# Assignment 4

CSCI2244-Randomness and Computation
Due Saturday, February 24, at 11:30PM

Some problems about random variables. For the coding parts of the problem, the dice simulation that I will present in class on Feb. 15 and post online should be helpful.

## 1 Words

Think 'hypergeometric distribution'. (You don't need this fancy term to solve this problem!)

Fifteen of the 100 most common words in English begin with the letter 't'. (Fun fact courtesy of Wikipedia.) You select seven distinct words from this list of 100 words at random. Let $X$ be the random variable denoting the number of words in your selection that begin with 't'. Formally, $X$ is a function from the set of all sets of seven of these words into the real numbers.

1. What is $X(\{\text{and, but, them, for,what,to, he, she}\})$?
2. What is $P(X = 3)$? What is $P(X = 0)$?
3. Let $k$ be a nonnegative integer. What is $P(X = k)$? (Don't just list the numerical values; give a formula or, at most, two formulas, that give the value of $P(X = k)$ for all nonnegative integers $k$.
4. What is $P(X \geq 3)$?
5. Draw the graph of the PMF of $X$ as a stem plot (use matplotlib to do this).
6. What is $E(X)$? You can compute this directly using the results above and the definition of expected value. Alternatively, you can use the very simple formula for the expected value of a random variable with hypergeometric distribution.

7. Write a function that returns a value in the range $0, \ldots, 7$ with the same distribution as $X$. (You do not need to simulate the card-drawing experiment—it is more efficient and simpler to generate a random value in the interval $[0, 1]$, then use the inverse of the CDF to generate the values.

8. Use your function to generate one thousand values with this distribution, compute their average, and compare the result to the expected value as computed above.

# 2 ..and more words.

Think 'binomial distribution' and 'Poisson approximation to binomial distribution'.

Approximately $0.27\%$ of the words in a very long English novel begin with the letter 'q'. Note that this is not the number of separate vocabulary items that begin with 'q', as in the first problem, because the a single word will typically have many different occurrences in the book.

The game consists of opening the book 1000 times to a random page, and selecting a random word from the page. The player gets one point if the selected word begins with the letter 'q'. Observe that the maximum number of points a player can receive is 1000 (although this is *extremely* unlikely), and the minimum is 0 (an event whose probability you will compute below). Let $Y$ be the random variable denoting the number of points a player of this game receives.

1. Write an exact expression for $P(Y = k)$.

2. If $k$ is relatively small, this is a case where the random variable can be well-approximated by a Poisson random variable. Write an approximate expression for $P(Y = k)$, using the Poisson distribution.

3. What are $P(Y = 0)$ and $P(Y = 1)$, both exactly, and in the Poisson approximation?

4. Write a function that simulates a game and returns the value of $Y$.

5. Plot the PMF of $Y$ at values less than or equal to 10, in three different ways: Use the exact formula, the Poisson-approximated formula, and the simulation from the preceding problem, run 1000 times. Plot the result as three parallel stem plots with an explanatory legend. To set this up, the

bases of the three stems at $k = 1$ should be at 0.9, 1, and 1.1, or something similar. You should see a pretty good matchup.