 = 0$	
$X_0 = 1$	c_{11}	c_{10}	M_1
$X_0 = 0$	c_{01}	c_{00}	M_0
	N_1	N_0	N

Test statistic

$$\chi^2(S) = \frac{(c_{11}c_{00} - c_{10}c_{01})^2 N}{(c_{11} + c_{10})(c_{11} + c_{01})(c_{10} + c_{00})(c_{01} + c_{00})}.$$

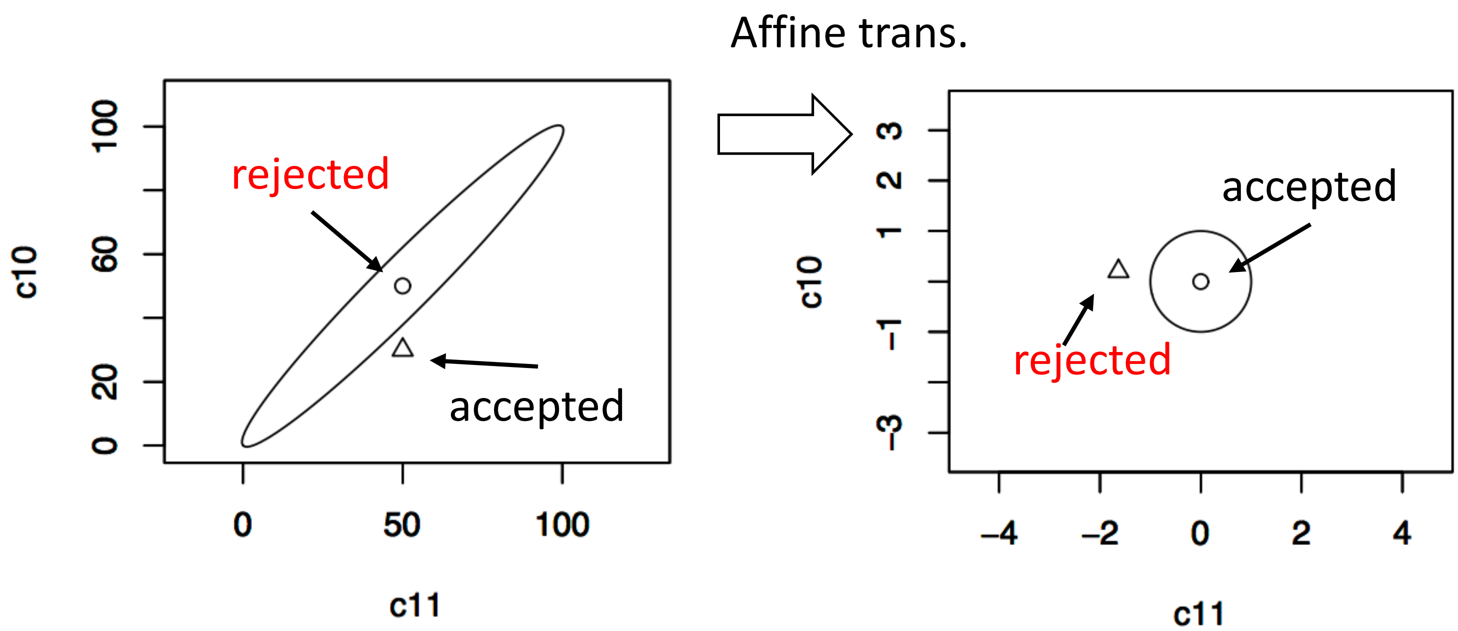
The test statistic is a function of c_{11} and c_{10}

Geometrical interpretation of statisitcs

- Chi-square test statistic $\chi^2(c_{11}, c_{10}) = \tau_\alpha.$

$$\iff Ac_{11}^2 + Bc_{10}^2 \quad + \quad 2Cc_{11}c_{10} + D(c_{11} + c_{10}) = 0,$$

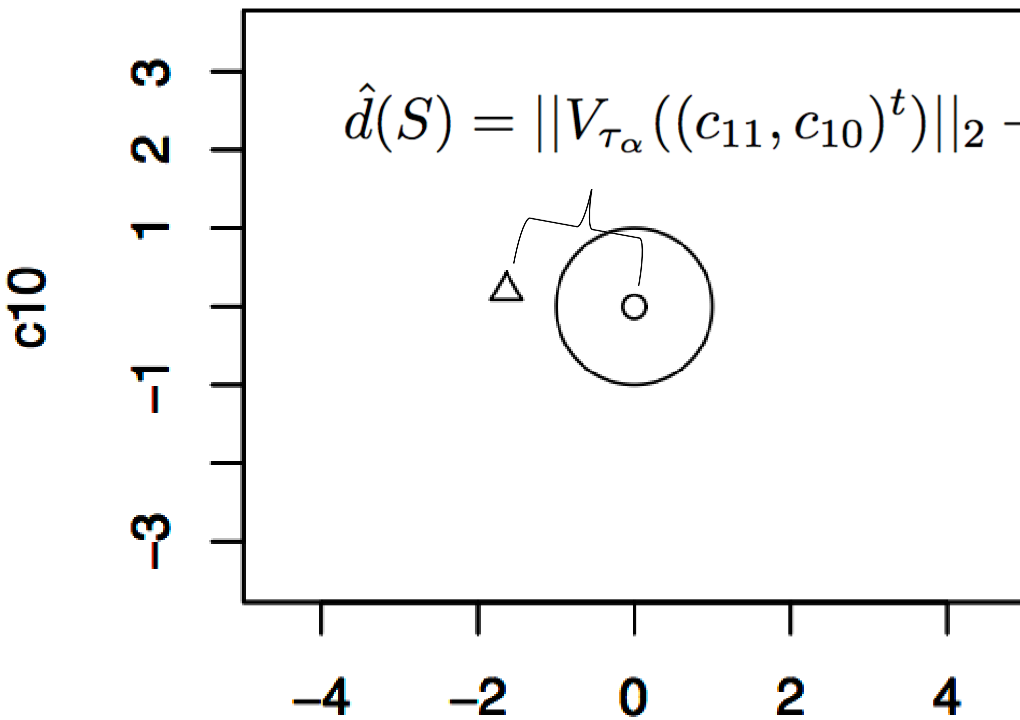
(forms an ellipsoid)



Unit circle mechanism

Test result:
- H_0 is accepted if $\|V((c_{11}, c_{10})^t)\|_2 + \text{noise} \leq 1$
- H_1 is accepted otherwise

$$\hat{d}(S) = ||V_{\tau_\alpha}((c_{11}, c_{10})^t)||_2 + \mathrm{Lap}(\frac{\Delta_{V,\alpha}(N_0, N_1)}{\epsilon}).$$



Randomize the distance from the origin and judge the result

Analysis of upper bound of type-II error

$$\Pr[M(S,\hat{\tau}_\alpha) = \mathrm{acc} | H_1 \text{ is true}]$$
$$\leq \sup_{P\in\mathcal{P}} \Big\{ \underbrace{\Pr_{S\sim P}[M(S,\hat{\tau}_\alpha) = \mathrm{acc} | \chi^2(S) > \hat{\tau}_\alpha + \gamma]}_{①} + \underbrace{\beta_{\hat{\tau}_\alpha+\gamma}}_{②} \Big\}$$

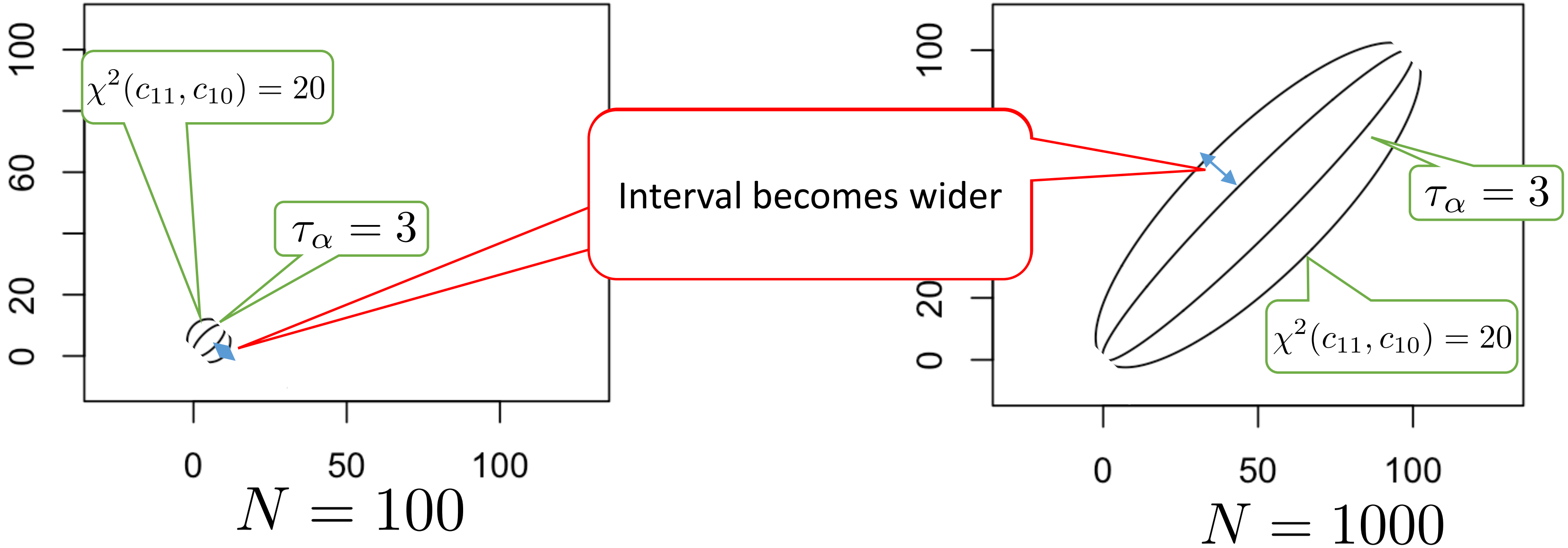
	① Gamma error	② Dependency on Sensitivity
Input perturbation (IP)		
Output perturbation (OP)	$O(1)$	$O(1)$
Unit circle mechanism (UCM)	$O(\exp(-\sqrt{N}))$	$O(1/\sqrt{N})$

- The type-II error of UCM is expected to decrease faster rate than OP
- The type-II error of IP cannot be analyzed

Why UCM has less gamma error?

- The cell of table (c_{11}, c_{10}) changes ± 1 when $S \to S'$
  - Then, the coordinate of cell move on (c_{11}, c_{10})-plane
- If we fix $\gamma = |\tau_\alpha - \chi^2(c_{11}, c_{10})|$, the interval between ellipse of τ_α and coordinate becomes wider, as N increases

– The randomized test statistics by UCM becomes less $S \to S'$ sensitive to noise as N increases



Contributions

1. Investigate the type-II error of DP mechanisms analytically
2. A novel DP mechanism with $O(\exp(-\sqrt{N}))$ type-II error
3. A novel DP mechanism that can control the family-wise error rate (FWER)

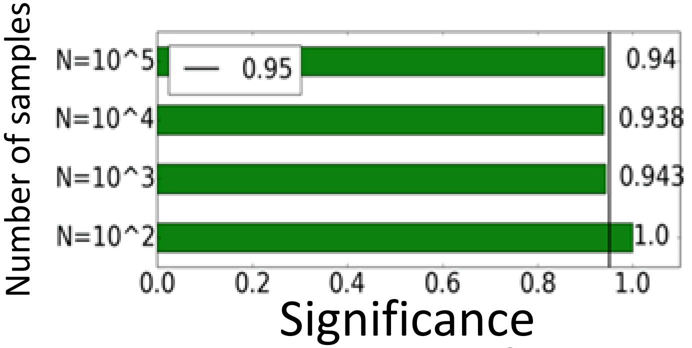
skipped

Experiment for single test

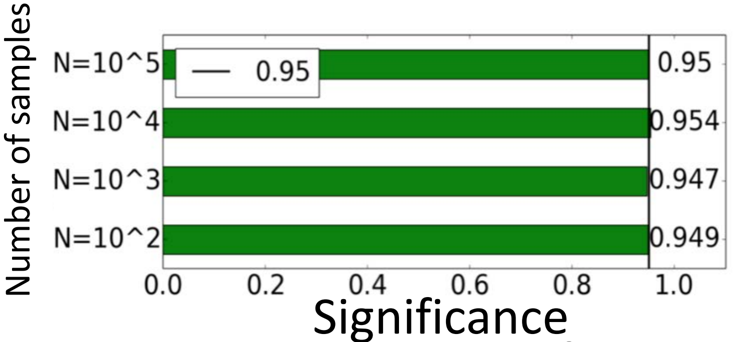
- We evaluate the significance and the power
  - Input perturbation
  - Output perturbation
  - Unit circle mechanism
- Type-I error is controlled by using MC sampling respectively
- Data
  - We sample 1000 contingency table from multinomial distribution
- Parameter
  - Significance level $\alpha = 0.05$
  - Privacy parameter $\epsilon = 0.1$

Significance result

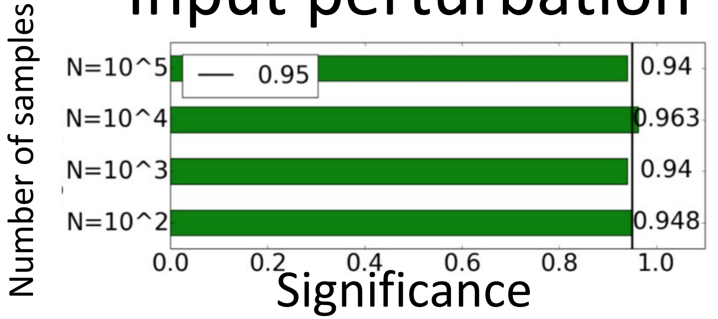
- Data: sample from $mult(0.25, 0.25, 0.25, 0.25)$
- Measure: significance = (1 - type-I error)



Input perturbation



Output perturbation

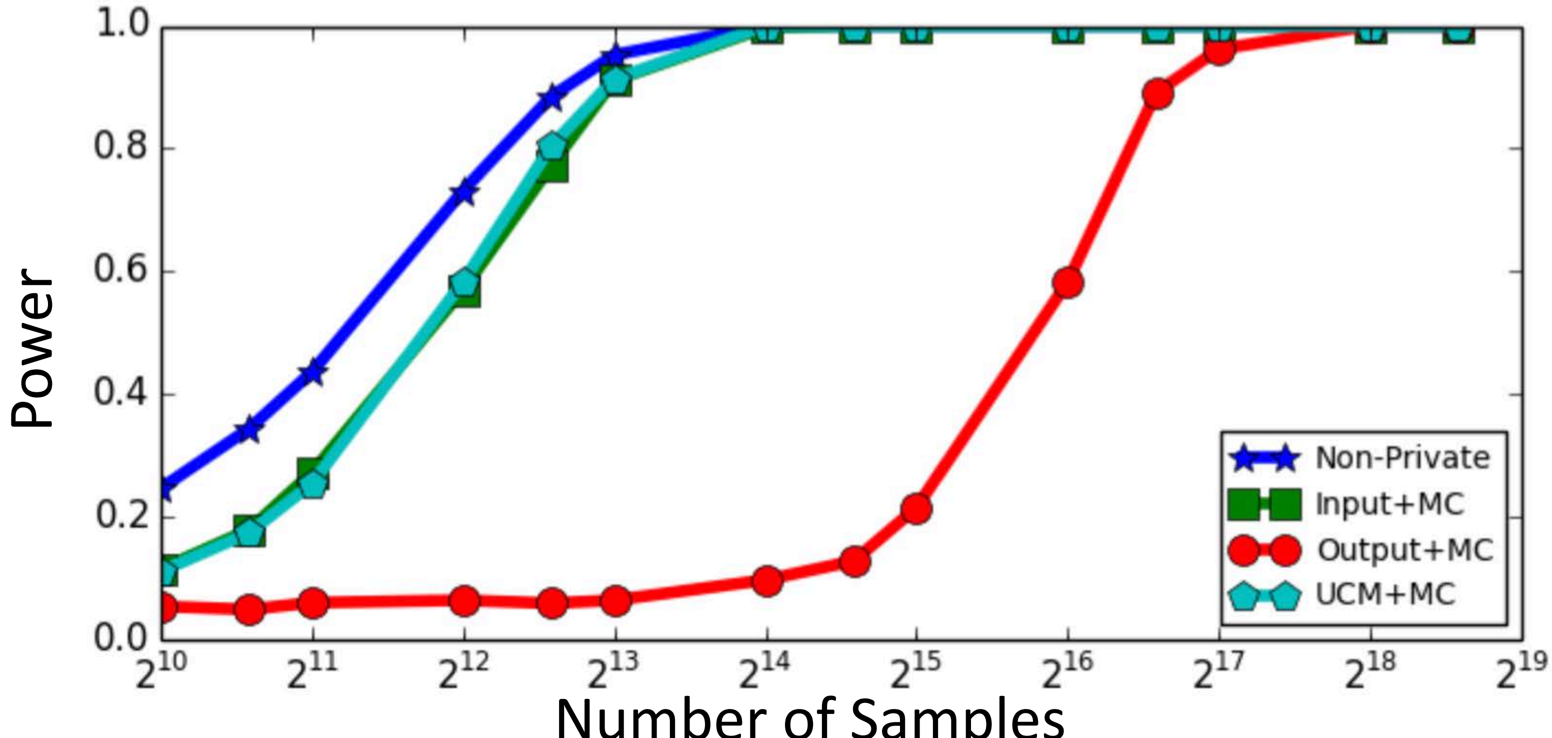


Unit circle mechanism

All mechanisms can properly control the significance at 0.95 for any sample size

Power result

- Data: sample from $mult(0.26, 0.24, 0.24, 0.26)$
- Evaluation: Power (1 – type-II error)



① UCM has a faster rate than OP
  - Because gamma error of UCM decreases as the sample size increases

② UCM has similar power to that of the IP
  - However type-II error of UCM is analyzed unlike IP

Conclusions and future work

We provide procedures for differential private chi-squared test and multiple version

- Contributions
  1. Investigate the upper bound of type-II error of OP and UCM
  2. A novel differentially private mechanism (unit circle mechanism)
     - Improve the dominated term of the type-II error from $O(1)$ to $O(\exp(-\sqrt{N}))$
  3. Framework of differential private multiple chi-squared test
     - Control the family-wise error rate (FWER) properly
- Future work
  - Investigate the upper bound of type-II error of IP