ECE 3502 / CS 3501: Foundations of Data Analysis **Due Tu 2/26**

Homework 2: Hypothesis Testing

Instructions: Submit a single Jupyter notebook (.ipynb) of your work to Collab by 11:59pm on the due date. All code should be written in Python. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

You may discuss the concepts with your classmates, but write up the answers entirely on your own. Do not look at another student's answers, do not use answers from the internet, and do not show your answers to anyone.

1. Write a Python function that computes the probability function for a hypergeometric random variable, X. (See the class notes and Wikipedia page for this formula.) Your function should take inputs:

   $N =$ number of available bits to select from
   $K =$ number of available bits that are 1
   $n =$ number of bits drawn at random
   $k =$ number of bits drawn that are 1

   Your function should return $P(X = k)$. Using your function, compute the following:

   (a) Recall the "lady drinking tea" example from class. Verify that your function gives the correct values for $k = 2, 3, 4$. (See the notes for the right answers!)

   (b) You are running an internet security firm trying to catch packets sent to a server by hackers. There are 100 packets sent to the server, with 10 of them from hackers, 90 from legitimate traffic. If you sample 50 packets at random, what is the probability that you will capture all 10 packets from the hackers?

   (c) What is the chance that you will capture at least half of the hackers' packets? That is, what is $P(X \geq 5)$? **Hint:** You are going to need to sum probabilities from multiple calls to your function.

2. Here we are going to test a hypotheses about cardiac measurements from the following data: `http://www.stat.ucla.edu/projects/datasets/cardiac.dat`

   Download this data set and load it into Python. It is just a CSV file, so you can load it the same way you have in the previous homework.

   To understand what the variables mean, read the description of the data set here: `http://www.stat.ucla.edu/projects/datasets/cardiac-explanation.html`

   You want to test the hypothesis that women are more likely to have hypertension (high blood pressure) than men. Hypertension is the variable `hxofHT` (be careful, `hxofHT = 0` indicates they **do** have hypertension) and `gender` is male = 0, female = 1.

(a) What is the 2×2 contingency table for this data? The rows of your table should be `gender` and the columns should be `hxofHT`. The four entries of the table will be counts from the data. For example, one entry will count the number of people who are both women (`gender = 1`) and have hypertension (`hxofHT = 0`), etc.

(b) Using your hypergeometric probability function from the previous question, compute the probability of getting *exactly* this table.

(c) If you want to test if women have hypertension more frequently than men, what is the null hypothesis?

(d) Again, using your hypergeometric probability function, perform the Fisher exact test to get a p value for the hypothesis that women have hypertension more frequently than men. Can you "reject the null hypothesis" with the threshold $p \leq 0.05$?

3. Now we are going to do a Bayesian hypothesis test of the difference in two Bernoulli random variables, $X_1 \sim \text{Ber}(\theta_1)$ and $X_2 \sim \text{Ber}(\theta_2)$. Remember from the lecture that given data, we write k_1 and k_2 for the total number of ones observed for n_1 samples from X_1 and n_2 samples from X_2, respectively. Assuming a uniform prior for θ_1 and θ_2, remember the posterior distributions are

$$\theta_1 \mid k_1 \sim \text{Beta}(k_1 + 1, n_1 - k_1 + 1),$$
$$\theta_2 \mid k_2 \sim \text{Beta}(k_2 + 1, n_2 - k_2 + 1).$$

(a) Model women getting hypertension as X_1 and men getting hypertension as X_2. Using the same data as above, what is the posterior probability that women have a higher chance of hypertension? In other words, use sampling from the Beta distribution posteriors above to estimate the probability:

$$P(\theta_1 > \theta_2 \mid k_1, k_2)$$

You should use at least 1 million samples to estimate this. How does this compare to the Fisher exact test result that you got above?

(b) Do the same analysis for our dementia classifier from HW 1. In that example, the classifier gets 31 out of 46 dementia cases correct and 51 out of 67 healthy cases correct. Model the classifier labels for the true dementia cases as X_1, and the classifier labels for the true healthy cases as X_2. In both, we'll stick to the convention that $X_i = 0$ means the classifier labeled it as healthy and $X_i = 1$ means the classifier labeled it as dementia. What is the probability that the classifier is more likely to label dementia as dementia than it is to label healthy as dementia?