Data Privacy in the Digital World

Tianen "Benjamin" Liu Dr. Nicole Dalzell

11/2/2019

What's Wrong with Our Data?

- 2 pieces of information for example: 1) there is only 1 person in a distant town A who has a certain disease; 2) Bob is from town A and checked in to the hospital
- Oftentimes, just removing sensitive information is not enough to protect the privacy of the data
- The sensitive information is identifiable when linked with another data set.
- 87% of individuals living in the US can be uniquely identified by using 3 data features: birth date, zip code, and gender

Differential Privacy - Definition in Practice

- I'm doing a survey about mental health that requires my sensitive information
- My college will release the data set for research but remove the sensitive information
- Still not private enough. My college will modify the data set
  - Captures the characteristics of the original data set while also making my information unidentifiable
  - Utility vs. privacy trade off
- To ensure unidentifiability: let the existence of one single answer make no difference on the probability of getting the released data set

Differential Privacy - Mathematical Definition

- Let I be the population whose data are collected
- d_i be the information given by person i
- $D_I = d_i | i \in I$ be the data set collected from all people in I
- Q be the privatized query run on a data set, and $R = Q(D_I)$ be the resultant modified data set released to the public.
- Ideally, since whether one person is in the data set does not impact the answers or data set released, we have

$$Q(D_{I-me}) = Q(D_I)$$

This should hold whenever, meaning the probability of $Q(D_{I-me})$ being equal to $Q(D_I)$ should be similar. Thus ϵ-differential privacy is defined as:

$$\frac{Prob(Q(D_I) = R)}{Prob(Q(D_{I\pm i}) = R)} \leq e^{\epsilon}, \text{ for small } \epsilon \geq 0$$

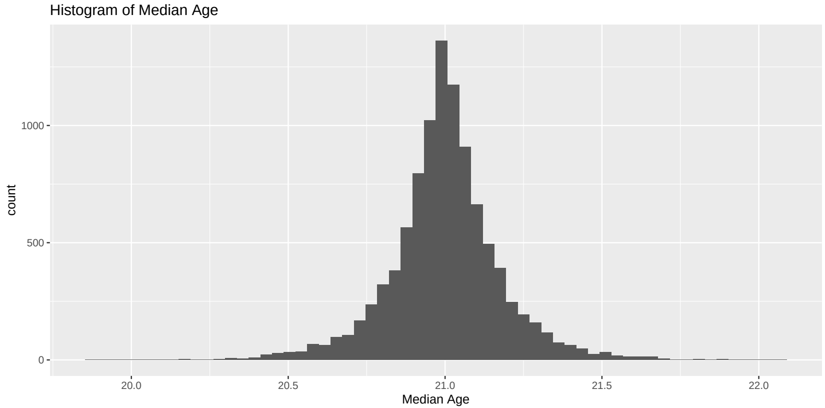
Differential Privacy Methods

Method 1: Laplacian Noise

- Add to the true answers noises drawn from the Laplace distribution: $TrueValue \pm Noise$
- Parameters: $Noise \sim Lap(\mu = 0, b = \frac{\Delta F}{\epsilon})$
- Tune the parameters to be differentially private
- Global sensitivity: $\Delta F = max_{(D_1, D_2)} |F(D_1) - F(D_2)|$, which means max difference in answers that adding or removing any individual from the data set can cause
- The released answers will have a Laplace distribution $Prob(R = x|D) = \frac{\epsilon}{2\Delta F} e^{-\frac{|x - F(D)|\epsilon}{\Delta F}}$ with ϵ-differential privacy

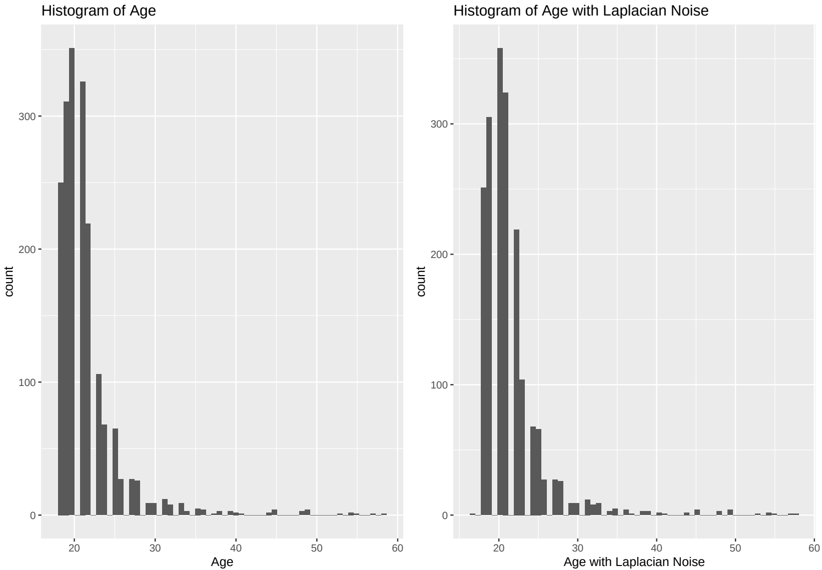
Method 1: Laplacian Noise - Binary Data & A Statistic

- Useful when releasing the count, mean, median,...
- Example: Query = Median Age in a data set. True median = 21.



Method 1: Laplacian Noise - Numeric Data

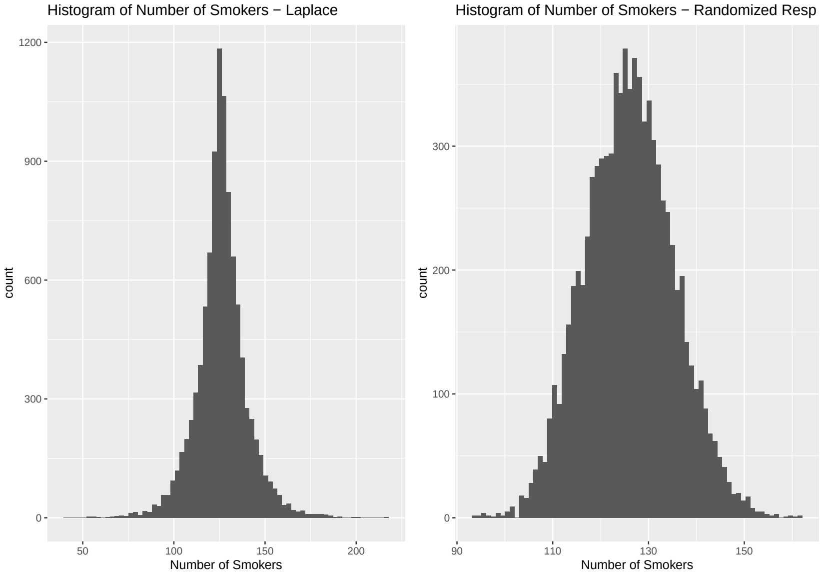
- add noise to each age and round to a whole number



Method 2: Randomized Response

- Useful when releasing the count in a binary data set
- Flip a biased coin with probability of heads α. If heads, then answer truthfully with d. If tails, flip a coin with probability of heads β and then answer with one response if heads and the other response if tails.
- We control the parameters α and β that satisfy differential privacy using an extreme case of definition: $\frac{P[Q(d_{yes},\alpha,\beta)=Yes]}{P[Q(d_{no},\alpha,\beta)=Yes]} \leq e^{\epsilon}$ from which we can get $ln(\frac{\alpha+(1-\alpha)\beta}{1-(\alpha+(1-\alpha)\beta)}) \leq \epsilon$.
- For numeric data, we can also flip one coin with probability α, and report with a Laplacian noise if tail.

Method 2: Randomized Response



Evaluation: Utility vs. Privacy

- Take differentially private output of the mean for example
- Utility: $\frac{|output - real|}{\epsilon\sqrt{n}}$, the percentage of the outputs that are useful
- Exponential Mechanism: $Pr[M_q^\epsilon(D) = o] = \frac{exp(\frac{\epsilon q(D,o)}{2\Delta q})}{\sum_{o' \in O} exp(\frac{\epsilon q(D,o')}{2\Delta q})}$, the probability of an output

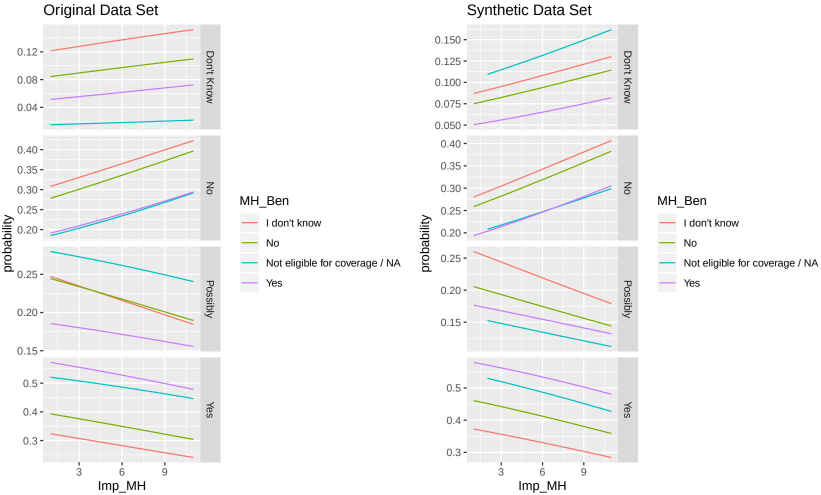
Synthetic Data

Synthesizing Open Sourcing Mental Illness Dataset

- Using `synthpop`, we can create a synthetic data set
- Control: order, method, restrictions...
- does not ensure ϵ-differential privacy

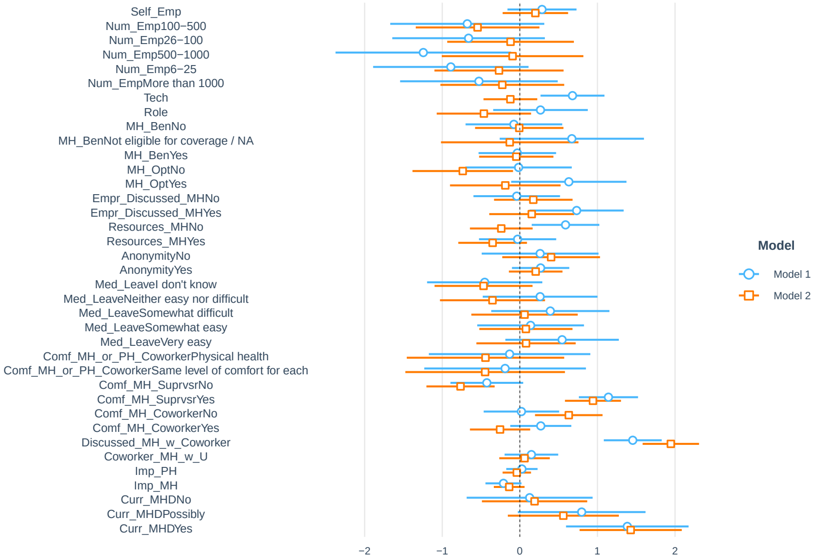
Synthesizing Open Sourcing Mental Illness Dataset

- The probability one currently has a mental heath disorder vs. how much the company values mental health



Synthesizing Open Sourcing Mental Illness Dataset

- Confidence intervals of parameters of Logistic Regressions using original (Model 1) and synthetic (Model 2) data set



Wake Forest University Healthy Minds Dataset

- Mostly categorical variables that are not necessarily correlated, thus `synthpop` does not work well
- Synthesize a subset of variables and move on to the whole data set.

Reference

- [1] Open Sourcing Mental Illness Ltd. OSMI Mental Health in Tech Survey. 2017, 2018. url: https://osmihelp.org/research.
- [2] David McClure and Jerome P. Reiter. Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data“. In: Trans. Data Privacy 5.3 (Dec. 2012), pp. 535-552. issn: 1888-5063. url: http://dl.acm.org/citation.cfm?id=2423656.2423658.
- [3] Christine Task. Privacy-preserving Datamining: Differential Privacy And Applications. June 2014.
- [4] Roberto Agostino Vitillo. Differential Privacy for Dummies. July 2016. url: https://robertovitillo.com/2016/07/29/differential-privacy-for-dummies/.