MATH 210, PROBLEM SET 5

DUE BY E-MAIL TO HAO ZHANG BY 5 P.M. APRIL 14.

Email of Hao Zhang: zhangphy@sas.upenn.edu

1. Minimizing covid-19 tests

This problem is a variation on Example 5.2.7 in deGroot's book. Suppose that we would like to test some large number N of members of a population for covid-19. The idea is that if the probability p that a person has the disease is small, one can usually find out which of a group of N people have the disease using considerably less than N tests. This is highly relevant at the present time, in view of the shortage of test kits. In fact, this method is just now coming into use against covid-19: see

https://youtu.be/vxs11ryS9Dg

Suppose that the probability that a person chosen at random will have covid-19 is some number p, where $0 \leq p \leq 1$. Let S be the set of all ways one can choose an ordered list of N people from the population. For $i = 1, \ldots, N$, let

$$X_i : S \to \{0, 1\}$$

be the Bernoulli random variable which for a given $s = (s_1, \ldots, s_N)$ returns $X_i(s) = 1$ if s_i has covid-19 and returns $X_i(s) = 0$ if s_i does not have covid-19. We'll assume that the X_i are independent random variables, each with density function

$$f_{X_i} = f : \mathbb{R} \to \mathbb{R} \quad \text{defined by} \quad f(1) = p \quad \text{and} \quad f(0) = 1 - p.$$

The naive way to find which elements of the list represented by s have covid-19 is to evaluate $X_i(s)$ for $i = 1, \ldots, N$. In other words, one tests every person on the list individually, requiring N tests. Here is a more efficient method when p is small.

1. Fix an integer $m \geq 1$ which we will assume divides N. We divide $\{1, \ldots, N\}$ into N/m disjoint subsets A_j of size m, where $j = 1, \ldots, N/m$. Explain why $Z_j = \sum_{i \in A_j} X_i$ is a random variable $Z_j : S \to \mathbb{R}$ which represents the number of people $\{s_i : i \in A_j\}$ which have covid-19.
2. Explain why Z_j is a binomial random variable with parameter m and p. In other words,
$$\text{Prob}(Z_j = k) = \binom{m}{k} p^k (1-p)^{m-k} \quad \text{for} \quad 0 \leq k \leq m.$$
3. Explain why we can determine whether $Z_j(s) > 0$ for a given $s = (s_1, \ldots, s_N) \in S$ with just one test by combining samples taken from the $s_i \in A_j$ into one sample and then testing this sample. Show that the random variable Y_j defined by $Y_j(s) = 1$ if $Z_j(s) > 0$ and $Y_j(s) = 0$ if $Z_j(s) = 0$ is a Bernoulli random variable with parameter $1 - (1-p)^m$. In other words, show
$$\text{Prob}(Y_j = 0) = (1-p)^m \quad \text{and} \quad \text{Prob}(Y_j = 1) = 1 - (1-p)^m.$$

4. Show that $Y = \sum_{j=1}^{N/m} Y_j$ is the random variable defined by letting $Y(s)$ for $s = (s_1, \ldots, s_N)$ be the number of A_j such that some $s_i \in A_j$ tests positive for covid-19. For each j such that $Y_j(s) = 1$, we go ahead and test all $s_i \in A_j$ for covid-19. Explain why this amounts to doing $m \cdot Y(s)$ tests in addition to the N/m tests needed to determine $Y_1(s), \ldots, Y_{N/m}(s)$. Show that the expected number of tests is then
$$N/m + mE(Y) = N/m + m\sum_{j=1}^{N/m} E(Y_j) = N \cdot (1/m + 1 - (1-p)^m).$$
Deduce from this that if there is a divisor m of N such that $1/m < (1-p)^m$, then the expected number of tests needed to determine which people represented by a given $s = (s_1, \ldots, s_N)$ are sick is less than N.

   (Comment: In example 5.2.7 of de Groot's book, he takes $p = 0.002$, $N = 1000$ and $m = 100$. Then $N \cdot (1/m + 1 - (1-p)^m) = 191$ is considerably less than $N = 1000$.)

5. Suppose $m^2 \geq N$. Show that for a given $s = (s_1, \ldots, s_N)$, the only way that the above procedure can require more than N tests in order to find all of the s_i which have covid-19 is for $Y_j(s)$ to be 1 for $j = 1, \ldots, N/m$. What is the probability more than N tests will be needed as a function of N, m and p?

Extra Credit: What can you say about real numbers $m > 0$ where the function
$$N \cdot (1/m + 1 - (1-p)^m)$$
has a local minimum? Can you use this and Wolfram alpha to do better than 191 for the expected number of tests needed to treat the case in which $N = 1000$ and $p = 0.002$? Here you may need to make an adjustment of the method to allow m to not be a divisor of N.