Computer Science 124/224. Submit by 8 PM. Hardcopy. MS2933. May 4.

Use Python or R. Neatly type your report. In your each of your answers, describe your methods: what is your reasoning, and what is your pseudo-code. I am interested in your logic and not the correct solution. Attach any relevant equations and plots (if you have any). Attach your code at the end of your report. This report (except the code) must not be unreasonably long. No answers will be given in office hour or discussion; only useful suggestion and dumb jokes will be told. Work alone, but you are welcome to collaborate.

1. Suppose you are investigating 200 SNPs $s_1 ... s_{200}$ and their association to disease $\mathtt{d}$ with $\mathbb{P}(\mathtt{d}) = 0.10$. Suppose s_1 is the only causal SNP.

   For simplicity, assume a person has only 1 chromosome, so that 1000 people will give you 1000 chromosomes.

   (a) You randomly collect 1000 independent individuals as your sample. How many cases do you observe in your sample? Your sample here will be unbalanced; it has more controls than cases.

   (b) Suppose $\mathbb{P}(A_1|\mathtt{d}) = 0.95$ and $\mathbb{P}(A_1|\text{not } \mathtt{d}) = 0.25$, where the A_1 is the minor allele of SNP s_1. Obviously, $\mathbb{P}(A_i|\mathtt{d}) = 0.5$ and $\mathbb{P}(A_i|\text{not } \mathtt{d}) = 0.5$, where $i \neq 1$, that is because only s_1 is causal.

   Use 1 to denote existent of allele A_i at SNP s_i. Randomly assign 200 allele statuses to the cases and controls (full sample size is 1000).

   Your random person i in your hypothetical data should look as follows (of course, due to the random process, your data must not look exactly as below)

   Person i: $11111000000001000100\cdots$

   As a sanity check, among the cases in your sample, the s_1 should have about 95% of the value 1. Then, among your controls, the s_1 should have about 25% of the value 1.

   (c) As stated above, s_1 is causal SNP. But assume you do not know this fact yet. And you use all 200 SNPs. Use your hypothetical data, will you reject the null hypothesis? Use $\alpha = 0.05$.

   (d) Use your hypothetical data, find the correlations among the 200 SNPs. How many pairs have absolute correlation $|r_{ij}|$ at least 0.10. Use absolute correlation higher than 0.1, and apply greedy algorithm to find the best set of markers. Since the 200 SNPs are in fact uncorrelated, the "best set" is garbage information. In fact, optimal solution for the best set does not exist and would not make sense here.

   (e) Suppose you now know s_1 is causal, so you will only study s_1. Analytically find power. Use simulation to find power.

2. Usually, the people in your sample are likely to be related.

   (a) Create a new 1000 samples. However, make 20% cases to have correlation approximately τ among each other, but zero correlation to the other 80%. Then likewise, make 20% controls with correlation approximately τ. Set $\tau = 1$.

   Keep in mind the criteria that $\mathbb{P}(A_1|\mathtt{d}) = 0.95$ and $\mathbb{P}(A_1|\text{not } \mathtt{d}) = 0.25$, and $\mathbb{P}(\mathtt{d}) = 0.10$.

Making 20% of the cases (or controls) with $\tau = 1$ is trivial. You can easily replicate one person 20% of the times in your cases (or controls).

With this new sample, repeat `Q1c, d, e`. Do you observe any differences?

(b) Increase 20% to 50% in both cases and controls. With this new sample, repeat `Q1c, d, e`. Do you observe any differences?

(c) Extra Credit. Create new 1000 samples. Set $\tau = 0.5$. So, 20% of your cases (similarly controls) should have $r_{ij} \approxeq 0.5$. The sampling process here is not trivial; how can you create correlated individuals?

Hint: you and your parents should have correlation approximately 0.5; this statement seems obvious, but proving its validity is not easy.

With this new sample, repeat `Q1c, d, e`. What do you observe?