HW 1: Probability Review and Reidentification Attacks

CS 208 Applied Privacy for Data Science, Spring 2022

Version 1.0: Due Weds, Feb. 2, 5pm.

Instructions: Submit a single PDF file to Gradescope containing your solutions, plots, and analyses. Submit any code files and notebooks separately on Gradescope. Make sure to list all collaborators and references.

1. **Probability Review**

   (a) Let $S \sim \mathrm{Bin}(n, p)$ be a binomial random variable. That is, $S = X_1 + X_2 + \cdots + X_n$, where $X_1, \ldots, X_n$ are independent $\{0, 1\}$-valued Bernoulli random variables where $\Pr[X_i = 1] = p$ (i.e. coin tosses where the probability of heads is p). Calculate the standard deviation $\sigma[S]$. (Hint: recall that if X and Y are independent random variables, then $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$, where Var denotes the variance.)

   (b) Let $Z_1, \ldots, Z_k$ be independent random variables that are drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, let $M = \max\{|Z_1|, |Z_2|, \ldots, |Z_k|\}$ and let $\Phi : \mathbb{R} \to [0, 1]$ be the CDF of a standard normal $\mathcal{N}(0, 1)$ distribution. Show that for every $t > 0$

   $$\Pr[M \geq t\sigma] = 2 \cdot (1 - \Phi(t)^k) \leq 2k \cdot (1 - \Phi(t)).$$

   (c) It is known that for all $x \geq 0$, we have

   $$1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{x} \cdot e^{-x^2/2}.$$

   Using this fact and Part 1b, show that for $t = \sqrt{2 \ln k + 7}$, we have

   $$\Pr[M \geq t\sigma] < .01,$$

   where M is defined as in Part 1b.

   (d) Let $S_1, \ldots, S_k$ be independent $\mathrm{Bin}(n, p)$ random variables. The Central Limit Theorem (CLT) implies that as $n \to \infty$, each $Y_i = (S_i - \mathrm{E}[S_i])/\sigma[S_i]$ converges in distribution to a standard $\mathcal{N}(0, 1)$ normal distribution. Pretending that Y_i is actually a normal distribution (i.e. ignoring the rate of convergence), show that

   $$\Pr\left[\max_i |S_i - pn| \geq \sqrt{2 \ln k + 7} \cdot \sqrt{p(1 - p)n}\right] < .01.$$

   Comment: While we have ignored the rate of convergence in the Central Limit Theorem here, similar bounds with slightly worse constants can be proven rigorously using "Chernoff-Hoeffding Bounds," provided that $p(1 - p)n \geq c \log k$ for an appropriate constant c

   (e) Review the definitions of asymptotic notation in Section 1 notes or Section 3.1 of the Cormen-Leiserson-Rivest-Stein text.

   Fill in the table below with T (true) or F (false) to indicate the relationship between f and g. For example, if $f = O(g)$, the first cell of the row should be T.

f	g	O	o	Ω	ω	Θ
$n^3 + 2n + 1$	$5n^2 + 4$					
$6\sqrt{n \log n}$	$2n$					
$1/n$	$e^{1/n} - 1$					
2.5^n	$n^2 2^n$					
$\ln n^2$	$(\log n) + 5$					

Above and throughout the course, log denotes the logarithm base 2, and ln denotes the logarithm base e.

2. **Reidentification Attack**

In the GitHub repo,[1] you will find the Public Use Micro Sample (PUMS) dataset from the 2000 US Census `FultonPUMS5full.csv`. This is a sample from the "Long Form" from Georgia residents, which contained many more questions than the regular questionnaire, and was randomly assigned to some individuals during the decennial Census. (It has since been replaced by a continuously collected survey known as the *American Community Survey.*)

Also in that folder is the codebook file for the PUMS dataset that lists the variables available in the release. Note this is the 5% sample which means that five percent of records are randomly sampled and released. Assume that there was no disclosure avoidance techniques applied to this data.

In the style of Latanya Sweeney's record linkage reidentification attack,[2] in this problem you will propose a reidentification attack on the PUMS dataset by identifying demographic variables that, if known from another auxiliary source, could uniquely identify individuals. Note that while Sweeney used zipcodes as the geographic indicator, individuals in this Census release are identified by Public Use Microdata Areas (PUMAs) which are Census constructed geographic areas that contain at least 100,000 individuals.

(a) Create a new Jupyter notebook and read in the PUMS dataset.

(b) Determine the variables that you would match across the auxiliary source and the PUMS dataset.

i. In your notebook, plot the distribution or histograms of each of these variables.

ii. In your notebook, write a short function to calculate what fraction of individuals in the data supplied are unique, given a set of features/variables in the dataset. [3]

iii. Using your function, and your proposed reidentification attack using an auxiliary source, what is the fraction of unique individuals in the dataset you could attempt to reidentify from your proposed attack?

iv. Recall that this is a 5% sample from the full Census data. As a "back-of-the-envelope" calculation, roughly approximate what fraction of individuals would you expect to be unique if you could instead run your function on the entire Census dataset? Write a few

[1] https://github.com/opendp/cs208/tree/master/data

[2] https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1748-720X.1997.tb01885.x

[3] Note there is also a short subset of the data in the file `FultonPUMS5sample100.csv` which might be useful for testing purposes as you write your function.

sentences stating the assumptions underlying your calculation.[4] Write a few sentences stating the assumptions underlying your calculation. Your logic is more important than the accuracy of the number itself.

v. Assuming your answer from part (iv.) is lower than part (iii.), what does that mean? That is, if your proposed attack reidentifies a unique individual from the sample data, *what is the probability that you have actually reidentified the correct individual?* In a sentence or two, discuss any implication for privacy protections that are afforded by random sampling.

[4] Hint: There are many ways to go about this, either analytically with some simplifying assumptions, or numerically with a simulation. Analytically, if an individual has a p chance of being unique among N individuals, then think about what assumption you'd make to be able to say they have a p^k chance of being unique among kN individuals. Numerically, you could instead plot the value your function from part (iii.) gives you as you use subsamples of the available data and increase the sample size up to the current size of the data, and then try to project that curve out to where it would be with 20 times that amount of data.