差分プライバシーの母数の決め方

星野　伸明

金沢大学・経済学経営学系

2021 年 3 月 12 日

本研究は科研費、統計数理研究所共同利用研究経費の助成を受けている。

概要

1. 差分プライバシーとは
   - データの真値を不確実にする程度を母数 ϵ で管理
2. データの度数表現とランダムネスによる保護
   - 度数の真値をランダム観測から推定
3. 離散の情報量不等式
   - 真値の推定量の分散の下限
4. ϵ の決め方
   - 真値を当てたという主張の有意水準を管理

差分プライバシー (DP) とは何だろうか

- 問題意識：そのまま公にするとプライバシーの問題があるデータをマスク（変換）して発行。いかなる意味で「安全」?
- DP は確率的なマスクの安全性基準。
  - 確率的なマスクの例）加法的なランダムノイズ、ランダムスワッピング、サンプリング、synthetic data など。
  - DP は k-匿名など決定論的なマスキングを危険とみなす。場合によって確定的な逆変換が出来てしまうことを防ぐには、発行データをランダムとするのが効果的。
  - 決定論的なマスクは精度保証が難しい（有為抽出と同じ)。
- DP が保護するものは、素直に考えて元データの真値である。
  - ある個体が元データに含まれるか否かの”presence”を保護しているのでは（必ずしも）ない。

差分プライバシーの定義

- （保護したい）データセット：$D \in \mathcal{D}$
- ランダムなマスク機構：A
  - A は D を公開データへ変換する：$A(D) \in \mathcal{A}$
- $D + \Delta$ が D と一単位だけ違うデータセットとする。
- 【定義】以下の式を満たす A を ϵ-差分プライベートと呼ぶ。
  $\forall S \subseteq \mathcal{A}, \forall D \in \mathcal{D}, \forall (D + \Delta) \in \mathcal{D},$

$$\frac{\mathrm{P}_A(A(D + \Delta) \in S)}{\mathrm{P}_A(A(D) \in S)} \leq \exp(\epsilon).$$

  - Δ の有無が公開データを観測してもはっきりしない、と解釈される。

プライバシーとの関係：Δ の解釈

- Dwork(2006) は Δ を特定個体の D への加入 (メンバーシップ) と解釈。暗号論の”presence”概念が下敷き。
- しかしこの解釈は限られた状況でしか成立しない。
  1. Δ が個体と解釈できる場合でも、データセットに同じ変化を及ぼす任意の個体である。
     - 例）D がゲノムデータセットなら、異なる個体は異なる Δ と対応。母集団一意だから。
     - 例）D がコロナ感染者の性別のみのデータセットなら、異なる個体が同じ Δ になる。
  2. 個体の加入以外の要因で変化するデータで解釈破綻。
     - 例）D が星野の週間飲酒回数の時系列データだとして、Δ は個体ではありえない。
     - Δ をレコード値の一単位変化とみなしても DP は定義可能

(Dwork et al., 2006)。

- 個体加入の曖昧化は DP の特殊ケースである。一般に Δ は D の値の単位変化と考えるべき。
- つまり保護対象は D であり Δ ではない。
  1. データの真値が当てにくくなるという意味でプライバシー保護になっている。統計学的な説明を後述。
  2. このように考える方が情報保護を設計しやすい。
     - 守りたいものを D に置くだけ。そうしてますよね？
     - 個体識別と加入を分離可能。
       * Δ を個体の加入と解釈すれば、任意の個体の加入を曖昧にすることで特定の個体の加入を曖昧にすることになる。個体の加入は明らかになってもよいが識別は曖昧にしたい場合、過保護。

D の表現：離散属性空間

- 現実のデータは全て離散（有限）属性空間の一点とみなせる。
  - 連続変数も観測結果は有限桁で離散。なお可算無限でも理論的には問題ないが、計算機は有限集合しか扱えない。
  - 例）(身長 (cm), 体重 (kg)) $\in (0, 1, \ldots, 300) \times (0, 1, \ldots, 300)$
  - 例）(検索に用いた単語) $\in \{$ 入力可能な文字列 $\}$
- D は属性空間の点のインデクスの度数ベクトル。
  1. 仮名付きなどレコード番号に情報がある場合、各レコードをインデクスの度数ベクトル $(0, 0, \ldots, 1, \ldots, 0)$ で表現する。これを順に並べたのが D で $\vec{n} = (n_1, n_2, \ldots, n_J) \in \{0, 1\}^J$ と表現。
  2. レコード番号が無情報なら、D は各インデクスの総度数 $\vec{n} = (n_1, n_2, \ldots, n_J) \in {\mathbb{N}_0}^J$ で表現される。

D の保護：例

- $\vec{n}$ を正確に推定できれば、個体属性が正確に判明することがある。
  1. A さんと B さんの先週の飲酒回数 (0,1-6,7 回以上) が $\vec{n} = (0, 0, 1, 1, 0, 0)$ なら、A さんは確実に飲み過ぎ。
  2. A 村在住の弁護士たちの年収度数分布が (低, 中, 高)$=(0, 0, 3) = \vec{n}$ なら、A 村在住弁護士は皆、年収が確実に高い。$(0, 1, 2)$ なら低くないことが確実。
  3. 先月の調査で A 村在住弁護士の年収が $(3, 4, 5)$ であった。今月の調査で $(3, 5, 5)$ なら、新しく引っ越してきた H さんの年収は確実に中。

ランダムマスクの出力

- D にランダムマスク A を適用した公開データセットも離散属性空間の点の集合である。
  - 例）身長にラプラースノイズを加えた結果は、計算機で扱うなら離散値。
- レコード番号が無情報だとして、各インデクスの度数の集合 $\vec{m} = (m_1, m_2, \ldots, m_K)$ で $A(D)$ を表す。
- 問題：$\vec{m}$ を見て $\vec{n}$ がどの程度正確に推定出来るか？
  - 例）A がサンプリングなら、$\vec{m}$ が標本度数で母集団度数 $\vec{n}$ を推定。
  - 例）A が模造なら、$\vec{m}$ が synthetic data で母集団度数 $\vec{n}$ を推定。

統計的推定の精度評価

- 一般に考えて $\vec{n}$ の任意の関数 $g(\vec{n})$ が推定対象。
  - 例）属性 j を持つ母集団個体数を推定したいなら $g(\vec{n}) = n_j$ と考える。
- 観測の関数 $\gamma(\vec{m})$ が $g(\vec{n})$ の推定量とする。
- γ は不偏推定量とする。つまり $\mathrm{E}(\gamma(\vec{m})) = g(\vec{n})$。
- 注目）相関係数の絶対値は 1 以下：

$$1 \geq \frac{\mathrm{Cov}(\gamma, \psi)^2}{\mathrm{V}(\gamma)\mathrm{V}(\psi)}$$

  - $\psi = \mathrm{P}(\vec{m}; \vec{n} + \Delta)/\mathrm{P}(\vec{m}; \vec{n}) - 1$ とすれば $\mathrm{Cov}(\gamma, \psi) = g(\vec{n} + \Delta) - g(\vec{n})$ になる。

情報量不等式（離散）

- Hammersley-Chapman-Robbins inequality:

  Suppose that $\vec{m} \sim \mathrm{P}(\vec{m};\vec{n})$ and $\mathrm{P}(\vec{m};\vec{n}) > 0$ for all $\vec{m}$. If $\vec{n}$ and $\vec{n}+\Delta$ are two values for which $g(\vec{n}) \neq g(\vec{n}+\Delta)$, then for any unbiased estimator γ of $g(\vec{n})$,

$$\mathrm{V}(\gamma) \geq [g(\vec{n}+\Delta) - g(\vec{n})]^2 / \mathrm{E}[\frac{\mathrm{P}(\vec{m};\vec{n}+\Delta)}{\mathrm{P}(\vec{m};\vec{n})} - 1]^2.$$

  - The support of $\vec{m}$ is assumed to be common between the cases of $\vec{n}$ and $\vec{n}+\Delta$.
  - 注）この不等式はそれほどシャープではない。

- つまり尤度比 $\mathrm{P}(\vec{m};\vec{n}+\Delta)/\mathrm{P}(\vec{m};\vec{n})$ が 1 から離れるほど正確に推定可能。DP はこの比を 1 近辺に制限。

n_j の不偏推定量の分散の下限

- ϵ-DP$\Leftrightarrow \forall(\vec{m}, \vec{n}, \Delta)$

$$\mathrm{P}(\vec{m}; \vec{n} + \Delta)/\mathrm{P}(\vec{m}; \vec{n}) \leq \exp(\epsilon). \tag{1}$$

- (1) 式は下記の不等式の十分条件である (N.B., $\epsilon > 0$)。

$$\mathrm{E}[\frac{\mathrm{P}(\vec{m}; \vec{n} + \Delta)}{\mathrm{P}(\vec{m}; \vec{n})} - 1]^2 \leq (\exp(\epsilon) - 1)^2.$$

Proposition 1 *If* (1) *holds then the lower bound of the variance of the unbiased estimator of* n_j *is* $1/(\exp(\epsilon) - 1)^2$.

ϵ	.01	.1	.5	1	2	3
$\mathrm{V}(\hat{n_j}) \geq$	9900.4	90.4	2.38	.339	.024	.003
S.D.$\geq$	99.5	9.51	1.54	.582	.155	.055

Table 1: Accuracy of $\hat{n_j}$

ϵ の設定—既存研究

- $\vec{n}$ の不確実性はどの程度なら許容される？
- プライバシー予算 ϵ の設定は未解決問題。
  - Google は $\epsilon = \log(3) \doteq 1.1$、Apple は $\epsilon = 1 \sim 2$ を利用か。
  - 既存研究 (Lee and Clifton (2011), Li et al. (2019)) は D に事前分布を仮定。いずれにせよラプラースノイズに依存した議論。一般性に欠ける。ケース依存で ϵ を決めると ϵ にケースの情報が乗る。
- 尤度比（＝ベイズファクター）の大きさには目安が存在。

Table 2: Interpretation of the Bayes factor (Kass and Raftery, 1995)

$\log_{10} \exp(\epsilon)$	$\exp(\epsilon)$	ϵ	Evidence against null
0 to 1/2	1 to 3.2	0 to 1.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	1.2 to 2.3	Substantial
1 to 2	10 to 100	2.3 to 4.6	Strong
> 2	> 100	> 4.6	Decisive

Table 3: Interpretation of the Bayes factor (Evett et al., 2000)

$\exp(\epsilon)$	ϵ	Evidence against null
1 to 10	0 to 2.3	Limited evidence to support
10 to 100	2.3 to 4.6	Moderate evidence to support
100 to 1000	4.6 to 6.9	Moderately strong evidence to support
1000 to 10000	6.9 to 9.2	Strong evidence to support
> 10000	> 9.2	Very strong evidence to support

- D を母数と見て尤度比の漸近分布を使うことも考えられる。
  - Δ が一次元なら、$2\log(\mathrm{P}(A(D+\Delta)\in S)/\mathrm{P}(A(D)\in S))$ は自由度 1 の χ^2 分布に漸近的に従う。
    - 「漸近」の意味は、公開データが iid でサイズが大きいということ。現実的でない。
  - χ^2 分布の $(1-\alpha)$ 分位点と 2ϵ が等しいとすれば、直感的な α で ϵ が定まる。
  - α=10%, $\epsilon = 1.353$; α=5%, $\epsilon = 1.921$; α=1%, $\epsilon = 3.317$.

ϵ の設定についての提案

- 提案手法の利点：
  1. 漸近論を使わない：中心極限定理の妥当性が低い状況。
  2. 事例毎の個別評価を必要としない：ϵ に事例の情報が乗らない。
- Adversary は n_j の真値が非負整数であることを知っている。$\hat{n_j}$ は実数値を取り得るので、$\mathrm{P}(|\hat{n_j} - n_j| \geq 1/2)$ は n_j を間違えて推定する確率と考える。
- $\hat{n_j}$ が n_j の不偏推定量として、チェビシェフの不等式より

$$\mathrm{P}(|\hat{n_j} - n_j| \geq 1/2) \leq 4\mathrm{V}(\hat{n_j}).$$

  – 注）チェビシェフの不等式はシャープ。

- Proposition 1 より $\mathrm{V}(\hat{n_j}) \geq (\exp(\epsilon) - 1)^{-2}$ である。つまり最強の adversary でも

$$\mathrm{P}(|\hat{n_j} - n_j| \geq 1/2) \leq \frac{4}{(\exp(\epsilon) - 1)^2} =: \alpha_\epsilon. \tag{2}$$

- α_ϵ は過誤確率の上限なので、α_ϵ より大の有意水準で正確な推定が主張されうる。（例：$H_0 : n_j = 100$ が正しいのに $H_1 : n_j = 1$（母集団一意）と主張するような、第一種の過誤の確率は α_ϵ 以下である。）
  - α_ϵ より小さい有意水準での主張を否定する理屈。
  - Adversary の主張についての deniability を α_ϵ で管理。
- $\alpha_\epsilon = 10\%$, $\epsilon = 1.991$; $\alpha_\epsilon = 5\%$, $\epsilon = 2.297$; $\alpha_\epsilon = 1\%$, $\epsilon = 3.045$.
  - これまでの例とかけ離れてはいない。

参考文献

- 情報量不等式の話など：

  Hoshino, N. (2020) A firm foundation for statistical disclosure control. *Japanese Journal of Statistics and Data Science*, Vol. 3, 721–746.