Simulation-based, Finite-sample Inference for Privatized Data

Jordan Awan [1]   Zhanyu Wang [1]

[1]Department of Statistics, Purdue University

Motivation and Contributions

- Private mechanisms output noisy statistics with complex sampling distributions and intractable likelihood functions. However, the privacy mechanism and data generating model are often **easy to sample** from, enabling simulation-based, indirect inference.

We expand the repro sample method (Xie and Wang, 2022) for finite-sample inference:

- We ensure that the coverage/type I errors account for Monte Carlo errors;
- We give efficient algorithms to numerically compute confidence intervals and p-values;
- We apply it to many private inference problems and compare it to other methods.

Differential Privacy (DP)

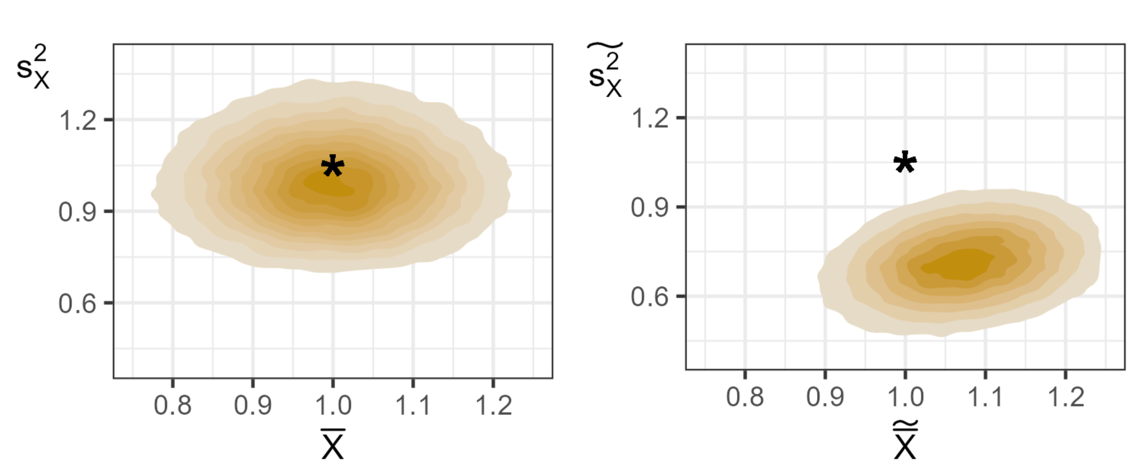
- M maps a dataset $D \in \mathcal{X}^n$ to a random variable. $d(\cdot,\cdot)$: Hamming distance.
- M is ε-DP if $\mathrm{P}(M(D) \in S) \le e^{\varepsilon}\mathrm{P}(M(D') \in S)$ for all measurable sets S and $d(D, D') \le 1$.
- M is μ-Gaussian DP (μ-GDP) if the hypothesis test $H_0: Z \sim M(D), H_1: Z \sim M(D')$ is never easier than $H_0: Z \sim N(0,1), H_1: Z \sim N(\mu, 1)$ for $d(D, D') \le 1$.

- $M(D) = g(D) + (\sup_{d(D,D')\le 1} \|g(D) - g(D')\|) \cdot \xi$ satisfies ε-DP if $\xi \sim \mathrm{Laplace}(0, 1/\varepsilon)$ (Laplace Mechanism, ℓ_1 norm), μ-GDP if $\xi \sim N(0, 1/\mu^2)$ (Gaussian Mechanism, ℓ_2 norm).
- (Objective perturbation) $\tilde{\theta} = \mathrm{argmin}_\theta(\frac{1}{n}\sum_{i=1}^n f(x_i, \theta) + c\|\theta\|_2^2 + \frac{\xi^\intercal \theta}{n})$ is ε-DP for $\xi \sim F_{f,c,\varepsilon}$.

DP statistics have a very different sampling distribution from non-private statistics.

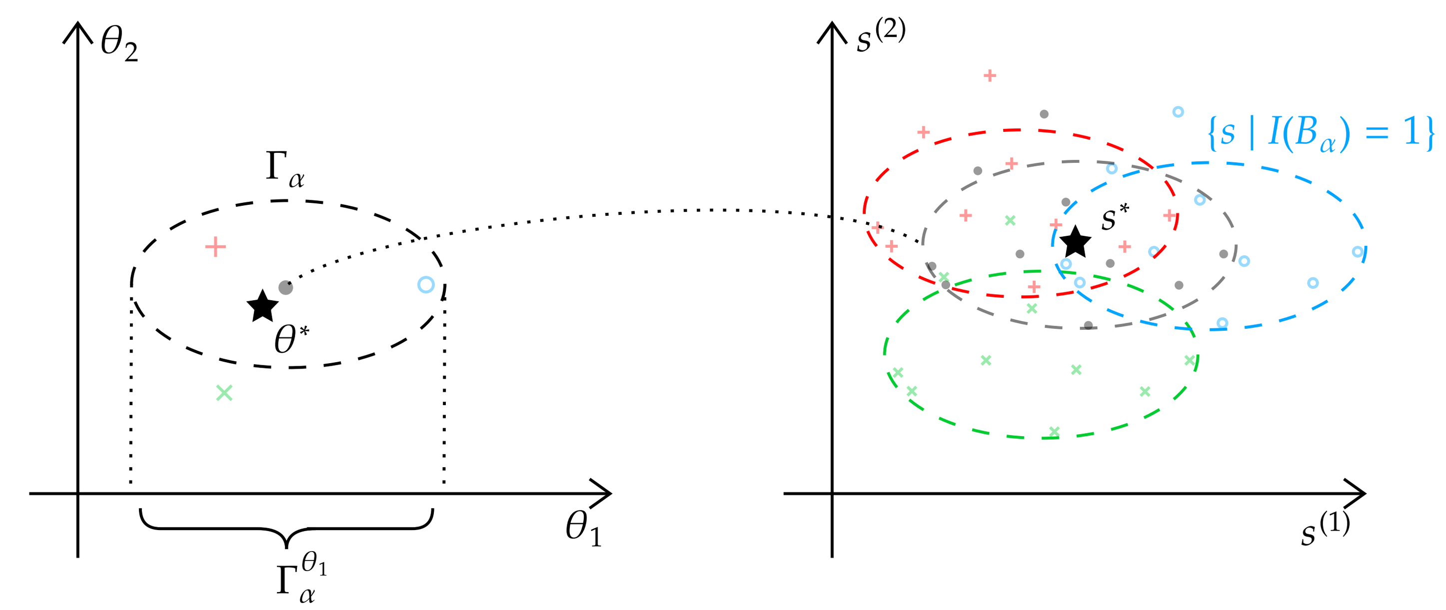
<u>Example</u>: location-scale normal

- Observe $X = (x_1, \ldots, x_n) \overset{\text{iid}}{\sim} N(\mu^*, \sigma^{*2})$ and use it to build a confidence set for (μ^*, σ^*).
- Non-private statistic: $(\bar{X} = \frac{\sum_{i=1}^n x_i}{n}, s_X^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1})$,
- Data clamping $X_c = ([x_1]_L^U, \ldots, [x_n]_L^U)$, $[x_i]_L^U = \max(\min(x_i, U), L)$.
- Private statistic: $(\widetilde{\bar{X}} = \bar{X}_c + \frac{U-L}{n\varepsilon}N_1, \widetilde{s_X^2} = s_{X_c}^2 + \frac{(U-L)^2}{n\varepsilon}N_2)$ is $(\sqrt{2}\varepsilon)$-GDP, $N_1, N_2 \overset{\text{iid}}{\sim} N(0,1)$.
- Compare the distributions of $(\widetilde{\bar{X}}, \widetilde{s_X^2})$ and $(\bar{X}, s_X^2)$. $(\mu^*, \sigma^*, n, U, L, \varepsilon) = (1, 1, 100, 3, 0, 1)$. Very different sampling distributions $\Rightarrow$ difficult to build a confidence set from $(\widetilde{\bar{X}}, \widetilde{s_X^2})$.



Simulation-based confidence sets and p-values

- Observe $s^* \overset{d}{=} G(\theta^*, u)$. G: generating equation, θ^*: true parameter, $u \sim P$: random seed.
- Generate more seeds $u_i \overset{\text{iid}}{\sim} P, i = 1, \ldots, R$. Then for each θ, simulate $s_i(\theta) = G(\theta, u_i)$.
- Use $\{s_i\}_{i=1}^R$ to build B_α to cover s^*, since $\{s_i(\theta)\}_{i=1}^R$ is close to $s^* \Rightarrow \theta$ is close to θ^*.



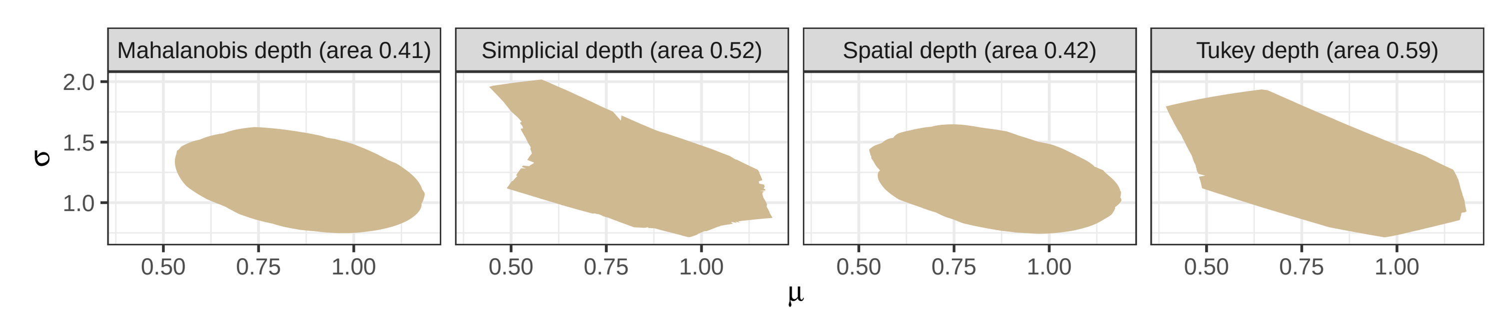
- If $B_\alpha(\theta; \{u_i\}_{i=1}^R)$ is a prediction set for $s \sim F_\theta$, $\Gamma_\alpha := \{\theta | s^* \in B_\alpha\}$ is a confidence set for θ^*. Formally, if $P_{s\sim F_\theta}(s \in B_\alpha(\theta)) \ge 1 - \alpha$ for all θ, then Γ_α is a $(1-\alpha)$-confidence set for θ^*.

Each θ in the confidence set generates a prediction set covering the observed statistic.

Theorem 1: Confidence set from simulated (repro) samples

$\underline{s} = (s^*, \{s_i(\theta)\}_{i=1}^R)$. $\{T_{(i)}^\theta\}_{i=1}^{R+1}$: order statistics of $T(s^*; \underline{s}), \ldots, T(s_R(\theta); \underline{s})$. T is permutation-invariant in $\underline{s}$. $\Gamma_\alpha(s^*, u) := \{\theta \mid T(s^*; \underline{s}) \in [T_{(\lfloor\alpha(R+1)\rfloor+1)}^\theta, T_{(R+1)}^\theta]\}$ is a $(1-\alpha)$-confidence set. If $\Gamma_\alpha(s^*, u)$ is an interval, find it using binary search and $\hat{\theta}_{\text{init}} := \mathrm{argmax}_{\theta\in\Theta} \#\{T_{(i)}^\theta \le T(s^*; \underline{s})\}$.

- **Key insights:** 1) include s in $\underline{s}$ to ensure exchangeability from permutation-invariance, and 2) get a prediction set from order statistics, similar to conformal prediction (Vovk et al., 2005).
- Most statistical depths are permutation-invariant: unusual points have lower depth. E.g., Mahalanobis depth: $T(x; X) = [1 + (x - \mu_X)^\intercal \Sigma_X^{-1}(x - \mu_X)]^{-1}$, (μ_X, Σ_X) is sample (mean,cov). Below is a comparison among different depths when $s = (\widetilde{\bar{X}} = 1, \widetilde{s_X^2} = 0.75)$.



Theorem 2: Hypothesis testing p-value

T is a depth function. $p = \frac{1}{R+1}\sup_{\theta\in\Theta_0} \#\{i | T_{(i)}^\theta \le T(s^*; \underline{s})\}$ is a p-value for $H_0: \theta^* \in \Theta_0$. For easier optimization, replace $\#\{i | T_{(i)}^\theta \le T(s^*; \underline{s})\}$ with a continuous objective function, i.e., $p = \min\left\{1, \frac{1}{R+1}\left\lfloor \sup_{\theta\in\Theta_0}\left[\#\{i | T_{(i)}^\theta \le T(s^*; \underline{s})\} + T(s^*; \underline{s})\right]\right\rfloor\right\}$.

Simulations: location-scale normal, linear and logistic regression

- Repro compared to parametric bootstrap (PB) in two tasks:
  1) constructing CI for location-scale normal μ^* and σ^* (PB by Du et al., 2020);
  2) hypothesis testing for linear regression coefficient β_1^* (PB by Alabi and Vadhan, 2022).
  PB does not have good coverage or type I error due to the bias from data clamping, i.e., $[x_i]_L^U$.

		μ^*	σ^*
Repro Sample	Coverage	0.989	0.998
	Width	0.599	0.758
Parametric Bootstrap	Coverage	0.688	0.003
	Width	0.311	0.291

Table 1. 95% CIs for location-scale normal using the same settings in the Example.

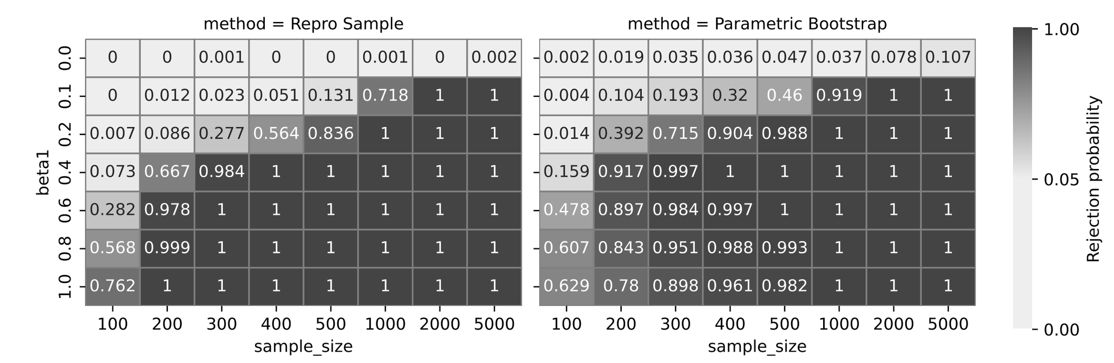


Figure 1. The rejection probability for hypothesis testing on $H_0: \beta_1^* = 0$ and $H_1: \beta_1^* \ne 0$ in $Y = \beta_0^* + X\beta_1^* + \epsilon$ where $\alpha = 0.05$.

- Repro compared to DP-CI-ERM (Wang et al., 2019) in logistic regression (for coefficient β_1^*).
  1. Repro allows for arbitrary privacy mechanisms; DP-CI-ERM requires to use a specific mechanism,
  2. Repro gives finite sample coverage guarantees; DP-CI-ERM gives an asymptotic guarantee,
  3. Repro CI is for the true parameter; DP-CI-ERM CI is for the regularized population risk minimizer.

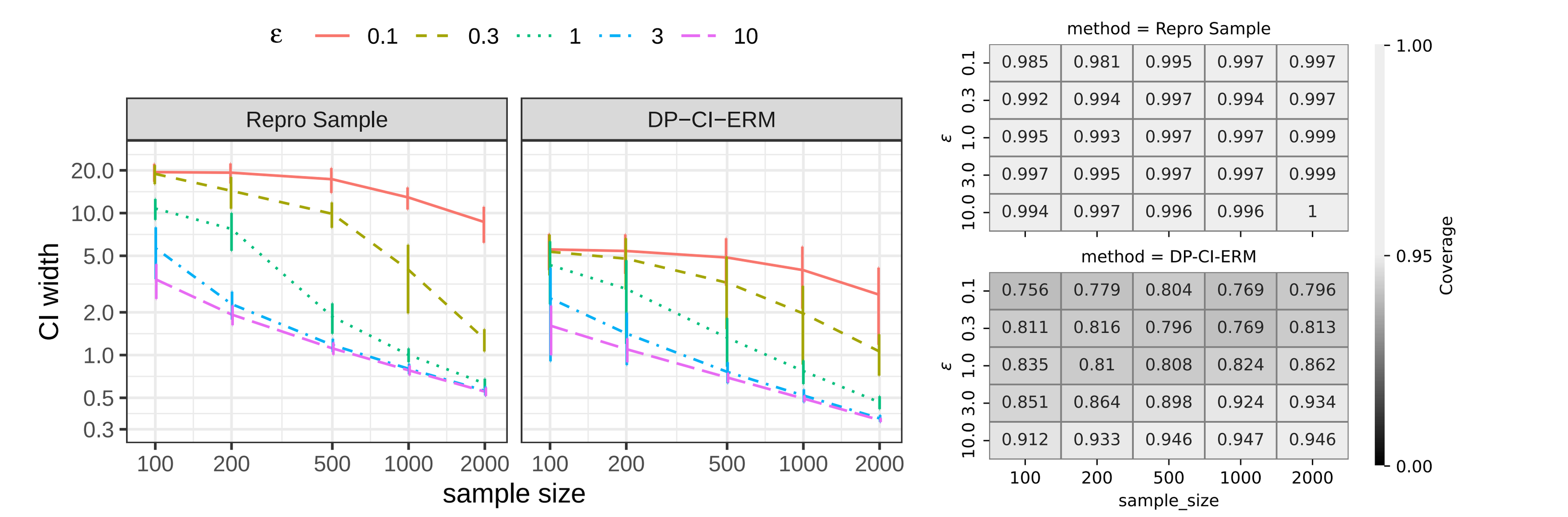


Figure 2. For logistic regression, the model is $y_i | x_i \sim \mathrm{Bern}(1/(1 + e^{-(\beta_0^* + \beta_1^* x_i)}))$, and we assume $x_i \in [-1, 1]$ modeled by $x_i \overset{\text{iid}}{\sim} 2 * \mathrm{Beta}(a^*, b^*) - 1$. We build the 95% CIs for β_1^* under $(\beta_0^*, \beta_1^*, a^*, b^*) = (0.5, 2, 0.5, 0.5)$.

Repro has larger width but better (valid) coverage compared to existing methods.

Discussions and References

- A limitation of repro is that the resulting confidence set may not be an interval.
- The ideal test statistic for use in repro would be a pivot not depending on the nuisance parameters η, which can avoid the over-coverage issue and save the optimization in η.

Alabi, D. and Vadhan, S. (2022). Hypothesis testing for differentially private linear regression. In Advances in Neural Information Processing Systems, volume 36.
Du, W., Foot, C., Moniot, M., Bray, A., and Groce, A. (2020). Differentially private confidence intervals. arXiv preprint arXiv:2001.02285.
Vovk, V., Gammerman, A., and Shafer, G. (2005). Algorithmic learning in a random world, volume 29. Springer.
Wang, Y., Kifer, D., and Lee, J. (2019). Differentially private confidence intervals for empirical risk minimization. Journal of Privacy and Confidentiality, 9(1).
Xie, M.-g. and Wang, P. (2022). Repro samples method for finite-and large-sample inferences. arXiv preprint arXiv:2206.06421.

arXiv:2303.05328
github.com/Zhanyu-Wang/Simulation-based_Finite-sample_Inference_for_Privatized_Data
wang4094@purdue.edu