CS 3190: Foundations of Data Analysis **Due: September 5 at 11.50pm**

Homework 1: Probability and Bayes' Rule

Instructions: Your answers are due at 11:50pm submitted on canvas. You must turn in a pdf through canvas. I recommend using latex (`http://www.cs.utah.edu/~jeffp/teaching/latex/`, see also `http://overleaf.com`) for producing the assignment answers. If the answers are too hard to read you will lose points, entire questions may be given a 0 (e.g. **sloppy pictures with your phone's camera are not ok, but very careful ones are**)

Please make sure your name appears at the top of the page.

You may discuss the concepts with your classmates, but write up the answers entirely on your own. **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.**

---

1. **[20 points]** Using the probability table below for the random variables X and Y, derive the following values:

   (a) $\textbf{Pr}(X \neq 0)$

   (b) $\textbf{Pr}(X = 0 \text{ OR } Y = 0)$

   (c) $\textbf{Pr}(Y = 1 \mid X = 1)$

   (d) Are X and Y independent? Explain why.

	$X = 0$	$X = 1$
$Y = 0$	2/3	1/8
$Y = 1$	1/6	1/24

2. **[25 points]** An "adventurous" track athlete has the following running routine every morning: He takes a bus to a random stop, then hitches a ride, and then runs all the way home. The bus, described by a random variable B, has four stops where the stops are at a distance of 1, 4, 10, and 12 miles from his house – the first three stops have probability $1/6$ of occurring. The 12 mile stop has probability $1/2$ of occurring. Then the random hitchhiking takes him further from his house a uniformly distributed number of miles on the distances -4 to 5; that is it is represented as a random variable H with pdf described

$$f(H = x) = \begin{cases} 1/9 & \text{if } x \in [-4, 5] \\ 0 & \text{if } x \notin [-4, 5] \end{cases}$$

What is the expected distance between his home and the place where he started his run?

3. **[30 points]** Consider independently rolling two fair die D_1 and D_2; each has a probability space of $\Omega = \{1, 2, 3, 4, 5, 6\}$ which each value equally likely.

   (a) What is the probability that $D_2 - D_1 = 2$?

   (b) What is the expected value of $D_2 + D_1$? Hint: Try using linearity of expectation.

4. **[25 points]** This problem has a data set D1.csv, which is available in Canvas in the assignment description as well as in the files folder [homework/HW1]. Each row of the data set is a realization of a random vector (X_i, Y_i) where $Y_i = X_i(1 - X_i) + \epsilon_i$ where X_i is uniformly distributed on $[0, 1]$ and ϵ_i is normally distributed with mean 0 and variance 1/1000 and is independent of X_i. The rows are generated independently of one another. The first column gives the realizations of the X values. The second column gives the realizations of the Y values.

   (a) Use Python to load $D1.csv$ and compute and report the pearson sample correlation r_{xy}

   (b) Make a scatter plot of y on the vertical axis and x on the horizontal axis for D1.csv. Show the scatter plot.

(c) Without doing any mathematical calculations, and just using the plot you made for guidance, what do you think is probably the relation between $\textbf{Var}(Y_i)$ and $\textbf{Var}(Y_i|X_i = .5)$? Note that the value $\textbf{Var}(Y_i|X_i = .5)$ is known as the conditional variance of Y_i given we know $X_i = .5$.

(d) When the conditional variance is much smaller than the marginal variance, this is a good indication that Y_i and X_i are correlated in some way. But the correlation need not be linear. Recall that the pearson correlation is a measure of what specific kind of correlation?