Statistics of Big Data, Spring 2015

Homework 1

Due date: 19 April 2015

1. **Releasing summaries of GWAS does not preserve privacy**[1]

   Assume the simplified GWAS-like setting we described in class:

   - We have 1000 cases, 1000 controls, $q = 10^5$ measures on each "genome".
   - All measures are i.i.d Bernoulli(0.5), in particular the disease has no genetic association of any kind.
   - We release summaries which are the proportion of 1's in each (or some) of the q measures for cases and controls separately. Denote these summaries by $ca_1, ..., ca_q$ for cases and $co_1, ..., co_q$ for controls.

   Now we are given a single "genome" x (binary vector of length p), and we want to find out whether this is a case, a control, or not in our study. The statistic we propose for testing this is the log-likelihood ratio statistic:

   $$\Lambda(x) = \sum_{i=1}^{q} x_i \log(ca_i) + (1 - x_i) \log(1 - ca_i) - x_i \log(co_i) - (1 - x_i) \log(1 - co_i)$$

   (a) Considering the three cases: x is a case from our study, x is a control from our study, or neither, argue that in each case the q terms in the summary are independent and identically distributed.

   (b) Explain why you expect this distribution to have a positive mean in the first case, negative in the second and zero in the third. Estimate the mean and variance of this distribution in each of the three cases empirically. You can repeat several times the following list of steps:

      i. Create two matrices of size 1000×10^5 of "cases" and "controls" as described above, and a third similar matrix of "neither"

      ii. Choose at random a few rows from each matrix, and calculate the q different values of $x_i \log(ca_i) + (1 - x_i) \log(1 - ca_i) - x_i \log(co_i) - (1 - x_i) \log(1 - co_i)$ for each of these rows.

      Having accumulated many examples of the expression for each scenario, you can empirically estimate its mean and variance.

   (c) (* Extra credit) Using a normal approximation and a Taylor expansion, derive approximate analytical expressions for the mean and variance, and compare them to your empirical estimates.

   (d) Now rely on the central limit theorem (CLT) to describe the distribution of $\Lambda(x)$ in each of the three cases, as a function of p. Now assume we are testing $H_0 : x$ is case vs. $H_A : x$ is neither case nor control at level 0.01. At what p do we get power of 0.9 according to this approximation? Explain.

[1] Our setting is similar to that discussed in the paper by Jacobs et al. (2009) pointed from the class home page, which can be used as reference for this problem.

2. **Is the Laplace mechanism useful?**

Consider a collection of patient records from a hospital, which contains $n = 10^6$ records. Two researchers ask the following queries:

- Researcher A wants to know the number n_1 of patients whose diagnosis is flu. Assume he knows from past experience that $n_1 \sim Pois(\lambda = 10^5)$ (this can be thought of as a prior distribution).
- Researcher B wants to compute the number n_2 of patients whose diagnosis is lupus. Similarly in this case $n_2 \sim Pois(\lambda = 10)$.

To answer these queries, the database uses a differentially private algorithm that adds Laplace-distributed random noise to the answer, using the Laplace mechanism to guarantee ϵ-differential privacy, at level $\epsilon = 0.1$.

(a) Explain clearly what is guaranteed for every patient by the use of this privacy mechanism with $\epsilon = 0.1$.

(b) Compute the parameter of the Laplace noise that needs to be added, and calculate its standard deviation.

(c) Compare the standard deviation of the noise to that of the prior in both cases, and also calculate explicitly the probability that the added noise will be bigger in absolute value than λ in each case. What do you conclude about the usefulness of the Laplace mechanism for these tasks?

(d) (* Extra credit) Give an expression for the posterior expected value of n_2 given the reported noisy value X_2, make it as explicit as you can. Calculate it for the case $X_2 = 20$.

(e) Assume now that researcher B states that she only cares whether $n_2 > 15$ or $n_2 \leq 15$. Thus she asks to have only a binary variable Y_2 returned.

i. Prove directly from the definition of differential privacy that reporting the right answer with probability that is $\exp(\epsilon)$ times the probability of reporting the wrong answer preserves ϵ-differential privacy.

ii. The researcher now proposes to use an exponential mechanism u with $u(n_2, Y_2) = 1$ if Y_2 is "correct" about which side of of 15 the value of n_2 is on, and $u(n_2, Y_2) = 0$ otherwise. Calculate the sensitivity of u which we denoted as δ_u in class (the maximum change in u from changing one record in the database).
Hint: If you get a trivial answer, you are on the right track...

iii. Conclude that this exponential mechanism reports the correct value as Y_2 with probability that is $\exp(\epsilon/2)$ times higher than the probability it reports the wrong value. What do you conclude about the usefulness of this exponential mechanism?