Creation and Analysis of Differentially-Private Synthetic Datasets

Anne-Sophie Charest
acharest@stat.cmu.edu

IAB Workshop
Nuremberg, Germany
June 30, 2011

Outline

1. Published Results
   From Charest, Anne-Sophie (2010) "How Can We Analyze Differentially-Private Synthetic Datasets?", *Journal of Privacy and Confidentiality*: Vol. 2: Iss. 2, Article 3.

2. Work in Progress
   Relaxation of Differential Privacy

Part 1 - Published Results

Context

Setup:

- Consider the creation of synthetic datasets.
- Want to achieve differential privacy.
- Use a perturbed version of multiple imputation to do so.

Context

Setup:

- Consider the creation of synthetic datasets.
- Want to achieve differential privacy.
- Use a perturbed version of multiple imputation to do so.

Research question:

- How should analysts obtain inferences from differentially-private synthetic datasets?

Context

Setup:

- Consider the creation of synthetic datasets.
- Want to achieve differential privacy.
- Use a perturbed version of multiple imputation to do so.

Research question:

- How should analysts obtain inferences from differentially-private synthetic datasets?
- In particular, can we use the combining rules developed for multiply imputed synthetic datasets when we analyze differentially-private datasets created with multiple imputations?

The Multiple Imputation (MI) Approach

It was first suggested by Rubin (2003) to generate synthetic datasets using the framework of Multiple Imputation.

Multiple Imputation:

- Proposed to deal with non-response in surveys (Rubin, 1993).
- Write $Y = (Y_{obs}, Y_{mis})$, the observed and missing part of the data matrix for the sampled respondents.
- The analyst draws Y_{mis} from the posterior predictive distribution of $Y_{mis}|Y_{obs}$.
- After drawing M independent sets of values for Y_{mis}, we obtain M completed datasets (Y_{obs}, Y^{m}_{mis}), $m = 1, \cdots, M$.

The Multiple Imputation (MI) Approach

It was first suggested by Rubin (2003) to generate synthetic datasets using the framework of Multiple Imputation.

Multiple Imputation:

- Proposed to deal with non-response in surveys (Rubin, 1993).
- Write $Y = (Y_{obs}, Y_{mis})$, the observed and missing part of the data matrix for the sampled respondents.
- The analyst draws Y_{mis} from the posterior predictive distribution of $Y_{mis}|Y_{obs}$.
- After drawing M independent sets of values for Y_{mis}, we obtain M completed datasets (Y_{obs}, Y_{mis}^{m}), $m = 1, \cdots, M$.

To get completely synthetic datasets, we use the same idea and generate Y_{syn} from the posterior predictive distribution $Y|Y_{obs}$.

Combining Rules for MI

Key Idea: Having more than one synthetic dataset allows to estimate the variability introduced because of the SDL mechanism and account for it in our inferences.

Combining Rules for MI

Key Idea: Having more than one synthetic dataset allows to estimate the variability introduced because of the SDL mechanism and account for it in our inferences.

In Practice: Suppose we have M completely synthetic datasets and we want to estimate one parameter of interest Q. We obtain from each of the datasets an estimate q_m of Q and an estimate v_m of the variance of this estimator.

Combining Rules for MI

Key Idea: Having more than one synthetic dataset allows to estimate the variability introduced because of the SDL mechanism and account for it in our inferences.

In Practice: Suppose we have M completely synthetic datasets and we want to estimate one parameter of interest Q. We obtain from each of the datasets an estimate q_m of Q and an estimate v_m of the variance of this estimator.

Then,

$$\begin{aligned} \widehat{Q} &= \overline{q}_M \\ \widehat{Var(\widehat{Q})} &= T_M = (1 + 1/M) * b_M - \overline{v}_M \\ \text{or} \quad & T_M^* = max(0, T_M) + \frac{n_{syn}}{n}\overline{v}_M I[T_M < 0] \end{aligned}$$

where $\overline{q}_M = \frac{1}{M}\sum_m q_m$; $\overline{v}_M = \frac{1}{M}\sum_m v_m$; $b_M = \frac{1}{M-1}\sum_m (q_m - \overline{q}_M)^2$

Reiter (2003) shows that such inference is accurate.

Differential Privacy

Formal definition (Dwork, 2006):
A randomized function κ gives ϵ-**differential privacy** if and only if for all datasets B_1 and B_2 differing on at most one element, and for all $S \subseteq \text{range}(\kappa)$,

$$\exp(-\epsilon) \leq \frac{Pr[\kappa(B_1) \in S]}{Pr[\kappa(B_2) \in S]} \leq \exp(\epsilon)$$

Differential Privacy

Formal definition (Dwork, 2006):
A randomized function κ gives ϵ-**differential privacy** if and only if for all datasets B_1 and B_2 differing on at most one element, and for all $S \subseteq \text{range}(\kappa)$,

$$\exp(-\epsilon) \leq \frac{Pr[\kappa(B_1) \in S]}{Pr[\kappa(B_2) \in S]} \leq \exp(\epsilon)$$

- Smaller values of ϵ provide stronger privacy guarantees.
- For synthetic data, the randomized function κ takes as input the real dataset and generates a synthetic dataset.
- If we want M synthetic datasets, generate each with ϵ/M differential privacy.

Case Study: Beta-Binomial Synthetizer

Consider the simple case of publishing Y where $Y \sim \text{Bin}(n, p)$. The following algorithm was proposed to generated a differentially-private dataset $\tilde{Y}$ (Abowd and Vilhuber, 2008):

Case Study: Beta-Binomial Synthetizer

Consider the simple case of publishing Y where $Y \sim \text{Bin}(n, p)$. The following algorithm was proposed to generated a differentially-private dataset $\tilde{Y}$ (Abowd and Vilhuber, 2008):

Given a dataset Y, sample, for $i = 1, \ldots, m$,

$$\begin{aligned} \tilde{p}_i &\sim \text{Beta}(\alpha_1 + Y, \alpha_2 + n - Y) \\ \tilde{Y}_i &\sim \text{Binomial}(\tilde{n}, \tilde{p}_i) \end{aligned}$$

Case Study: Beta-Binomial Synthetizer

Consider the simple case of publishing Y where $Y \sim \text{Bin}(n, p)$. The following algorithm was proposed to generated a differentially-private dataset $\tilde{Y}$ (Abowd and Vilhuber, 2008):

Given a dataset Y, sample, for $i = 1, \ldots, m$,

$$\begin{aligned} \tilde{p}_i &\sim \text{Beta}(\alpha_1 + Y, \alpha_2 + n - Y) \\ \tilde{Y}_i &\sim \text{Binomial}(\tilde{n}, \tilde{p}_i) \end{aligned}$$

The parameters α_1, α_2 are deterministically chosen based on the sample size, n, and the level of differential privacy desired, ϵ.

We can interpret this synthetic data generation process as generating from a perturbed posterior predictive distribution, where we implicitly use a prior distribution of Beta(α_1,α_2) instead of a belief prior for p.

First Results

Can we use the combining rules developed for multiply imputed synthetic datasets when we analyze differentially-private datasets created with multiple imputations?

First Results

Can we use the combining rules developed for multiply imputed synthetic datasets when we analyze differentially-private datasets created with multiple imputations?

NO

First Results

Can we use the combining rules developed for multiply imputed synthetic datasets when we analyze differentially-private datasets created with multiple imputations?

NO

Bias of q_m

$$E[q_m|x] = \frac{\alpha_1 + x}{\alpha_1 + \alpha_2 + n} \neq \frac{x}{n}$$

To obtain differential privacy, we need $\alpha_1 + \alpha_2 \geq 0$.
(e.g. If $\tilde{n} = 100$, $\epsilon = 2$ (0.1), then $\alpha_j \geq 15.65$ (950)).

First Results

Can we use the combining rules developed for multiply imputed synthetic datasets when we analyze differentially-private datasets created with multiple imputations?

NO

Bias of q_m

$$E[q_m|x] = \frac{\alpha_1 + x}{\alpha_1 + \alpha_2 + n} \neq \frac{x}{n}$$

To obtain differential privacy, we need $\alpha_1 + \alpha_2 \geq 0$.
(e.g. If $\tilde{n} = 100$, $\epsilon = 2$ (0.1), then $\alpha_j \geq 15.65$ (950)).

- Avergaing over possible datasets x from a prior distribution does not in general fix this problem.
- The bias does not reduce as n increases.

Is the Bias Important in Practice?

Table: Relative Bias (in %) of $\overline{q}_M$ as estimator of p (100,000 simulations, $n = 100$, $\tilde{n} = 100$)

ϵ	p	Real data	M=1	M=2	M=5	M=10
2	0.25	0.12	23.88	53.84	80.30	90.05
2	0.50	0.05	0.05	-0.03	0.03	-0.00
250	0.25	0.01	0.05	-0.04	0.00	0.05

Same for the variance...

Table: Relative bias (in %) of T_M and T_M^* as estimators of the variance of $\overline{q}_M$. (100,000 simulations, $n = 100$, $\tilde{n} = 100$)

p	ϵ	M	Variance of $\overline{q}_M$ (x 10^{-2})	Relative Bias of T_M (%)	Relative Bias of T_M^* (%)
0.25	2	2	22.40	-44.71	54.35
0.25	2	5	6.42	-9.44	251.00
0.25	2	10	3.05	-26.61	503.95
0.50	2	2	23.57	-39.12	63.09
0.50	2	5	7.09	-6.79	225.99
0.50	2	10	3.12	-10.82	466.29
0.25	250	2	39.42	-54.29	-14.66
0.25	250	5	30.35	-38.10	-15.33
0.25	250	10	25.46	-26.51	-16.71

Note: T_M is however negative 11% to 50% of the time.

So, How Can we Analyze Such Datasets?

We could try to modify the combining rules.

Instead, we create an inferential model which takes into account the synthetic datasets generation mechanism:

$$\begin{aligned} p &\sim \text{Beta}(\gamma_1, \gamma_2) \\ y &\sim \text{Binomial}(n, p) \\ \tilde{p}_i &\sim \text{Beta}(\alpha_1 + y, \alpha_2 + n - y), \text{ for } i = 1, \ldots, M \\ \tilde{y}_i &\sim \text{Binomial}(m, \tilde{p}_i), \text{ for } i = 1, \ldots, M \end{aligned}$$

So, How Can we Analyze Such Datasets?

We could try to modify the combining rules.

Instead, we create an inferential model which takes into account the synthetic datasets generation mechanism:

$$\begin{aligned}
p &\sim \text{Beta}(\gamma_1, \gamma_2) \\
y &\sim \text{Binomial}(n, p) \\
\tilde{p}_i &\sim \text{Beta}(\alpha_1 + y, \alpha_2 + n - y), \text{ for } i = 1, \dots, M \\
\tilde{y}_i &\sim \text{Binomial}(m, \tilde{p}_i), \text{ for } i = 1, \dots, M
\end{aligned}$$

The parameters in this model can be estimated with a Metropolis-Hastings algorithm, with some Gibbs sampling steps.

We assume that α_1 and α_2 are public.

Some Results

Table: Comparison of the posterior distribution obtained with the synthetic datasets and the posterior distribution obtained with the real dataset. ($x = 30$, $n = 100$, $\tilde{n} = 100$, $\epsilon = 2$, 1000 simulations)

M	Posterior mean	Relative bias of posterior mean (%)	Variance of the posterior distribution ($\times 10^{-3}$)
1	0.311	0.76	6.30
2	0.309	0.49	7.50
5	0.312	0.85	11.70
10	0.322	1.86	15.88

True posterior distribution :
mean = 0.3039; variance = 0.0002053.

Some results (2)

Table: Comparison of the posterior distribution obtained with the synthetic datasets and the posterior distribution obtained with the real dataset. ($x = 30$, $n = 100$, $\tilde{n} = 100$, $M = 1$, 1000 simulations)

ϵ	Posterior mean	Relative bias of posterior mean (%)	Variance of the posterior distribution ($\times 10^{-3}$)	Ratio to variance from true dataset
0.1	0.485	18.09	77.07	37.54
0.5	0.365	6.14	33.75	16.44
1	0.315	1.14	15.63	7.61
2	0.311	0.72	8.18	3.98
3	0.310	0.61	6.55	3.19
250	0.312	0.83	5.81	2.83

Conclusions from Case Study

- We must create new methods for inference from differentially-private synthetic datasets.

Conclusions from Case Study

- We must create new methods for inference from differentially-private synthetic datasets.
- Directly incorporating the data generation model in the analysis seems a promising method.

Discussion

Conclusions from Case Study

- We must create new methods for inference from differentially-private synthetic datasets.
- Directly incorporating the data generation model in the analysis seems a promising method.
- There is even then a loss in utility incurred by requiring a confidentiality guarantee.

Discussion

Conclusions from Case Study

- We must create new methods for inference from differentially-private synthetic datasets.
- Directly incorporating the data generation model in the analysis seems a promising method.
- There is even then a loss in utility incurred by requiring a confidentiality guarantee.

Note that we obtain similar conclusions when considering the more general case of vectors of counts.

Part 2 - Relaxation of Differential Privacy

General Idea

- **Differential privacy controls the worst-case scenario,** i.e. the intruder knows all of the dataset except for one observation, and the released synthetic dataset is the one which gives out the most information about this particular individual in this circumstance.

General Idea

- **Differential privacy controls the worst-case scenario,** i.e. the intruder knows all of the dataset except for one observation, and the released synthetic dataset is the one which gives out the most information about this particular individual in this circumstance.

  $\rightarrow$ It is necessary to add a lot of noise to the dataset to satisfies differential privacy.

- **Proposed relaxations**
  - $\delta - \epsilon$ differential privacy

General Idea

- **Differential privacy controls the worst-case scenario,** i.e. the intruder knows all of the dataset except for one observation, and the released synthetic dataset is the one which gives out the most information about this particular individual in this circumstance.

  $\rightarrow$ It is necessary to add a lot of noise to the dataset to satisfies differential privacy.

- **Proposed relaxations**
  - $\delta - \epsilon$ differential privacy
  - $\delta - \epsilon$ probabilistic differential privacy

General Idea

- **Differential privacy controls the worst-case scenario,** i.e. the intruder knows all of the dataset except for one observation, and the released synthetic dataset is the one which gives out the most information about this particular individual in this circumstance.

  $\rightarrow$ It is necessary to add a lot of noise to the dataset to satisfies differential privacy.

- **Proposed relaxations**
  - $\delta - \epsilon$ differential privacy
  - $\delta - \epsilon$ probabilistic differential privacy

  I am considering a version of probabilistic differential privacy.

Probabilistic differential privacy

From Machanavajjhla et al. (2008):

Let κ be a randomized algorithm and let $\mathcal{S}$ be the set of all outputs of κ. Let $\epsilon > 0$ and $0 < \delta < 1$ be constants. We say that κ satisfies (ϵ, δ)-probabilistic differential privacy if for all tables D,

$$P(\mathcal{A}(D) \in \text{Disc}(D, \epsilon)) \leq \delta)$$

Probabilistic differential privacy

From Machanavajjhla et al. (2008):

Let κ be a randomized algorithm and let $\mathcal{S}$ be the set of all outputs of κ. Let $\epsilon > 0$ and $0 < \delta < 1$ be constants. We say that κ satisfies (ϵ, δ)-probabilistic differential privacy if for all tables D,

$$P(\mathcal{A}(D) \in \text{Disc}(D, \epsilon)) \leq \delta)$$

where $\text{Disc}(D, \epsilon)$ is the disclosure set of D, that is

$$\left\{ S \in \mathcal{S} \mid \exists X_1, X_2 \in \mathcal{D}, |X_1 \setminus X_2| = 1 \wedge \left| \ln \frac{P(\mathcal{A}(X_1) = S)}{P(\mathcal{A}(X_2) = S)} \right| > \epsilon \right\}.$$

Probabilistic differential privacy

From Machanavajjhla et al. (2008):

Let κ be a randomized algorithm and let $\mathcal{S}$ be the set of all outputs of κ. Let $\epsilon > 0$ and $0 < \delta < 1$ be constants. We say that κ satisfies (ϵ, δ)-probabilistic differential privacy if for all tables D,

$$P(\mathcal{A}(D) \in \text{Disc}(D, \epsilon)) \leq \delta)$$

where $\text{Disc}(D, \epsilon)$ is the disclosure set of D, that is

$$\left\{ S \in \mathcal{S} \mid \exists X_1, X_2 \in \mathcal{D}, |X_1 \setminus X_2| = 1 \wedge \left| \ln \frac{P(\mathcal{A}(X_1) = S)}{P(\mathcal{A}(X_2) = S)} \right| > \epsilon \right\}.$$

where the probability P is over the distribution of the synthetic datasets for a given observed dataset.

Relaxation

$\delta - \epsilon$ probabilistic differential privacy ensures that for any possible dataset the probability that the output synthetic dataset is in the disclosure set of level ϵ of that dataset is bounded above by δ.

Relaxation

$\delta - \epsilon$ probabilistic differential privacy ensures that for any possible dataset the probability that the output synthetic dataset is in the disclosure set of level ϵ of that dataset is bounded above by δ.

But, several of the possible datasets have very low probability of occurence.

Relaxation

$\delta - \epsilon$ probabilistic differential privacy ensures that for any possible dataset the probability that the output synthetic dataset is in the disclosure set of level ϵ of that dataset is bounded above by δ.

But, several of the possible datasets have very low probability of occurence.

Instead, we consider a version of probabilistic differential privacy where we control $P(D, \mathcal{A}(D) \mid \mathcal{A} \in \text{Disc}(D, \epsilon))$
where the probability is over the joint distribution of the observed dataset D and the synthetic dataset $\mathcal{A}(D)$.

Relaxation

$\delta - \epsilon$ probabilistic differential privacy ensures that for any possible dataset the probability that the output synthetic dataset is in the disclosure set of level ϵ of that dataset is bounded above by δ.

But, several of the possible datasets have very low probability of occurence.

Instead, we consider a version of probabilistic differential privacy where we control $P(D, \mathcal{A}(D) \mid \mathcal{A} \in \text{Disc}(D, \epsilon))$
where the probability is over the joint distribution of the observed dataset D and the synthetic dataset $\mathcal{A}(D)$.

We can write

$$P(D, \mathcal{A}(D)) = \underbrace{P(\mathcal{A}(D)|D)}_{\text{Synthesizer}} \qquad \underbrace{P(D)}_{\text{Need a prior for D}}$$

Example

Create $\tilde{Y}$ for an observed dataset Y using the beta-binomial synthesizer with $n = 5, \tilde{n} = 5, \alpha_1 = \alpha_2 = 0.5$.

Table: Values of the log of the differential privacy ratio

	$\tilde{Y} = 0$	$\tilde{Y} = 1$	$\tilde{Y} = 2$	$\tilde{Y} = 3$	$\tilde{Y} = 4$	$\tilde{Y} = 5$
$Y = 0$ vs $Y = 1$	0.747	0.463	1.099	1.578	1.997	2.398
$Y = 1$ vs $Y = 2$	0.887	0.251	0.228	0.647	1.048	1.466
$Y = 2$ vs $Y = 3$	1.099	0.619	0.201	0.201	0.619	1.099
$Y = 3$ vs $Y = 4$	1.466	1.048	0.647	0.228	0.251	0.887
$Y = 4$ vs $Y = 5$	2.398	1.997	1.578	1.099	0.463	0.747

Example

Create $\tilde{Y}$ for an observed dataset Y using the beta-binomial synthesizer with $n = 5, \tilde{n} = 5, \alpha_1 = \alpha_2 = 0.5$.

Table: Values of the log of the differential privacy ratio

	$\tilde{Y} = 0$	$\tilde{Y} = 1$	$\tilde{Y} = 2$	$\tilde{Y} = 3$	$\tilde{Y} = 4$	$\tilde{Y} = 5$
$Y = 0$ vs $Y = 1$	0.747	0.463	1.099	1.578	1.997	2.398
$Y = 1$ vs $Y = 2$	0.887	0.251	0.228	0.647	1.048	1.466
$Y = 2$ vs $Y = 3$	1.099	0.619	0.201	0.201	0.619	1.099
$Y = 3$ vs $Y = 4$	1.466	1.048	0.647	0.228	0.251	0.887
$Y = 4$ vs $Y = 5$	2.398	1.997	1.578	1.099	0.463	0.747

Say you want $\epsilon = 2$.
If you think $Y \sim \text{Bin}(5, 0.1)$,

$$\begin{aligned}\delta &= P(Y = 0, \tilde{Y} = 5) + P(Y = 1, \tilde{Y} = 5) \\ &+ P(Y = 4, \tilde{Y} = 0) + P(Y = 5, \tilde{Y} = 0) \\ &= 0.004105469\end{aligned}$$

δ-ϵ equivalence

Because of the way that probabilistic differential privacy is defined, any one randomization procedure can be described with several (technically, infinitely many) sets of pairs (δ, ϵ).

δ-ϵ equivalence

Because of the way that probabilistic differential privacy is defined, any one randomization procedure can be described with several (technically, infinitely many) sets of pairs (δ, ϵ).

We run in some sort of $\delta - \epsilon$ equivalence, and it is hard to interpret the value of both coefficients.

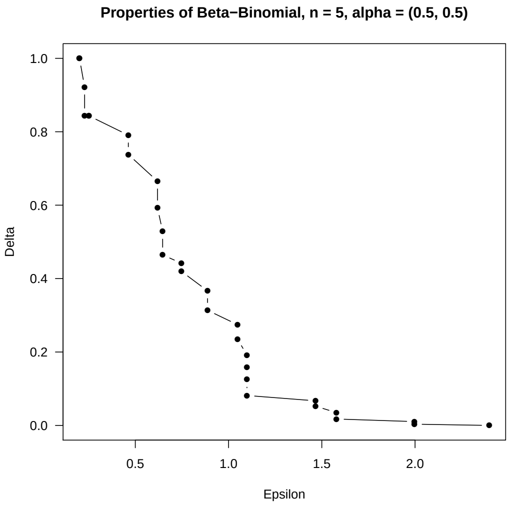
δ-ϵ equivalence

Because of the way that probabilistic differential privacy is defined, any one randomization procedure can be described with several (technically, infinitely many) sets of pairs (δ, ϵ).

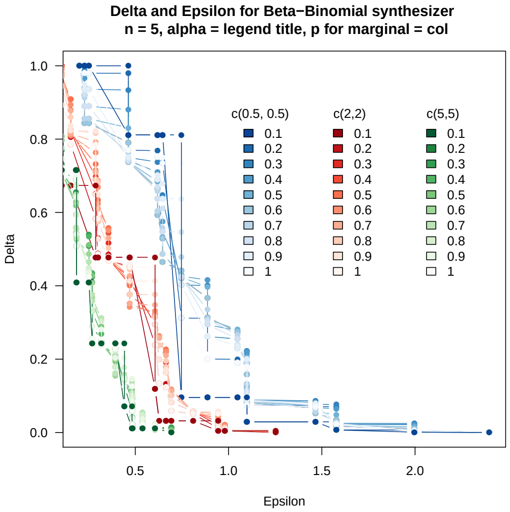
We run in some sort of $\delta - \epsilon$ equivalence, and it is hard to interpret the value of both coefficients.

Examples...

δ-ϵ equivalence (ctd)



δ-ϵ equivalence (ctd 2)



δ-ϵ equivalence (ctd)

Best parametrization?:

Choice 1:

$\epsilon = 0.6932$

$\delta = 0.0059$

Choice 2:

$\epsilon = 0.6061$ 13% smaller

$\delta = 0.0135$ twice as big

Which synthesizer to choose?

δ-ϵ equivalence (ctd)

Best parametrization?:

Choice 1:

$\epsilon = 0.6932$

$\delta = 0.0059$

Choice 2:

$\epsilon = 0.6061$ 13% smaller

$\delta = 0.0135$ twice as big

Which synthesizer to choose?

Choice 1 : $\alpha = 2$; Choice 2: $\alpha = 5$.

δ-ϵ equivalence (ctd)

Best parametrization?:

Choice 1:
$\epsilon = 0.6932$
$\delta = 0.0059$

Choice 2:
$\epsilon = 0.6061$ 13% smaller
$\delta = 0.0135$ twice as big

Which synthesizer to choose?
Choice 1 : $\alpha = 2$; Choice 2: $\alpha = 5$.

Impact of marginal for true dataset

Value of δ depends on choice of p for $x \sim \text{Binom}(n, p)$.
Example: $\epsilon = 0.75$

$$p = 0.1 \rightarrow \delta = 0.311$$
$$p = 0.9 \rightarrow \delta = 0.734$$

$\rightarrow$ might be necessary to have good priors for the marginal

Suggestions?
Comments?
Ideas?