ECE 3502 / CS 3501: Foundations of Data Analysis **Due Tu 3/22**

Homework 2: Maximum Likelihood and 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 or other sources, and do not show your answers to anyone. **Cite any sources you used outside of the class material (webpages, etc.), and list any fellow students with whom you discussed the homework concepts.**

1. You are maintaining a web server and need a probability model for the traffic on it. You've settled on a model for the time in between server requests with the following probability density function (pdf):

$$p(x;\alpha) = 2\alpha x \exp(-\alpha x^2)$$

   (a) Derive an equation for the maximum likelihood estimate (MLE) for the parameter α given a sample of data $x_1, x_2, \ldots, x_N$.

   (b) Download the data file "`traffic.csv`", which contains 10,000 samples from the above distribution of hypothetical server request time intervals. Use your MLE equation from above to compute the MLE of α for this data.

   (c) Plot a histogram of the data sample. Then plot the above pdf, $p(x;\alpha)$, using your MLE found in part (b) for the α parameter.

2. 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.

3. Here we are going to test a hypotheses about cardiac measurements from a study of cardiac disease contained in the file "`cardiac.csv`".

   To understand what the variables mean, read the description of the data set here: `http://tomfletcher.github.io/FoDA/homeworks/cardiac-explanation.txt`

   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 $p \leq 0.05$?