Statistics of Big Data, Fall 2023-24

Class notes 2

A motivating application: Genome-Wide Association Studies (GWAS)

The human genome can be thought of as a word of length 3×10^9 in a 4-letter alphabeth $ACGT$. Each person has two copies of this word, and the two copies are similar but not identical (similarly for copies from different individuals). A typical number cited is that two copies of the genome are about 99.9% identical, meaning several million letters different ($0.1\% = 3 \times 10^6$). A common representation of an individual genome (two copies) is as a *ternary* vector of length 3×10^9: $x_j \in \{0, 1, 2\}$, $j = 1 \dots 3 \times 10^9$.

GWAS basic idea: for a number of people ($n = 1000$ traditionally, today $n = 10^5$ is common), measure their “entire” genome and also a property (phenotype) of interest that has a heritable component, such as height or whether they have diabetes. Then look for statistical connections between each point in the genome and the pheonytpe, that is test:

$$H_{0i}: \text{ point (or region) } i \text{ has no statistical connection to the phenotype }, \; i = 1, \dots, 3 \times 10^9.$$

If we reject some of the null hypotheses we learned something about the genetics of this phenotype. In this naive view we have $p = 3 \times 10^9$ tests with $n \approx 1000$ observations — very wide data.

A few more details that will be important in our further consideration of this problem:

1. The genome has a unique and interesting correlation structure called *linkage disequlibrium (LD)* that is related to how inheritance actually works: Points in the genome that are physically close tend to be co-inherited and therefore are highly correlated (that is, two chromosomes that have the same letter in point i will tend to have the same letter in point $i + k$ for small $|k|$). In contrast, far away points in the genome are uncorrelated under some assumptions, or more weakly correlated under others.

2. Measuring all 3×10^9 in the genome is called *sequencing*. However because of the combination of only 0.1% difference and LD, it is considered that the entire “common” diversity of the genome can be represented by measuring much fewer locations, typically about 10^6 — this is called *genotyping*.

3. Additional interesting aspects that will not be widely considered in this course (⇒ Statistical Genetics course next semester):

   - Case-control sampling and its implications for statistical modeling and testing
   - The stochastic process of mutation and recombination and their spread in the population
   - Population structure and its implications for GWAS

This GWAS problem has been widely researched, and it has some important and interesting aspects related to core areas of our course. We will spend some time on several of these:

1. GWAS as an example of wide data with $n << p$, and implications for modeling it
2. Similar but slightly different view: GWAS as an example of multiple testing with high multiplicity, and the implications for how testing should be done
3. (TODAY) The privacy issues in releasing GWAS information for scientific research — how can we preserve the maximum useful information while preserving the privacy of study participants?

Privacy in big data

The basic problem: How to collect and publish data in a way that will be both:

- Useful for valid statistical analysis and scientific research
- "Safe" in terms of reasonably protecting the privacy of the individuals whose information was collected in the study

For most of the discussion we will assume standard tabular data with n individuals, each with p pieces of information, as in GWAS.

We are looking for non-trivial privacy protection, that will be robust against:

- Smart statistical analysis by the "privacy attackers"
- Availability of additional outside information in helping to identify participants and violate privacy

Unsatisfactory but common solution 1: anonymization

It seems reasonable that releasing the information without the identifying information of the participants like name, address, etc. will protect their privacy. This has been proven to fail in GWAS: it is enough for someone to have a tiny part of someone's genome to find whether that person is in the GWAS, and then know their entire genome.
Specific example: One of the most famous public genetic databases was called HapMap, which released increasingly detailed genomes of random individuals from 2005 onwards. The only personal details were sex, age and country/US state for each individual. In 2013 a Science paper demonstrated how the identities of some individuals in HapMap can be exposed:

- The male genome has a special small genetic pieces called the *Y-chromosome* that is directly inherited from father to son. It is widely used in relative search, and so millions of people have published some of their Y information online with their name and are actually finding relatives on their father's side.
- If a relative of a HapMap individual published their Y information online, we can identify that the HapMap individual is their relative on their paternal inheritance line — so probably have the same last name.

- The combination of age, country/state and last name is sometime enough to uniquely identify a person in the phone book and other publicly available sources.

Using this approach, they were actually able to positively identify several HapMap participants, meaning they now have their non-anonymized genomes, with severe consequences.

Unsatisfactory but common solution 2: releasing summaries

Assume now our $n = 2000$ GWAS samples are made of $n_1 = 1000$ *cases* who have some disease, say Type-I diabetes, and $n_2 = 1000$ samples of healthy *controls*. All their genomes are measured at $p = 10^6$ locations. It is widely recognized that in addition to analyzing this dataset separately, releasing it to the scientific community is of great interest, for example to combine with other studies and increase power.

The summary approach amounts to releasing only two tables of size $3 \times p$, one summarizing the statistics of the cases genotypes and one summarizing the statistics of the controls.

Now assume we have a genome of a specific individual, but we don't know whether they are a case (sick) or control (healthy). Two questions arise:

- If we know that this individual was in the current GWAS, can we find out whether they are a case or control?

- If we don't know whether this individual is in the study, can we separate the three options: not in the study/case/control?

The surprising(?) result is that we can typically positively answer the two questions above: not only identify whether the individual is a case or control if in the study, but also whether they were in the study at all.

For simplicity let's now assume that the disease studied is not genetic at all (say HIV), and a slightly simpler genotype structure: binary and independent coin tosses, and of length 10^5 only, to make the problem tougher (why?). So the data is matrices $X^{(case)}_{1000\times 10^5}$, $X^{(cont)}_{1000\times 10^5}$ with $X_{ij} \sim \text{Ber}(0.5)$ all i.i.d. The released information is a pair of vectors $\hat{p}_{case}, \hat{p}_{cont} \in [0,1]^{10^5}$.

Now assume we are given a "genome" $x \in \{0,1\}^{10^5}$ and we want to see whether it is a case, control, or not in our study. Let's start from the simpler problem:

$$H_0 : x \text{ case } \Leftrightarrow\ x_j \sim \text{Ber}(\hat{p}_{case,j}) \quad \text{vs} \quad H_1 : x \text{ control } \Leftrightarrow\ x_j \sim \text{Ber}(\hat{p}_{cont,j}).$$

Of course we can switch the hypotheses. Note these are simple hypotheses (specify the entire distribution).

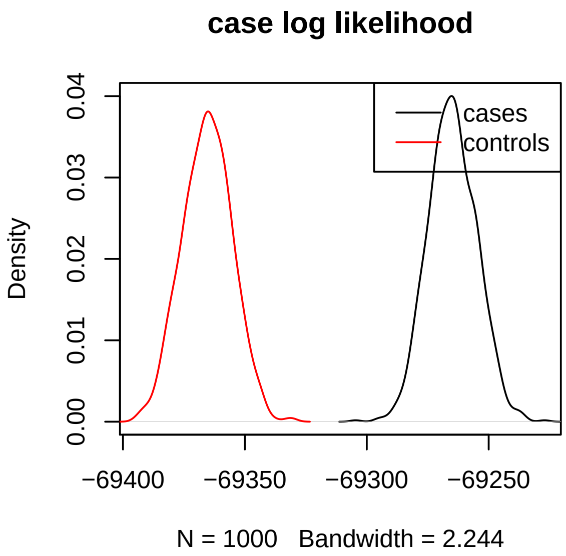
We can write the log-likelihood ratio for this problem:

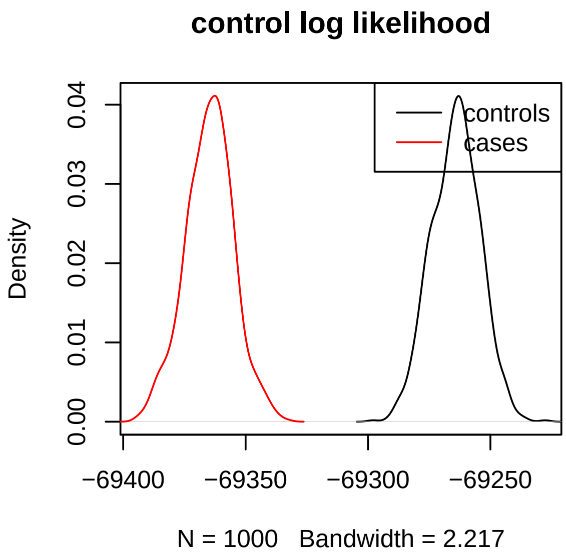
$$\lambda = \left[\sum_j x_j \log(\hat{p}_{case,j}) + (1-x_j)\log(1-\hat{p}_{case,j})\right] - \left[\sum_j x_j \log(\hat{p}_{cont,j}) + (1-x_j)\log(1-\hat{p}_{cont,j})\right].$$

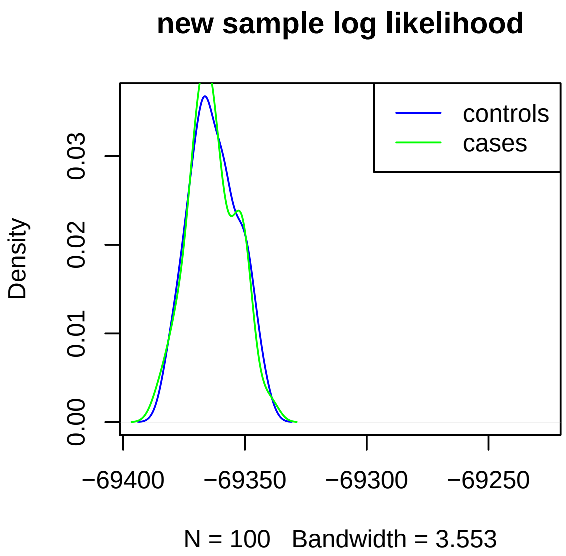
We can draw some data and present a histogram of the first sum (case likelihood) and second sum (control likelihood).

Next, we can consider the case of a third option that it is neither population, and not surprisingly the distribution of the two sums is the same and looks like the "wrong" distribution above.

The bottom line is, it turns out to be very easy to find out the hidden information about the disease status of the individual even if only summary statistics are released. Proving this rigorously will be part of the HW that will be given next week...


case log likelihood
Density
0.00
0.01
0.02
0.03
0.04
−69400
−69350
−69300
−69250
cases
controls
N = 1000   Bandwidth = 2.244


control log likelihood
Density
0.00
0.01
0.02
0.03
0.04
−69400
−69350
−69300
−69250
controls
cases
N = 1000   Bandwidth = 2.217


new sample log likelihood
Density
0.00
0.01
0.02
0.03
−69400
−69350
−69300
−69250
controls
cases
N = 100   Bandwidth = 3.553

Example of non-privacy invasion

It is important to also understand that we **can** reveal information about a person without it being an invasion of their privacy. For example, consider an imaginary study that shows that smoking hurts work productivity. After reading it the boss checks which of the employees smoke and fires them. So the fired employee is damaged by the results of the study that "taught" the manager that he is a bad employee. But assuming that the smoking status is not private, there was no violation of the employee's privacy, since his personal information was not used in any way in the study.

Hence harming X through information is not the same as harming X's privacy. There is privacy violation only if his information was used in the study and in that way it was exposed. Specifically this means "The study did not expose information on X" is not a good criterion for privacy.

The key to a proper definition: Limiting the information that can be exposed about *participants in our study* as a result of information we release. So we want to expose almost no additional information about the participants in our study, to any outside observer, whatever else they know about those participants or other people. Can we do this and still release information? It turns out we can, through the notion of Differential Privacy.

Differential Privacy

Main source for differential privacy material: The Algorithmic Foundations of Differential Privacy by Dwork and Roth

Differential privacy definition

Assume our n sampled objects (say individuals in GWAS) belong to a finite set $\mathcal{X}$ (for GWAS $\mathcal{X}$ can be the whole population). A *sample* is now $x \in \mathbb{N}^{\mathcal{X}}$, (or more simply $x \in \{0,1\}^{\mathcal{X}}$, where $x_i = 1$ if the ith element in $\mathcal{X}$ was selected for the sample and $x_i = 0$ if not). **Remarks:**

- The typical setting is where x is a sampling indicator, so $x_i = 1$ means the sample was selected and $x_i = 0$ otherwise. However, x can have different meaning, such as a indicator of who in the population has some property, in which case $x_i = 1$ is positive and $x_i = 0$ is negative. Any setting that can be put into this indicator framework is relevant.

- Using $\mathbb{N}$ allows a situation where samples were selected more than once, for example $x_i = 3$.

The sample size is:

$$n = \|x\|_1 = \sum_{i \in \mathcal{X}} |x_i|.$$

For two samples $x, y \in \{0,1\}^{\mathcal{X}}$, denote the distance between them as the number of samples that are in one and not the other:

$$d(x,y) = \|x - y\|_1 = \sum_{i \in \mathcal{X}} |x_i - y_i|.$$

Next we define the simplex on a finite set B:

$$\Delta(B) = \left\{ p \in \mathbb{R}^{|B|} \; : \; p_i \geq 0 \; \forall i \; , \; \sum_{i=1}^{|B|} p_i = 1 \right\}.$$

Definition of a randomized algorithm: Define a *draw probability function* $M : A \to \Delta(B)$, then $\mathcal{M}$ applied to A is a random algorithm with M if:

$$\mathcal{M}(a) = b \text{ w.p. } M(a)(b), \; \forall \, a \in A, \; b \in B.$$

In words, $\mathcal{M}$ gives random output that is distributed according to M.

Now we are ready to define differential privacy:
A randomized algorithm $\mathcal{M}$ applied on $\mathbb{N}^{\mathcal{X}}$ (all possible datasets) is (ϵ, δ)-differentially private if $\forall S \subseteq \mathit{Range}(\mathcal{M})(= B)$, and for any $x, y \in \mathbb{N}^{\mathcal{X}}$ such that $d(x,y) \leq 1$, we have:

$$\mathbb{P}(\mathcal{M}(x) \in S) \leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(y) \in S) + \delta.$$

In interpreting this definition, we can see the different roles of ϵ and Δ:

- To preserve $(0, \delta)$ privacy, we can release the full information of a random portion δ of the participants, since with probability $1 - \delta$ the difference between x, y is not released. So it can lead to complete privacy violation of a small portion of the participants.

- If we preserve $(\epsilon, 0)$, it means our confidence that a specific individual is in the sample cannot change by more than $\exp(\epsilon)$ depending on the results we get reported.

Thus, it is generally considered that $\delta = 0$ called $\epsilon-$privacy is the most relevant notion, and we will not consider the case $\delta > 0$ further.

Example of ϵ-privacy preservation: Randomized response

Assume we want to ask a set of people $\mathcal{X}$ whether they do something bad (say cheat on their taxes). We instruct them to do the following:

- Flip a coin (say a fair coin, but can be a general $Ber(q)$)
- If it comes out as heads, report the true answer
- If it comes out as tails, flip another fair coin, and answer yes if it comes heads, no otherwise

Thus, 50% (or more generally, q) of the answers are true and $1-q$ are randomly given as 50% true, and 50% false.

The statistician who analyzes the survey can easily conclude on the true percentage of cheaters via the unbiased estimate:

$$\hat{p}_{unbiased} = \frac{\hat{p} - (1-q)/2}{q},$$

where $\hat{p}$ is the observed positive rate in the surveys.

On the other hand, this approach guarantees $\left(\log\left(\frac{1+q}{1-q}\right), 0\right)$-differential privacy. We will show it for the specific case $q = 0.5$, where $\frac{1+q}{1-q} = 3$ for simplicity of notation.

In this setting $\mathcal{X} = \{1, \ldots, n\}$, and $x \in \{0,1\}^n$ is the identity of the true cheaters (note it is not a sampling indicator in this case). $\mathcal{M}(x)$ are the actual survey responses, and $\|x - y\| = 1$ means there is exactly one person different between x and y (cheater in one but not in the other), denote it by j and assume WLOG $x_j = 1, y_j = 0$. Since the coordinates are completely independent, and we know how the randomization works, it is easy to see that $\forall S$,:

$$\frac{\mathbb{P}(\mathcal{M}(x)_j = 0)}{\mathbb{P}(\mathcal{M}(y)_j = 0)} \leq \frac{\mathbb{P}(\mathcal{M}(x) \in S)}{\mathbb{P}(\mathcal{M}(y) \in S)} \leq \frac{\mathbb{P}(\mathcal{M}(x)_j = 1)}{\mathbb{P}(\mathcal{M}(y)_j = 1)}.$$

Given the randomization mechanism we can easily calculate:

$$\mathbb{P}(\mathcal{M}(x)_j = 1) = \frac{3}{4}, \ \mathbb{P}(\mathcal{M}(y)_j = 1) = \frac{1}{4} \implies \frac{\mathbb{P}(\mathcal{M}(x)_j = 0)}{\mathbb{P}(\mathcal{M}(y)_j = 0)} = \frac{1}{3}, \ \frac{\mathbb{P}(\mathcal{M}(x)_j = 1)}{\mathbb{P}(\mathcal{M}(y)_j = 1)} = 3.$$

The resulting $\log(3)-$differential privacy may not be a strong guarantee, in particular we know that a cheater is 3 times more likely to answer yes than no. What does it tell us about the probability of being a cheater given the answer is yes? Assuming the true proportion is r, we can write using Bayes rule:

$$\begin{aligned}\mathbb{P}(x_j = 1 | \mathcal{M}(x)_j = 1) &= \frac{\mathbb{P}(\mathcal{M}(x)_j = 1 | x_j = 1)\,\mathbb{P}(x_j = 1)}{\mathbb{P}((\mathcal{M}(x)_j = 1 | x_j = 1)\,\mathbb{P}(x_j = 1) + \mathbb{P}((\mathcal{M}(x)_j = 1 | x_j = 0)\,\mathbb{P}(x_j = 0)} = \\ &= \frac{3/4r}{3/4r + 1/4(1-r)} \leq 3r \ (\approx 3r \ \text{ if } r \text{ is small}),\end{aligned}$$

so the probability is still small, giving the person *plausible deniability.*