BF550: Fall 2020
Problem Set 6 is due by 12 pm on Monday, November 23

Problem 1
Let's pretend that we are designing a veterinary trial for a vaccine against Hepatitis E for pigs. The probability to become infected following an exposure is $p_c = 0.5$ for untreated pigs. The developer of the vaccine believes that this probability is reduced to $p_v = 0.1$ following vaccination. The control and treatment arms have the same number of pigs, N, and the statistical significance is evaluated via a permutation test. How should we choose N to ensure that we have approximately 90% chance of seeing a difference between the control and treatment arms that is significant at $\alpha = 0.05$ level?

This problem can be solved by a large number of methods. I suggest that you pursue a computational approach. That is implement classes or functions that simulate the outcomes in control and treatment arms; implement a class or function that performs a permutation test; and determine the probability of a statistically significant difference at $\alpha = 0.05$ level for a few values of N. Based on that plot, determine a suitable value of N that the problem asks for. Then perform simulations at this N and confirm that there is indeed about 90% chance that the trial passes the statistical test.

Problem 2
We are testing the expression levels of $n = 10^3$ genes for difference between control and disease. Let's assume that logarithms of the gene expression levels follow a normal distribution. We expect that about 10 genes have differences in the expression level of one standard deviation and 100 genes have differences of about 10% of the standard deviation. The rest of the genes do not show any difference in the expression level. What sample sizes do we need to find all the genes, with 100% and 10% differences in expression?

To solve this problem, you would need to create a class that simulates the results of an experiment. For each control, you can draw 10^3 Gaussian random variables with zero mean and unit variance. For each individual in the disease group, you should also draw 10^3 Gaussian random variables with unit variance. The means should be as specified above: 10 genes with mean 1, 100 genes with mean 0.1, and the rest with mean zero. The generated data should then constitute two $n \times N$ arrays: one for the control group and one for the disease group with N subjects in each. For the downstream analysis, you need to know the indexes of the 10 and 100 genes with different means.

Then you should implement a class that computes the observed differences in the mean expression of each gene and determines the corresponding p-values using the permutation test (or the t-test if you know how to do that). The p-values should be converted into q-values, and a cutoff of $\alpha = 0.05$ should be applied to determined genes with significantly different expression.

Finally, you should determine what fraction of genes with 100% difference and with 10% difference have been identified as statistically significant. Then, you can examine how these fractions depend on N and answer the main question of the problem.