MATH 210, PROBLEM SET 4

DUE BY E-MAIL TO HAO ZHANG BY 5 P.M. APRIL 2.

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

1. Deducing the number of mutations of covid-19.

After a new virus appears, mutations of it begin to occur. One way to estimate when the new virus appeared is to study how many mutations exist. The DNA sequence of the covid-19 virus consists of a sequence of approximately 29,000 basic nucleotides, each of which is represented in shorthand by one of the four letters A, T, C, G. So far, all samples of covid-19 taken from human subjects differ in at most 7 nucleotide positions. In other words, each covid-19 sample is represented by an ordered sequence

$$s = (s_1, \ldots, s_{29000})$$

of letters $s_i \in \{A, T, C, G\}$. There is a particular subset L of 7 numbers from $\{1, \ldots, 29000\}$ such that for all pairs

$$s = (s_1, \ldots, s_{29000}) \quad \text{and} \quad s' = (s'_1, \ldots, s'_{29000})$$

of virus strains, one has

$$s_i = s'_i \quad \text{if} \quad i \notin L.$$

The same L applies to all pairs s, s' of virus strains.

1. What is the maximal number of different strains of the virus? Explain your reasoning.
2. Suppose that on further investigation, it is discovered that every strain of the virus must have exactly 2 of the positions in L occupied by each of the letters A, T and C, with one position occupied by G. Now how many different strains of the virus could there be? Explain your reasoning. (Hint: Use the multinomial theorem.)

Extra Credit: We are assuming in problem #1 that there is one set L of 7 positions such that for all pairs s, s' of virus strains, the only positions where s and s' can differ lie in L.

A. Suppose that instead, we assume that for each virus strain s, there is a set $L(s)$ of 7 positions which can depend on s such that for all virus strains s', s and s' have the same nucleotides at all positions not in $L(s)$. Can you count the maximal number of virus strains which can occur?

B. Now suppose that for each pair s, s' of virus strains, there is a just a single position $P(s, s')$ which can depend on both s and s' such that s and s' have the same nucleotides at all positions not equal to $P(s, s')$. Now can you count how many virus strains there could be?

(Comment: For more information, click on

https://www.sciencemag.org/news/2020/01/mining-coronavirus-genomes-clues-outbreak-s-origins

The fact that the DNA of all known samples differs in at most a particular set of 7 places indicates that the virus crossed into humans very recently.)

2. Bayes theorem, false positives and false negatives

Suppose that when a new medical test is applied to a person who has covid-19, the probability that test will say they have the disease is 0.95. If the test is applied to a person who does not have covid-19, the probability that the test gives a false positive result is 0.05. Assume that the test always says either that the person has covid-19 or that they don't. In other words, the test never gives back "inconclusive" as the result.

3. Suppose that in the general population, one person out of every 10,000 people has covid-19. If a person selected at random has a positive reaction to the test, what is the probability that they actually do have covid-19?
4. How does your answer to problem #3 change if in the general population, one person our of every 100 has covid-19?