Oishi Banerjee and Cooper Raterink
CS 109

Problem Set #4
July 15, 2020

Problem Set #4
Due: 1:00pm on Monday, July 27

With problems by Mehran Sahami and Chris Piech

Written Problems

For each problem, briefly explain/justify how you obtained your answer. In fact, most of the credit for each problem will be given for the derivation/model used as opposed to the final answer. Make sure to describe the distribution and parameter values you used where appropriate. **Provide a numeric answer for all questions when possible.**

1. The **median** of a continuous random variable having cumulative distribution function F is the value m such that $F(m) = 0.5$. That is, a random variable is just as likely to be larger than its median as it is to be smaller. Find the median of X (in terms of the respective distribution parameters) in each case below.
   a. $X \sim \text{Uni}(a, b)$
   b. $X \sim \mathcal{N}(\mu, \sigma^2)$
   c. $X \sim \text{Exp}(\lambda)$

2. Users independently sign up for two online social networking sites, Lookbook and Quickgram. On average, 7.5 users sign up for Lookbook each minute, while on average 5.5 users sign up for Quickgram each minute. The number of users signing up for Lookbook and for Quickgram each minute are independent. A new user is defined as a new account, i.e., the same person signing up for both social networking sites will count as two new users.
   a. What is the probability that more than 10 new users will sign up for the Lookbook social networking site in the next minute?
   b. What is the probability that more than 13 new users will sign up for the Quickgram social networking site in the next 2 minutes?
   c. What is the probability that the company will get a combined total of more than 40 new users across both websites in the next 2 minutes?

3. Say that of all the students who will attend Stanford, each will buy at most one laptop computer when they first arrive at school. 40% of students will purchase a PC, 30% will purchase a Mac, 10% will purchase a Linux machine and the remaining 20% will not buy any laptop at all. If 15 students are asked which, if any, laptop they purchased, what is the probability that exactly 6 students will have purchased a PC, 4 will have purchased a Mac, 2 will have purchased a Linux machine, and the remaining 3 students will have not purchased any laptop?

4. Say we have two independent variables X and Y, such that $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$. Mathematically derive an expression for $P(X = k | X + Y = n)$, where k and n are non-negative integers.

Practice with the above problems for the midterm
Finish the below problems afterwards

5. Choose a number X at random from the set of numbers $\{1, 2, 3, 4, 5\}$. Now choose a number at random from the subset no larger than X, that is from $\{1, \ldots, X\}$. Let Y denote the second number chosen.
   a. Determine the joint probability mass function of X and Y.
   b. Determine the conditional mass function of X given $Y = i$. Do this for $i = 1, 2, 3, 4, 5$.
   c. Are X and Y independent? Justify your answer.

6. Let $X_1, X_2, \ldots$ be a series of independent random variables which all have the same mean μ and the same variance σ^2. Let $Y_n = X_n + X_{n+1}$. For $j = 0$, 1, and 2, determine $\mathrm{Cov}(Y_n, Y_{n+j})$. Note that you may have different cases for your answer depending on the value of j.

7. Our ability to fight contagious diseases depends on our ability to model them. One person is exposed to llama-flu. The method below models the number of individuals who will get infected.

```
from scipy import stats
"""
Return number of people infected by one individual.
"""
def num_infected():
  # most people are immune to llama flu.
  # stats.bernoulli(p).rvs() returns 1 w.p. p (0 otherwise)
  immune = stats.bernoulli(p = 0.99).rvs()
  if immune: return 0

  # people who are not immune spread the disease far by
  # making contact with k people (up to 100).
  spread = 0
  # returns random # of successes in n trials w.p. p of success
  k = stats.binom(n = 100, p = 0.25).rvs()
  for i in range(k):
    spread += num_infected()

  # total infections will include this individual
  return spread + 1
```

What is the expected return value of `numInfected()`?

Continued on next page. . .

8. Consider the following recursive function:

```
def recurse():
    x = np.random.choice([1,2,3])   # equally likely values 1,2,3
    if (x == 1): return 3
    elif (x == 2): return (5 + recurse())
    else: return (7 + recurse())
```

   Let Y = the value returned by `recurse()`. We previously computed $E[Y] = 15$. What is $\text{Var}(Y)$?

9. You go on a camping trip with two friends who each have a mobile phone. Since you are out in the wilderness, mobile phone reception isn't very good. One friend's phone will independently drop calls with 20% probability. Your other friend's phone will independently drop calls with 30% probability. Say you need to make 6 phone calls, so you randomly choose one of the two phones and you will use that *same* phone to make all your calls (but you don't know which has a 20% versus 30% chance of dropping calls). Of the first 3 (out of 6) calls you make, one of them is dropped. What is the conditional expected number of dropped calls in the 6 total calls you make (conditioned on having already had one of the first three calls dropped)?

Coding Problem

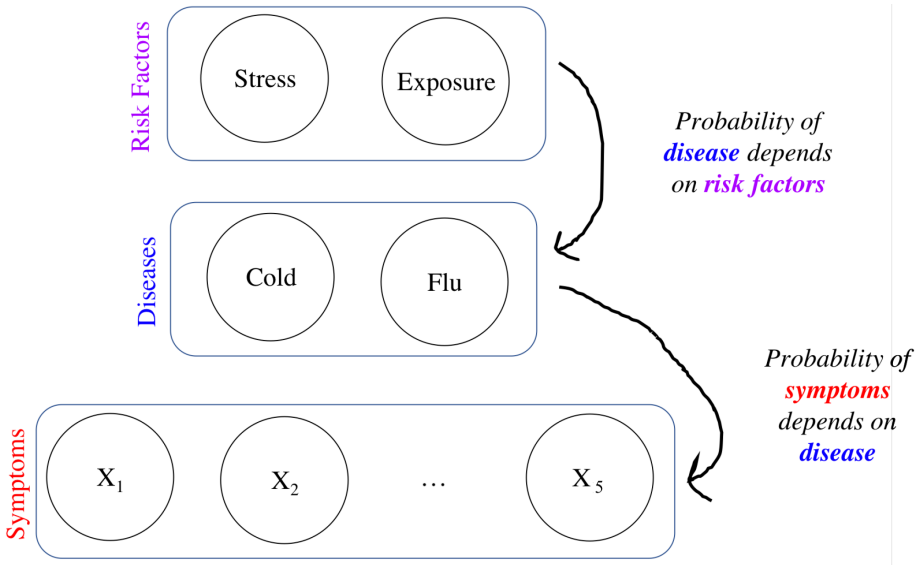
Download the starter code for this problem from the Problem Set #4 webpage. For parts (a) and (b), submit your completed file on Gradescope under "PSet 4 - Coding". Your code will be autograded. We expect you to follow these guidelines:

- Do not use global variables.
- You may define helper functions if you wish.
- Your code should not print anything.
- You should not implement the helper functions already provided to you in `probabilities.py`.

10. We are writing a WebMD program that is slightly larger than the one we worked through in class. In this program we predict whether a user has a flu ($F = 1$) or cold ($C = 1$) based on knowing any subset of 5 potential binary symptoms (e.g., headache, sniffles, fatigue, cough, fever) and a subset of binary risk factors (exposure, stress).

    The file `probabilities.py` implements the below functions. You **should not** re-define these functions as we may use different probabilities when grading.

    - The functions `probStress()` and `probExposure()` return the prior probabilities for stress and exposure, respectively.
    - The functions `probCold(s, e)` and `probFlu(s, e)` return the probability that a patient has a cold or flu, given the state of the risk factors stress (`s`) and exposure (`e`).



- The function `probSymptom(i, f, c)` returns the probability that the `i`th symptom (X_i) takes on value 1, given the state of cold (`c`) and flu (`f`): $P(X_i = 1 | F = f, C = c)$, for $i = 1, \ldots, 5$.

We would like to write code that computes the probability $P(\text{Flu} = 1 | \text{Exposure} = 1, X_2 = 1)$: the probability of flu *conditioned on observing* that the patient has had exposure to a sick friend and that they are experiencing Symptom 2 (sniffles):

a. Implement the function `inferProbFlu()` that computes the probability $P(\text{Flu} = 1 | \text{Exposure} = 1, X_2 = 1)$ using **Rejection Sampling**. This function takes one parameter `ntrials`, which is the number of observations to generate:

```
def inferProbFlu(ntrials=1000000) # P(Flu = 1 | Exposure = 1 and X2 = 1)
```

b. (Reach) Implement the function `inferProbFluExact()` that computes the probability $P(\text{Flu} = 1 | \text{Exposure} = 1, X_2 = 1)$ without using sampling.

Note: Causality implies the following: (1) risk factors are independent; (2) all diseases are independent of one another conditioned on risk factors; and (3) all symptoms are independent of one another conditioned on knowing the state of diseases.

$$P(S = s, E = e) = P(S = s)P(E = e) \tag{1}$$

$$P(C = c, F = f | S = s, E = e) = P(C = c | S = s, E = e) \cdot P(F = f | S = s, E = e) \tag{2}$$

$$P(\text{symptoms} | F = f, C = c) = \prod_j P(X_j = k_j | F = f, C = c) \tag{3}$$

"Reach" means I don't *expect* CS109 students to be able to solve the problem. But thinking about it will be useful. Show your work. If you get stuck (> 15 mins), explain what is hard and include your explanation in your writeup, not in your code.

11. **[Written, Extra Credit]** Consider a bit string of length n, where each bit is independently generated and has probability p of being a 1. We say that a *bit switch* occurs whenever a bit differs from the one preceding it in the string (if there is a preceding bit). For example, if $n = 5$ and we have the bit string 11010, then there are 3 bit switches. Find the expected number of bit switches in a string of length n. (Hint: You might find it helpful to use a set of indicator (Bernoulli) variables that are defined in terms of whether a bit switch occurred in each *position* of the string. And in case you're wondering why we care about bit switches, the number of bit switches in a string can be one indicator of how compressible that string might be—for example, if the bit string represented a file that we were trying to ZIP.)