HW 1: Probability Review and Reidentification Attacks

CS 2080 Applied Privacy for Data Science, Spring 2025

Version 2: Due Friday, Feb. 7, 5pm.

Instructions: Submit a PDF file containing your written responses as well as a zip file with your code in their respective assignments on Gradescope. Read the section "Collaboration & AI Policy" in the syllabus for our guidelines regarding the use of LLMs and other AI assistance on the assignments.

1. **Probability Review**

   (a) Let $S \sim \text{Bin}(n, p)$ 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* $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]$*, 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 $t > 0$

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

   (c) Now show that for every $t > 0$,

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

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

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

   Using this fact and Parts 1b and 1c, 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.

   (e) Let $S_1, \ldots, S_k$ be independent $\text{Bin}(n, p)$ random variables. The Central Limit Theorem (CLT) implies that as $n \to \infty$, each $Y_i = (S_i - \text{E}[S_i])/\sigma[S_i]$ 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 in the CLT[1]), show that

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

(f) 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
$n^2 + 3n + 7$	$10n^3 + 5n$					
$\log(n^{\sqrt{n}})$	$4\sqrt{n \log n}$					
$n + 2\log n$	n					
3^n	$n^3 2^n$					
$\log(n^3 + 1)$	$(\log n) + 10$					

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,[2] 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,[3] 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. For instructions on setting up a programming environment, installing Jupyter, and running your first notebook, see the section 0 notes. It is also fine if you prefer to work on Google Colab or other python IDEs.

[1] 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

[2] `https://github.com/opendp/cs208/tree/main/spring2025/data`

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

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

i. Write a function that takes in a dataset and a set of features/variables for that dataset, and returns the fraction of individuals in the dataset who are unique with respect to the specified variables. [4]

ii. 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?

Note on the auxiliary source: You do not need to find a specific external dataset for the auxiliary source. You could simply explain what is the auxiliary knowledge that you need as an adversary to make the reidentification attack successful by:

- Providing a list of three potential auxiliary sources.
- Arguing how the auxiliary knowledge needed for your attack could be found in these sources, which could simply be suggesting that a certain set of variables and individuals are likely to be present in the auxiliary sources.

iii. 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 sentences stating the assumptions underlying your calculation.[5] Your logic is more important than the accuracy of the number itself.

[4] 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.

[5] 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 roughly 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.

Codebook for Census PUMS 5 Percent CS208 Datasets

`X.1`	Deprecated, removed from dataset
`state`	The US State of residence.
`puma`	The Public Use Microdata Area, a Census constructed region of about 100,000 persons.
`jpumarow`	Deprecated, removed from dataset
`serialno.household`	Deprecated, removed from dataset
`sex`	0: Male, 1: Female.
`age`	Age in years.
`educ`	1: No schooling completed, 2: Nursery school to 4th grade, 3: 5th grade or 6th grade, 4: 7th grade or 8th grade, 5: 9th grade, 6: 10th grade, 7: 11th grade, 8: 12th grade, no diploma, 9: High school graduate, 10: Some college, but less than 1 year, 11: One or more years of college, no degree, 12: Associate degree, 13: Bachelor's degree, 14: Master's degree, 15: Professional degree, 16: Doctorate degree.
`income`	Person's total income.
`latino`	0: Not Hispanic or Latino, 1: Hispanic or Latino.
`black`	0: Not Black or African American, 1: Black or African American, alone or in combination with one or more other races.
`asian`	0: Not Asian, 1: Asian, alone or in combination with one or more other races.
`married`	0: Presently married, not separated, 1: Widowed, divorced, separated, never married.
`divorced`	0: Married or not married but not divorced, 1: Divorced and not remarried.
`uscitizen`	0: Not a citizen of the United States, 1: Citizen of United States.
`children`	0: No own minor children living in residence, 1: Lives with own minor children.
`disability`	0: Without a disability, 1: With a disability (sensory, physical, mental)
`militaryservice`	0: No military service, 1: Past or present active duty service, or training for reserves or National Guard.
`employed`	0: Unemployed or not in labor force, 1: Employed, including armed services.
`englishability`	0: Spoken English ability is "First Language", "Very Well" or "Well", 1: Spoken English ability categorized as "Not Well" or "Not at all".
`fips`	Federal Information Processing Standards County Code.