BU CS599
Foundations of Private Data Analysis
Spring 2023

Lecture 22: Inference and DP

Jonathan Ullman
NEU

Adam Smith
BU

Inference with DP

- Inference vs computation
- Confidence intervals
  - Estimating the bias of a coin
- Confidence intervals from complex algorithms
  - Estimating median from the binary-tree CDF
- Bootstrap-based approaches
- Topics not covered

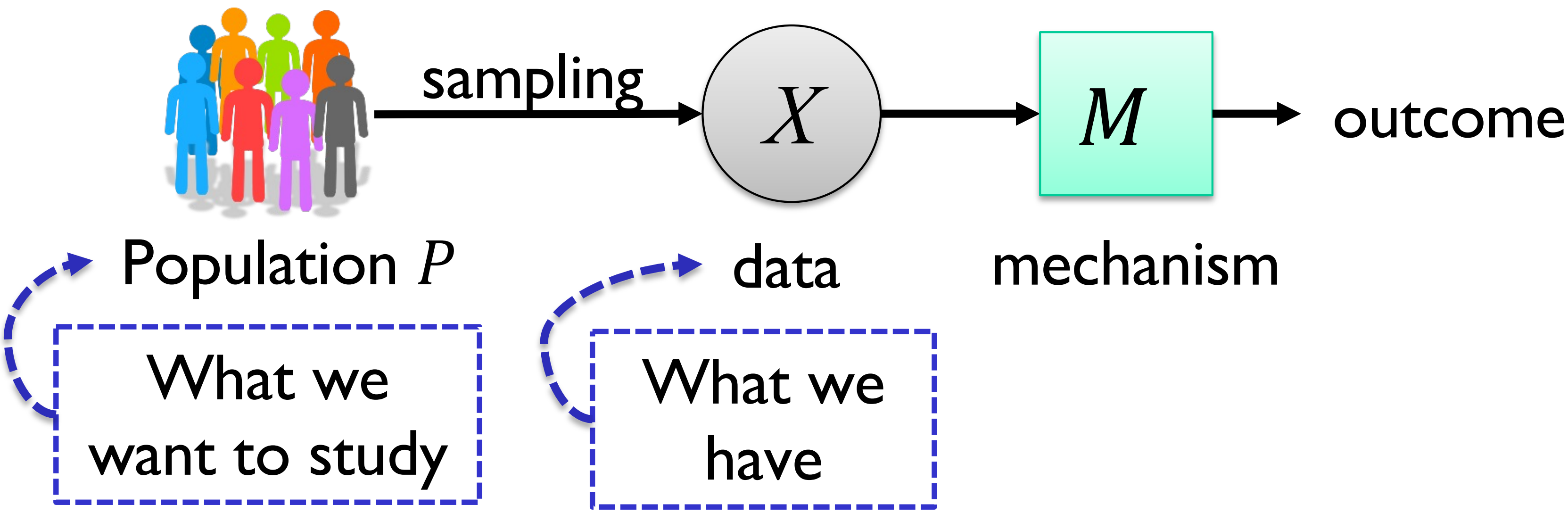
Inference versus computing a function

Subject	Peoria city, Illinois Estimate	Margin of Error	Percent	Percent Margin of Error
HOUSEHOLDS BY TYPE				
Total households	47,756	+/-1,640	47,756	(X)
Family households (families)	27,259	+/-1,641	57.1%	+/-3.2
With own children of the householder under 18 years	12,567	+/-1,332	26.3%	+/-2.7
Married-couple family	17,437	+/-1,657	36.5%	+/-3.3
With own children of the householder under 18 years	7,008	+/-1,155	14.7%	[illegible]
Male householder, no wife present, family	1,939	+/-634	4.1%	[illegible]
With own children of the householder under 18 years	1,038	+/-511	2.2%	[illegible]
Female householder, no husband present, family	7,883	+/-1,205	16.5%	[illegible]
With own children of the householder under 18 years	4,521	+/-1,038	9.5%	[illegible]
Nonfamily households	20,497	+/-1,804	42.9%	[illegible]
Householder living alone	17,685	+/-1,748	37.0%	[illegible]
65 years and over	5,917	+/-903	12.4%	[illegible]
Households with one or more people under 18 years	13,799	+/-1,360	28.9%	[illegible]
Households with one or more people 65 years and over	12,130	+/-935	25.4%	+/-2.8
Average household size	2.40	+/-0.07	(X)	(X)
Average family size	3.15	+/-0.13	(X)	(X)

47,756	+/-1,640
27,259	+/-1,641
12,567	+/-1,332

- American Community Survey
  - Covers ≈ 1% of the US population per year
  - Includes "ancestry, citizenship, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics"
- Meant to inform us about the population as a whole
  - Sample itself is not of interest

Statistical inference



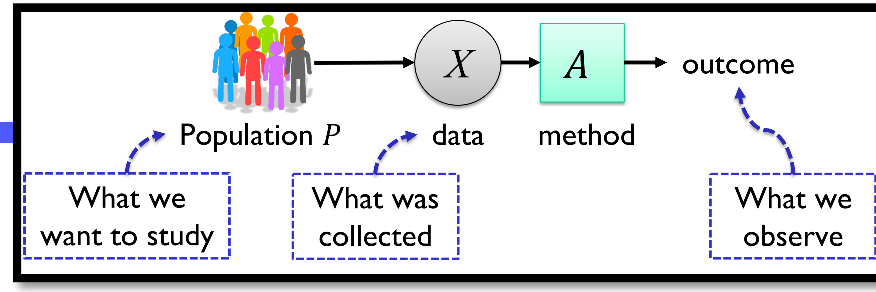
- Goal: Figure out something about P
  - ➢ Good classifier
  - ➢ Estimate for some parameter of P
    - Confidence interval: plausible range for the parameter
  - ➢ Test if P satisfies some hypothesis
    - E.g. smoking and lung cancer are independent

Two Settings

1. Externally specified mechanism
   - Census is using “TopDown”
   - How can social scientists draw inferences?

2. Algorithm design
   - What mechanisms make inference easy?
   - Are they good enough?

Theories of Inference



- Bayesian [lots of work]
  - ➢ Posit a prior Q on the data distribution P
  - ➢ Given $a = A_\epsilon(X)$, compute conditional distribution on $f(P)$

$$\Pr_{\substack{P \sim Q \\ X \sim P^n}} \left(f(P) = \theta \mid a = A_\epsilon(X)\right)$$

    - Incorporates all randomness, supports all inference tasks ☺
    - Often computationally hard ☹
    - Limited by prior. Social scientists suspicious ☹
- Frequentist [today]
  - ➢ Example: Find function $CI: a \to [low, high]$ such that

$$\forall P \in \mathcal{P}: \quad \Pr_{X \sim P^n} \Big(f(P) \in CI\big(A_\epsilon(X)\big)\Big) \approx 0.95$$

    - Often computationally simpler ☺
    - Correctness is often brittle ☹

Today: Two specific problems

Parametric estimation

- Estimating a coin's bias (Bernoulli)
  - $B(p)$: Output $\begin{cases} 1 & \text{w.p.} \quad p \\ 0 & \text{w.p.} \quad 1-p \end{cases}$
  - Given $X_1, \dots, X_n \sim_{iid} P = B(p)$

- Median
  - $X_1, \dots, X_n \sim_{iid} P$ on $[0,1]$ with CDF F
  - Want w such that $F(w) = \frac{1}{2}$
    (or $\inf\left\{w : F(w) \geq \frac{1}{2}\right\}$)

Bernoulli parameter estimation

- Say $X_1, \dots, X_n \sim Bern(p)$ so each $X_i \in \{0,1\}$
- We want an interval for p
- A frequentist confidence interval is an algorithm
  - Input: $x_1, \dots, x_n$ and parameter $\beta \in (0,1)$
  - Output: a, b

Two goals

- **Validity/coverage**: for all $p \in [0,1]$:

$$\Pr_{\substack{\boldsymbol{X}=(X_1,\dots,X_n)\sim B(p) \\ i.i.d.}} \left(p \in [a(\boldsymbol{X}), b(\boldsymbol{X})]\right) \geq 1 - \beta$$

- **Size**: Want $b - a$ as small as possible
  - E.g. in expectation

Bernoulli parameter estimation

- Say $X_1, \dots, X_n \sim Bern(p)$ so each $X_i \in \{0,1\}$
- **Validity/coverage**: for all $q \in [0,1]$:

$$\Pr_{\substack{\boldsymbol{X}=(X_1,\dots,X_n)\sim B(p)\\ i.i.d.}}(p \in [a(\boldsymbol{X}), b(\boldsymbol{X})]) \geq 1-\beta$$

Typical strategy for parametric estimation: Given $\boldsymbol{x}$,

1. Compute $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Let $a(\boldsymbol{x}) = \min\left\{q : \Pr_{\substack{Y_1,\dots,Y_n\sim B(q)\\ i.i.d.}}(\bar{Y} > \bar{x}) \geq \frac{\beta}{2}\right\}$

   $b(\boldsymbol{x}) = \max\left\{q : \Pr_{\substack{Y_1,\dots,Y_n\sim B(q)\\ i.i.d.}}(\bar{Y} < \bar{x}) \geq \frac{\beta}{2}\right\}$

In practice, often use upper bounds on tail probabilities

- Looser bounds lead to larger intervals

Validity

Proof:

- Two ways to be invalid: either $p < a(\boldsymbol{X})$ or $p > b(\boldsymbol{X})$
- Look at $\Pr_{\vec{X} \sim_{iid} B(p)} \left(p < a(\boldsymbol{X}) \right)$
  - Recall $a(\vec{x}) = \min \left\{ q : \Pr_{\substack{Y_1, \ldots, Y_n \sim B(q) \\ i.i.d.}} (\bar{Y} > \bar{x}) \geq \frac{\beta}{2} \right\}$

QED

Same proof works if we use upper bound on tails

- E.g. Chernoff bounds, or
  CLT: $\bar{X} \approx Z$ where $Z \sim N\left(p, \frac{p(1-p)}{n}\right)$. Ok for $n \gg \frac{1}{p(1-p)}$

Validity (with proof filled in)

Proof:

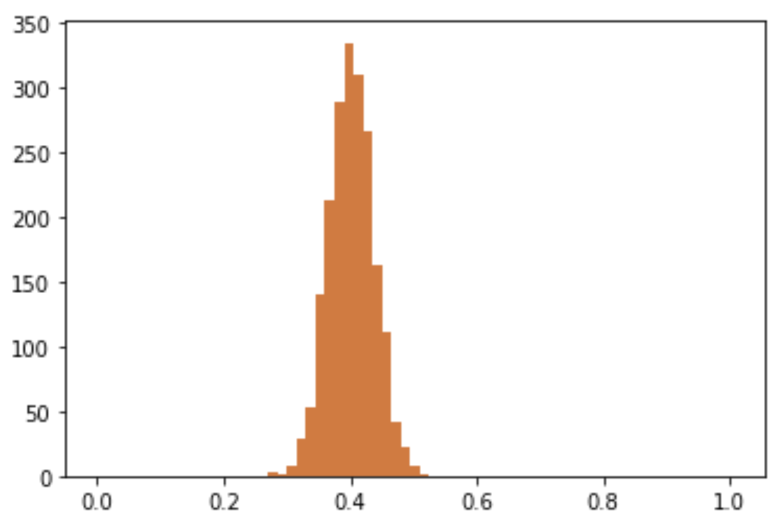
- Two ways to be invalid: either $p < a(\boldsymbol{X})$ or $p > b(\boldsymbol{X})$
- Look at $\Pr_{\vec{X} \sim_{iid} B(p)} \left(p < a(\boldsymbol{X})\right)$
  - ➢ Recall $a(\vec{x}) = \min\left\{q: \Pr_{\substack{Y_1,\ldots,Y_n \sim B(q) \\ i.i.d.}} (\bar{Y} > \bar{x}) \geq \frac{\beta}{2}\right\}$
  - ➢ If $p < a(\boldsymbol{X})$ then $\Pr_{\substack{Y_1,\ldots,Y_n \sim B(p) \\ i.i.d.}} (\bar{Y} > \bar{x}) < \frac{\beta}{2}$
    - By definition!
- Similarly, probability that $p > b(\boldsymbol{X})$ is at most $\frac{\beta}{2}$. QED

Same proof works if we use upper bound on tails

- ➢ E.g. Chernoff bounds, or
  CLT: $\bar{X} \approx Z$ where $Z \sim N\left(p, \frac{p(1-p)}{n}\right)$. Ok for $n \gg \frac{1}{p(1-p)}$

General strategy

- **Sampling distribution** of a statistic $g(\boldsymbol{X})$ for distribution P is the distribution you observe in the sample.

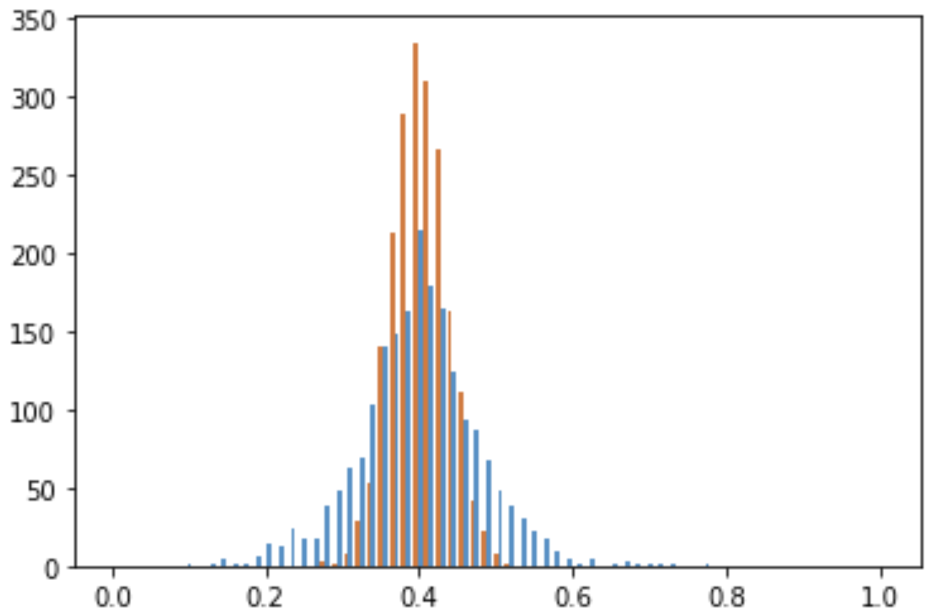


Sampling distribution of $\bar{X}$ where $\boldsymbol{X} \sim_{(iid)} B(0.4)$ and $n = 200$

- General approach: look how sampling distribution might have given rise to observed value

DP Confidence Intervals

- Given $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$,
  Run existing DP algorithm $M(\boldsymbol{x})$ to approximate $\bar{x}$
  - Example: $M(\boldsymbol{x}) = \bar{x} + Z$ where $Z \sim Lap\left(\frac{1}{\epsilon n}\right)$



Sampling distributions of $\bar{X}$ and $M(\boldsymbol{X})$ where $\boldsymbol{X} \sim_{(iid)} B(0.4)$ and $n = 200$ and $\epsilon = 0.1$

- How should we compute a confidence interval for p?

DP confidence intervals

- Approach #1:
  - Given $m = M(\boldsymbol{x}) = \bar{x} + Z$ where $Z \sim Lap\left(\frac{1}{\epsilon n}\right)$
  - Let

$$a(m) = \min\left\{q \colon \Pr_{\substack{Y_1,\ldots,Y_n \sim B(q) \\ i.i.d.}} (\bar{Y} \geq m) \geq \frac{\beta}{2}\right\}$$

$$b(m) = \max\left\{q \colon \Pr_{\substack{Y_1,\ldots,Y_n \sim B(q) \\ i.i.d.}} (\bar{Y} \leq m) \geq \frac{\beta}{2}\right\}$$

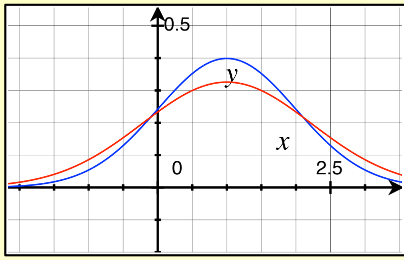
- Multiple choice: This approach produces

a) Valid intervals that are wider than they need to be

b) Valid intervals that are narrower than they need to be

c) Invalid intervals because they are too wide

d) Invalid intervals because they are too narrow



DP confidence intervals

- Approach #2:
  - Given $m = M(\boldsymbol{x}) = \bar{x} + Z$ where $Z \sim Lap\left(\frac{1}{\epsilon n}\right)$
  - Let
$$a(m) = \min\left\{q: \Pr_{\substack{Y_1,\ldots,Y_n \sim B(q) \\ i.i.d.}} (M(\boldsymbol{Y}) \geq m) \geq \frac{\beta}{2}\right\}$$
$$b(m) = \max\left\{q: \Pr_{\substack{Y_1,\ldots,Y_n \sim B(q) \\ i.i.d.}} (M(\boldsymbol{Y}) \leq m) \geq \frac{\beta}{2}\right\}$$
- This approach is correct, but not obviously the best
  - In fact, adding integer version of Laplace is slightly better [GRS'08]
- Approximating $\Pr_{\substack{Y_1,\ldots,Y_n \sim B(q) \\ i.i.d.}} (M(\boldsymbol{Y}) \geq m)$ can be tricky
  - Loose overestimates lead to wide intervals
  - Loose underestimates yield invalid intervals
  - General approach: sampling

Asymptotics

- Central Limit Theorem: when p fixed and $n \to \infty$,
$$\frac{M(\boldsymbol{X}) - p}{\sqrt{p(1-p)n}} \to_D N(0,1)$$
just like $\bar{X}$.
    - So $M(X)$ is "as good as" $\bar{X}$ for statistical purposes as $n \to \infty$
- But when we do inference, we have a finite sample
    - We need to adjust for added noise
    - For large n, the adjustment is small
- We can quantify the cost in terms of …
    - Interval width of private v. nonprivate methods (for same n)
    - Increase in sample size needed (for same expected width)

Comparing sample sizes

- Bernoulli: For given confidence, intervals have width
  - Nonprivate with n samples: roughly $2\, z_{1-\beta/2} \cdot \frac{1}{\sqrt{p(1-p)n}}$
    where $z_{1-\beta/2}$ is the $1-\beta/2$ quantile of $N(0,1)$
  - Private with n' samples: roughly $2\, z_{1-\beta/2} \cdot \sqrt{\frac{1}{p(1-p)n'} + \frac{\sqrt{2}}{(\epsilon n')^2}}$
    - (This assumes Laplace behaves roughly like Normal)
  - Solving for n' to get the same width α, for constant p:
    $$n' = n + \Theta\left(\frac{1}{\epsilon^2}\right)$$
    - (Exercise ☺)
- For most models, we at best get statements of the form
  $$n' = \Theta(n_{nonprivate} + f(\epsilon, \alpha))$$
  - Example: For Gaussian mean with known covariance
    $$n' = \tilde{\Theta}\left(\frac{d}{\alpha^2} + \frac{d}{\epsilon\alpha}\right)$$
  - Open question for many models!

General points

- Adjustments above were possible only because we knew an exact description of M
  - Needed to compute $\Pr_{\substack{Y_1,\ldots,Y_n \sim B(q) \\ i.i.d.}} (M(\boldsymbol{Y}) \geq m)$
- Until 2010, Census methods for adding distortion were confidential
  - Users had to make inferences by taking estimates at face value
- Move to publicly described methods has caused controversy
  - Many did not understand distortion was added at all
  - New distortion is often larger than previously added

Inference with DP

- Inference vs computation
- Confidence intervals
  - Estimating the bias of a coin
- Confidence intervals from complex algorithms
  - Estimating median from the binary-tree CDF
- Bootstrap-based approaches
- Topics not covered

Median

- $X_1, \ldots, X_n \sim_{iid} P$ on $[0,1]$ with CDF F
- Median: w such that $F(w) = \frac{1}{2}$
  (or $\inf\left\{w: F(w) \geq \frac{1}{2}\right\}$)
- We've seen DP algorithms for median
  - Exp. Mech. $\Pr(Y = y) \propto \exp\left(-\left|rank_x(y) - \frac{n}{2}\right|\right)$
  - CDF tree estimator
    - Extract an estimate for median by looking where the estimated CDF crosses above ½
  - (also MWEM)
- What problems will we get?

Nonprivate CI's for median

- Let's first solve the problem without DP…
  - Let F be the CDF of P and m^* be its true median
  - Let F_x be the CDF of the sample
- Find two quantiles q_-, q_+ that contain the median with probability $1 - \beta$.

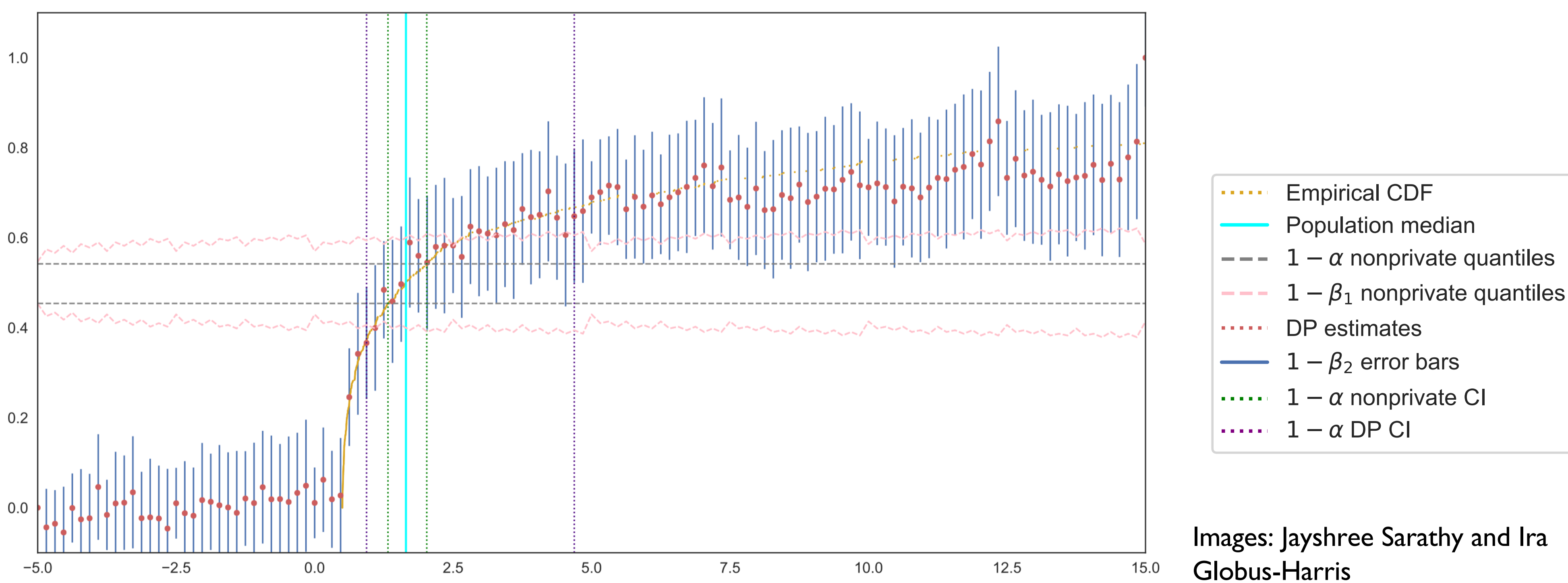
$$q_- = \sup\left\{q: \Pr_{\boldsymbol{X} \sim_{iid} P}\left(F_{\boldsymbol{X}}(m^*) \le q\right) \le \frac{\beta}{2}\right\}$$
$$= \sup\left\{q: \Pr_{Y \sim Bin\left(n, \frac{1}{2}\right)}\left(\bar{Y} \le q\right) \le \frac{\beta}{2}\right\}$$

  - q_+ is similar
- Given $\boldsymbol{x}$ with CDF $F_{\boldsymbol{x}}$, return

$$a(\boldsymbol{x}) = F_{\boldsymbol{x}}^{-1}(q_-) \quad \text{and}$$
$$b(\boldsymbol{x}) = F_{\boldsymbol{x}}^{-1}(q_+)$$

Using the CDF estimator

- Approach 1: For each w, find a confidence interval for w's quantile in the sample
  - Possible because we understand Gaussian noise for each x
  - $a =$ smallest value whose CI includes q_-
- Approach 2: For each w, find a confidence interval for w's quantile in the distribution
  - Possible because we understand Gaussian noise for each x and estimating the CDF at w can be viewed as Bernoulli estimation
  - $a =$ smallest value whose CI includes $1/2$



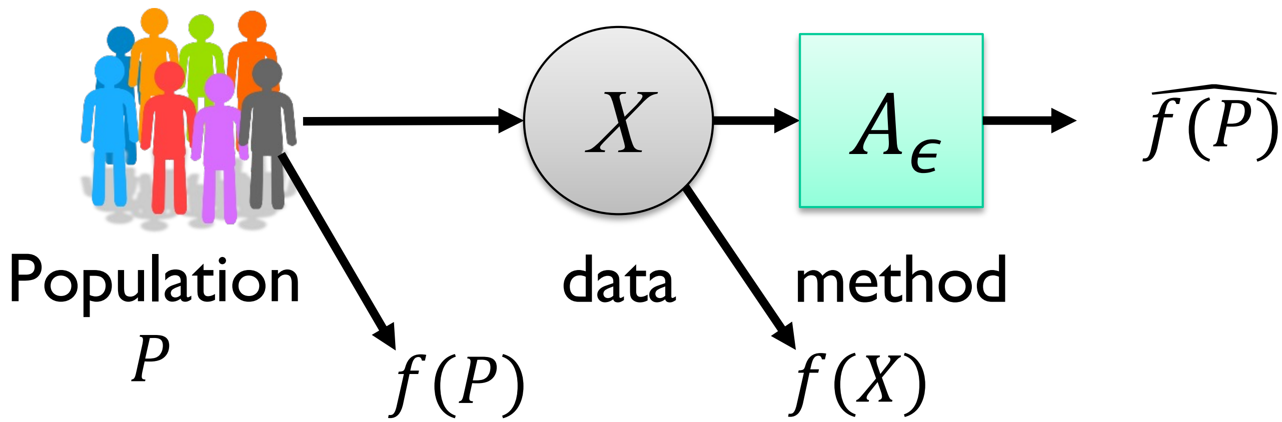
Images: Jayshree Sarathy and Ira Globus-Harris

Inference with DP

- Inference vs computation
- Confidence intervals
  - Estimating the bias of a coin
- Confidence intervals from complex algorithms
  - Estimating median from the binary-tree CDF
- Bootstrap-based approaches
- Topics not covered

Direct Estimation of Sampling Distribution

Sampling Distribution

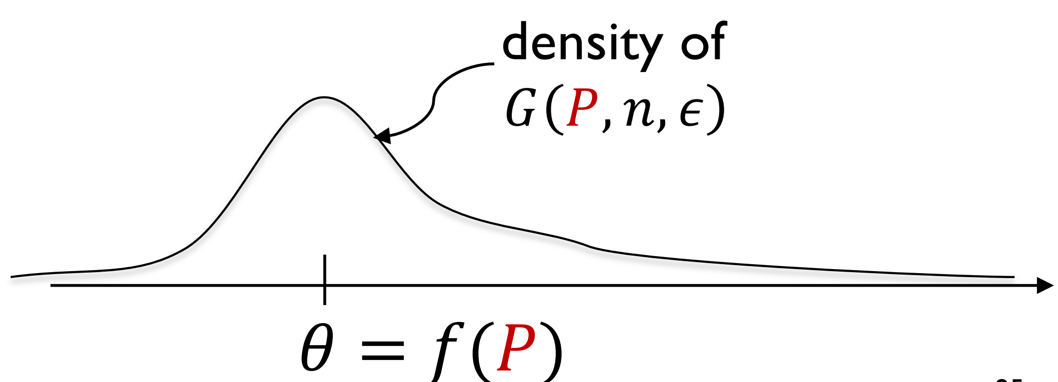


- Goal: CI for $f(P)$ from $A_\epsilon(X)$

- Intermediate goal: understand sampling distribution

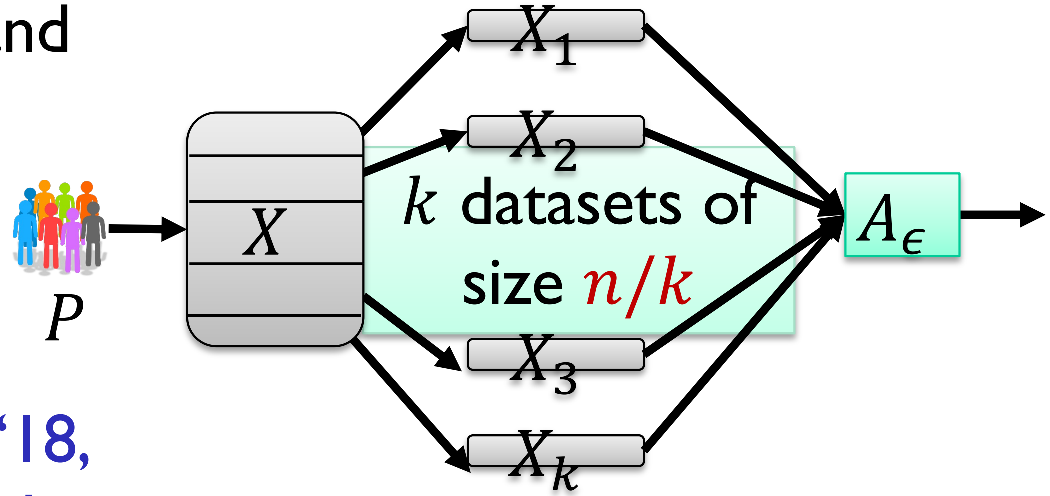
$$G(P, n, \epsilon)$$

of $A_\epsilon(X)$



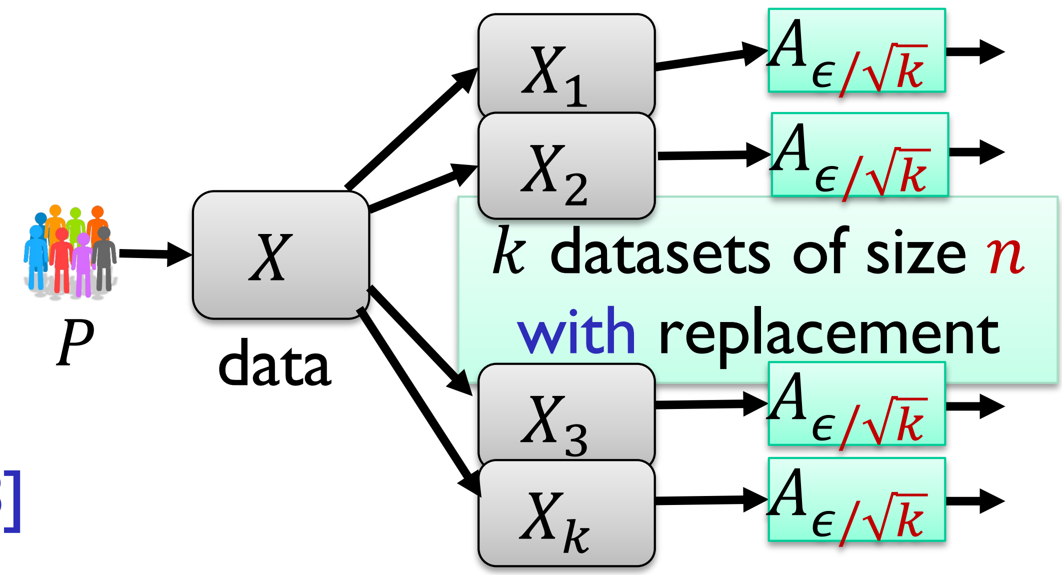
Direct Estimation of Sampling Distribution

Subsample and aggregate (smaller n) [NR**S**'07, **S**'11, Evans, King '18, Covington, He, Honaker, Kamath '21]



Samples from $G\left(P, \frac{n}{k}, +\infty\right)$ processed privately

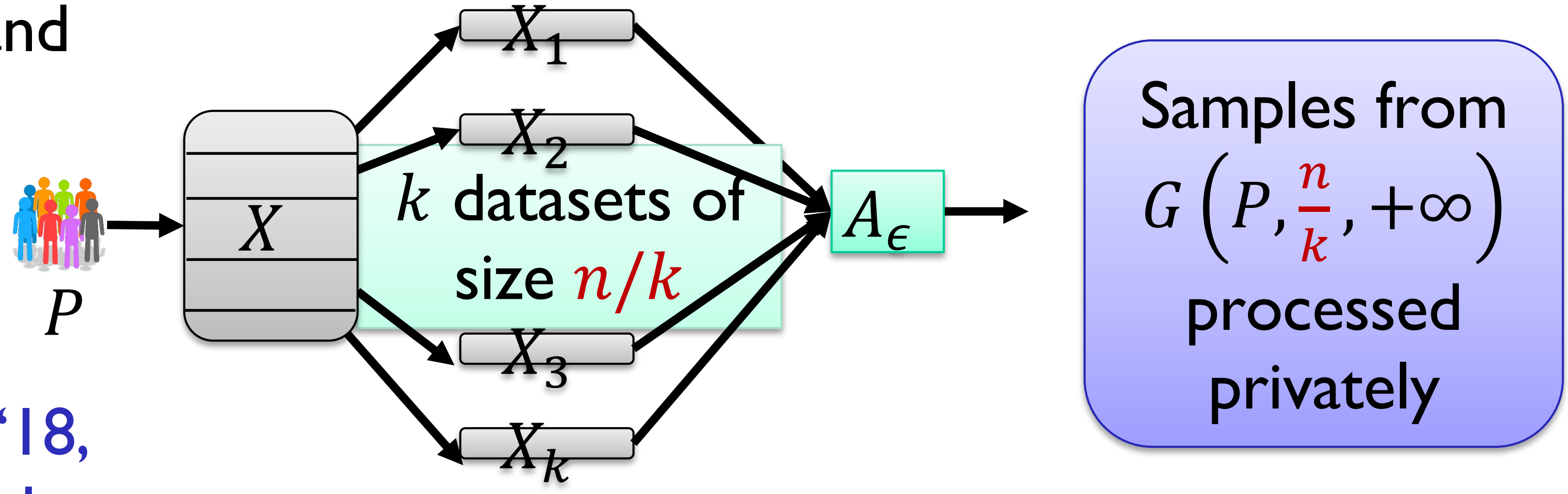
Bootstrap samples of same size (smaller ϵ) [Brawner-Honaker 18]



Samples from $G\left(P_X, n, \frac{\epsilon}{\sqrt{k}}\right)$ processed nonprivately

Direct Estimation of Sampling Distribution

Subsample and aggregate (smaller n) [NR**S**'07, **S**'11, Evans, King '18, Covington, He, Honaker, Kamath '21]



- Idea:
  - Assume a specific form for $G\left(P, \frac{n}{k}, +\infty\right)$ [e.g. Gaussian, χ^2]
  - Focus on estimation for distributions of that form
- Simple and sound ☺
- Highly specific and data-hungry ☹

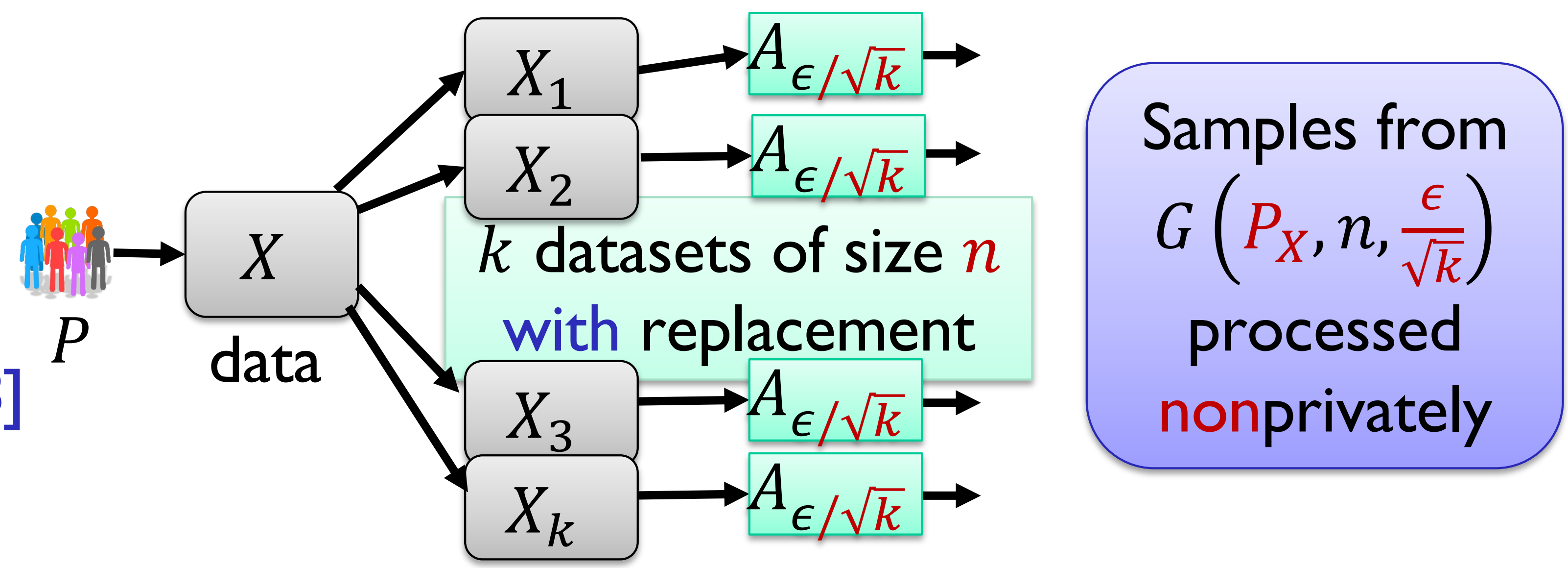
Direct Estimation of Sampling Distribution

- Booststrap theory suggests

$$G\left(P_X, n, \frac{\epsilon}{\sqrt{k}}\right) \approx G\left(P, n, \frac{\epsilon}{\sqrt{k}}\right)$$

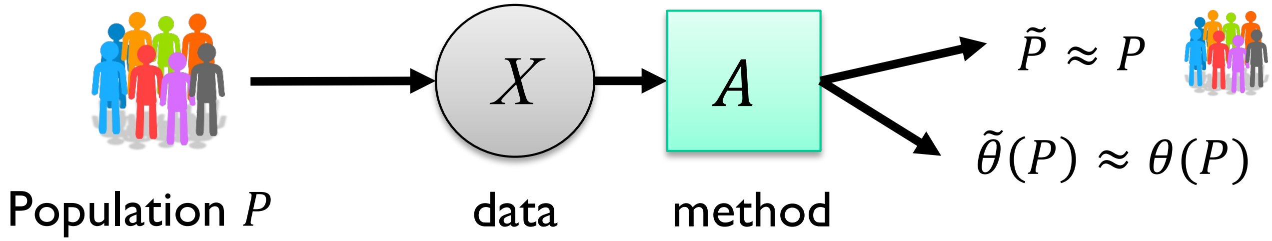
- If noise is additive, then infer mean and variance of $G(P, n, \epsilon)$

Bootstrap samples of same size (smaller ϵ) [Brawner-Honaker 18]

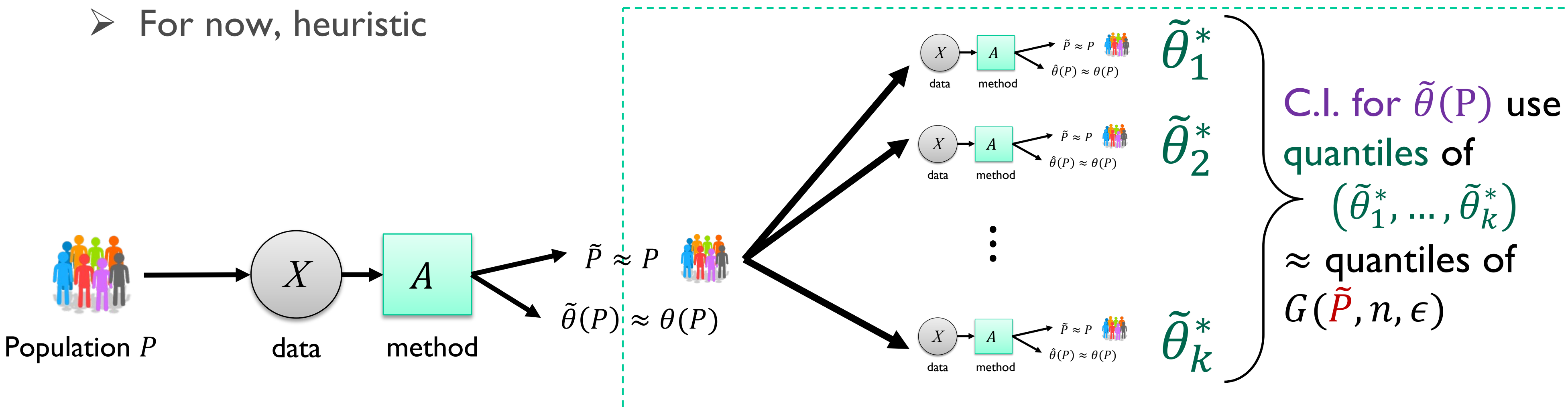


Model-based Bootstrap

"Model-based" Bootstrap



- What do we do in higher-dimensional settings?
- Many differentially private algorithms implicitly model the population
  - CDF estimators, synthetic data generators, …
- Heuristic: Use estimated model as basis for sampling distribution [Ferrando, Wang, Sheldon '21, Neunhoeffer, Sheldon, S. '22]
  - If $\tilde{P} \approx P$, then maybe $G(\tilde{P}, n, \epsilon) \approx G(P, n, \epsilon)$
  - Requires continuity of the sampling distribution
  - For now, heuristic



Example: Nonparametric Medians

- Two univariate distributions
  - Mixture of two normals
  - ADULT age data set (P= empirical distribution)
- Multivariate examples in progress

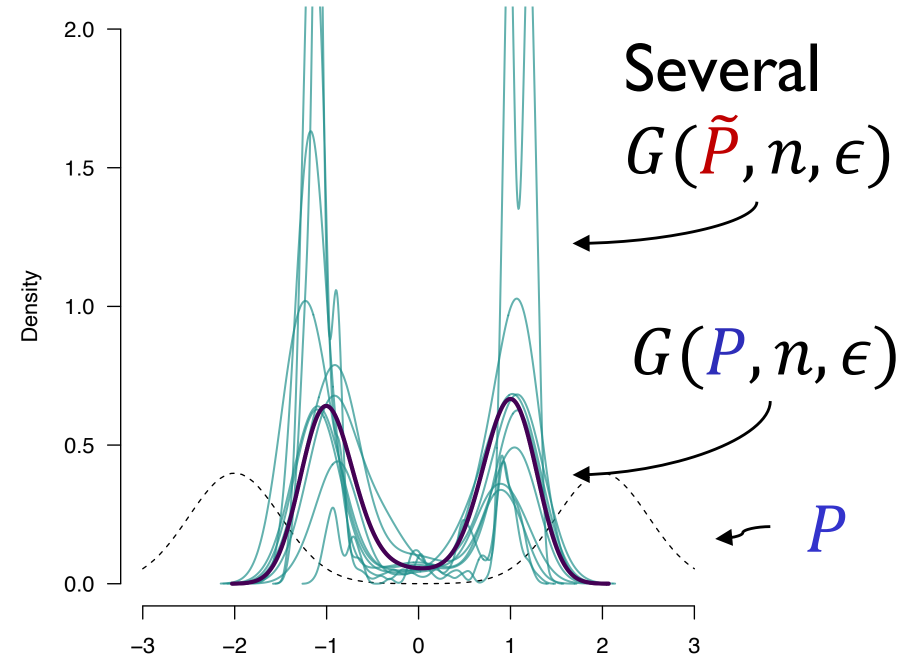
So far…

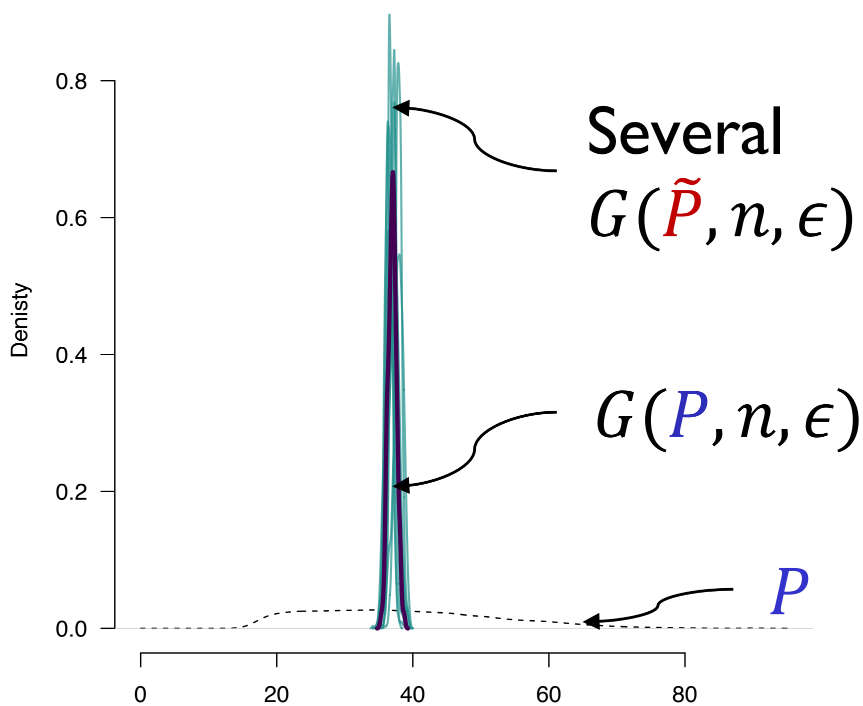
- Accurate coverage
  - But treating output naively undercovers
  - Still trying to find "bad" examples
- Narrower intervals than exact, conservative method
  [Drechsler, Globus-Harris, McMillan, Sarathy, S., 22]

Sampling distributions, n=1000

Bimodal data

- Sampling distributions $G(\tilde{P}, n, \epsilon)$ highly skewed
- Estimates $\tilde{\theta}_i^*$ are
  - Highly mean-biased in weird ways
  - Median-unbiased
-





ADULT age data

- Works well

Topics we did not cover

- Hypothesis tests and p-values
  - Basis for peer-review standards in many sciences
- Bayesian statistical approaches
- In "traditional" ML
  - Calibration of class probability estimates
  - Validity of prediction sets (set that contains correct class with high probability)
- Causal inference
- Data re-use
- Fairness to small subpopulations
- …