Differentially Private Chi-squared Test by Unit Circle Mechanism

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

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

(to appear in ICML 2017)
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
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.

Smoked marijuana last year…

Never tried …

Yes 1/4+p/2

No

Randomization preserves deniability

http://www.irasutoya.com/
**Definition 1** (\(\varepsilon\)-DP (Dwork et al., 2006)). Mechanism \(\mathcal{M} : S \rightarrow \mathcal{Y}\) provides \(\varepsilon\)-DP if, for any \(S \sim S'\) and \(Y \subseteq \mathcal{Y}\),

\[
\Pr[\mathcal{M}(S) \in Y] \leq \exp(\varepsilon) \Pr[\mathcal{M}(x') \in Y].
\]
χ² test for independence

<table>
<thead>
<tr>
<th></th>
<th>Smoker</th>
<th>Non-smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung cancer</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No lung cancer</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

• 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

   – 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

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

- $\beta_{\tau_\alpha}$: type-II error when one uses threshold $\tau_\alpha$
- $\mathcal{P} = \{ P : H_1 \text{ is true} \}$: 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 $\alpha$

\(\hat{\tau}_\alpha\):

- ① measures how often the mechanism wrongly accepts $H_0$ (type-II 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 $\gamma$ error is upper-bounded by \( \frac{1}{2} \exp \left( \frac{-\gamma \epsilon}{\Delta} \right) \)
  • Not decreases w.r.t. N

• Input perturbation
  • The $\gamma$ 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 variables

<table>
<thead>
<tr>
<th></th>
<th>$X_1 = 1$</th>
<th>$X_1 = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_0 = 1$</td>
<td>$c_{11}$</td>
<td>$c_{10}$</td>
</tr>
<tr>
<td>$X_0 = 0$</td>
<td>$c_{01}$</td>
<td>$c_{00}$</td>
</tr>
<tr>
<td></td>
<td>$N_1$</td>
<td>$N_0$</td>
</tr>
</tbody>
</table>

Test statistic

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

The test statistic is a function of $c_{11}$ and $c_{10}$. 
Geometrical interpretation of statistics

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

\[ A c_{11}^2 + B c_{10}^2 + 2C c_{11} c_{10} + D(c_{11} + c_{10}) = 0, \]
(forming an ellipsoid)

[Diagram showing geometric interpretation with rejected and accepted regions]

Affine trans.
Unit circle mechanism

Test result

\[ H_0 \text{ is accepted if } \|V((c_{11}, c_{10})^t)\|_2 + \text{noise} \leq 1 \]

\[ H_1 \text{ is accepted otherwise} \]

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

Randomize the distance from the origin and judge the result.
Analysis of upper bound of type-II error

\[
\Pr[M(S, \hat{\tau}_\alpha) = \text{acc} | H_1 \text{ is true}] \\
\leq \sup_{P \in \mathcal{P}} \left\{ \frac{\Pr[S \sim P, M(S, \hat{\tau}_\alpha) = \text{acc} | \chi^2(S) > \hat{\tau}_\alpha + \gamma]}{\beta \hat{\tau}_\alpha + \gamma} \right\}
\]

<table>
<thead>
<tr>
<th></th>
<th>① Gamma error</th>
<th>② Dependency on Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input perturbation (IP)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Output perturbation (OP)</td>
<td>(O(1))</td>
<td>(O(1))</td>
</tr>
<tr>
<td>Unit circle mechanism (UCM)</td>
<td>(O(\exp(-\sqrt{N})))</td>
<td>(O(1/\sqrt{N}))</td>
</tr>
</tbody>
</table>

- 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 \(\pm 1\) when \(S \rightarrow S'\)
  - 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 \(\tau_\alpha\) and coordinate becomes wider, as \(N\) increases
  - The randomized test statistics by UCM becomes less 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 $\text{mult}(0.25, 0.25, 0.25, 0.25)$

- **Measure:** significance = $(1 - \text{type-I error})$

Unit circle mechanism

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

- Data: sample from \( \text{mult}(0.26, 0.24, 0.24, 0.26) \)
- Evaluation: Power \((1 – \text{type-II error})\)

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