General Inferential Limits of Differential Privacy via Intervals of Measures

James Bailie

Department of Statistics
Harvard University

13 July 2023

13th International Symposium on Imprecise Probabilities: Theories & Applications
(ISIPTA 2023)

Ruobin Gong

Asst. Professor, Rutgers University

Privacy: a challenge in modern data curation

Modern data curators seek to meet two goals at once:

1. To **disclose** key statistics of the database, in accordance with its legal, policy, and ethical mandates.
2. To protect the **privacy** of individuals with trustworthy guarantees.

TITULO PRIMERO

De las estadísticas y su régimen

CAPITULO PRIMERO

Principios generales de la Función Estadística Pública

Artículo 4

1. La recogida de datos con fines estadísticos se ajustará a los principios de secreto, transparencia, especialidad y proporcionalidad.

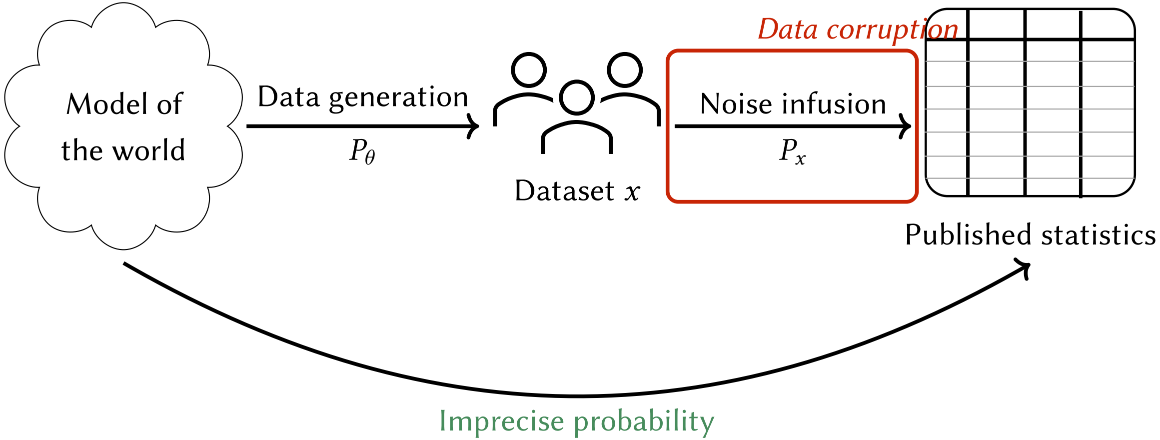
CAPITULO III

Del secreto estadístico

Artículo 13

1. Serán objeto de protección y quedarán amparados por el secreto estadístico los datos personales que obtengan los servicios estadísticos tanto directamente de los informantes como a través de fuentes administrativas.
2. Se entiende que son datos personales los referentes a personas físicas o jurídicas que o bien permitan la identificación inmediata de los

Privacy: a challenge in modern data curation



How can we encode the concept of *data privacy*?

Differential privacy (DP):

- A *family* of technical *standards*...
- which aim to quantify privacy as the change in P_x per change in x.
- **Ex**: Pure ϵ-differential privacy (Dwork et al., 2006).

Definition. A data-release mechanism $\{P_x\}$ satisfies (pure) ϵ-differential privacy if, for all x, x',

$$d_{\text{MULT}}(P_x, P_{x'}) \leq \epsilon d_{\text{HAM}}(x, x'),$$

where $d_{\text{MULT}}(P_x, P_{x'}) = \sup_S \left| \ln \frac{P_x(S)}{P_{x'}(S)} \right|$.

Applications: US Census, Apple, Facebook, LinkedIn, ...

An overview of results

We provide general limits on important statistical quantities in:

1. Likelihood inference (marginal probability of the observed, privatised data);
2. Frequentist inference (statistical power of Neyman-Pearson hypothesis testing);
3. Bayesian inference (prior predictive probability and posterior probability),

for arbitrary parameters, priors and data generating models, under ϵ-differential privacy.

The foundational tool we need

Interval of Measures

Definition (DeRobertis & Hartigan, 1981). Given measures L, U,

$$\mathcal{I}(L, U) = \{\mu : L(S) \leq \mu(S) \leq U(S) \ \forall S\}$$

is an *interval of measures.*

Theorem 5 (simplified)

$\{P_x\}$ is ϵ-DP – i.e. $d_{\text{MULT}}(P_x, P_{x'}) \leq \epsilon d_{\text{HAM}}(x, x')$ – if and only if

$$P_{x'} \in \mathcal{I}(L_{x,m\epsilon}, U_{x,m\epsilon}),$$

where $L_{x,m\epsilon} = e^{-m\epsilon} P_x$; $U_{x,m\epsilon} = e^{m\epsilon} P_x$; and $m = d_{\text{HAM}}(x, x')$.

Outlook

- We connect ϵ-DP to an IP concept – the interval of measures.
- We derive bounds on likelihood, frequentist and Bayesian inference which...
  1. Are near assumption free, optimal, and hence represents the limits of learning,
  2. Apply to both attackers and valid analysts,
  3. Generalise existing results.
- Further connections between IP & DP:
  - Are there other equivalences between IP objects & DP standards?
    - For example, Pufferfish DP which assumes a family of priors on the data x.
  - Can we use IP tools to...
    1. analyse DP statistics?
    2. develop new DP mechanisms?
    3. construct new DP standards?