David Varodayan
CS 109

Problem Set #4
February 5, 2020

Problem Set #4
Due: 1:00pm on Wednesday, February 19

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. A company owns 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?

2. 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. $X \sim \text{Exp}(\lambda)$

3. Let X, Y, and Z be independent random variables, where $X \sim \mathcal{N}(\mu_1, {\sigma_1}^2)$, $Y \sim \mathcal{N}(\mu_2, {\sigma_2}^2)$, and $Z \sim \mathcal{N}(\mu_3, {\sigma_3}^2)$.

   a. Let $A = X + Y$. What is the distribution (along with parameter values) for A?
   b. Let $B = 4X + 3$. What is the distribution (along with parameter values) for B?
   c. Let $C = aX - b^2Y + cZ$, where a, b, and c are real-valued constants. What is the distribution (along with parameter values) for C? Show how you derived your answer.

4. Say the lifetimes of computer chips produced by a certain manufacturer are normally distributed with parameters $\mu = 1.5 \times 10^6$ hours and $\sigma = 9 \times 10^5$ hours. The lifetime of each chip is independent of the other chips produced. What is the approximate probability that a batch of 100 chips will contain at least 65 whose lifetimes are less than 1.9×10^6 hours?

5. You roll 6 six-sided dice. How much more likely is a roll with [1 one, 1 two, 1 three, 1 four, 1 five, 1 six] than a roll with 6 sixes? Think of your dice roll as a multinomial.

6. You are testing software and discover that your program has a non-deterministic bug that causes catastrophic failure (aka a "hindenbug"). Your program was tested for 400 hours and the bug occurred **twice**.

   a. Each user uses your program to complete a three hour long task. If the hindenbug manifests they will immediately stop their work. What is the probability that the bug manifests for a given user?
   b. Your program is used by one million users. Use a normal approximation to estimate the probability that more than 10,000 users experience the bug. Use your answer from part (a).

7. 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?

Try to do the above before the midterm

And finish the below after the midterm

8. The joint probability density function of continuous random variables X and Y is given by:

$$f_{X,Y}(x,y) = c\frac{y}{x} \qquad \text{where } 0 < y < x < 1$$

   a. What is the value of c in order for $f_{X,Y}(x,y)$ to be a valid probability density function?
   b. Are X and Y independent? Explain why or why not.
   c. What is the marginal density function of X?
   d. What is the marginal density function of Y?
   e. What is $E[X]$?

9. A robot is located at the *center* of a square world that is 10 kilometers on each side. A package is dropped off in the robot's world at a point (x, y) that is uniformly (continuously) distributed in the square. If the robot's starting location is designated to be $(0, 0)$ and the robot can only move up/down/left/right parallel to the sides of the square, the distance the robot must travel to get to the package at point (x, y) is $|x| + |y|$. Let D = the distance the robot travels to get to the package. Compute $E[D]$.

10. 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.

11. 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 $\text{Cov}(Y_n, Y_{n+j})$. Note that you may have different cases for your answer depending on the value of j.

12. Consider a series of strings that independently get hashed into a hash table. Each such string can be sent to any one of $k + 1$ buckets (numbered from 0 to k). Let index i denote the i-th bucket. A string will independently get hashed to bucket i with probability p_i, where $\sum_{i=0}^{k} p_i = 1$. Let N denote the number of strings that are hashed until one is hashed to any bucket other than bucket 0. Let X be the number of that bucket (i.e. the bucket not numbered 0 that receives a string).

    a. Find $P(N = n), n \geq 1$.
    b. Find $P(X = j)$, $j = 1, 2, \ldots, k$.
    c. Show that N and X are independent.

13. Consider the following function, which simulates repeatedly rolling a 6-sided die (where each integer value from 1 to 6 is equally likely to be "rolled") until a value ≥ 3 is "rolled".

```
def roll():
  total = 0
  while(True):
    # randomInteger is equally likely to return 1, ..., 6
    roll = randomInteger(1, 6)
    total += roll

    # exit condition:
    if (roll >= 3):
      break
  return total
```

    a. Let X = the value returned by the function `roll()`. What is $E[X]$?
    b. Let Y = the number of times that the die is "rolled" (i.e., the number of times that `randomInteger(1, 6)` is called) in the function `roll()`. What is $E[Y]$?

Continued on the next page...

14. 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 the next page. . .

Coding/Written Problem

Download the starter code and data files for this problem from the Problem Set #4 webpage. For parts (a) and (b), submit your completed file on Gradescope under "PSet 4 - Coding". Parts (c) and (d) should be submitted as written answers in the same pdf as the other written problems.

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.

15. Did you know that computers can identify you not only by what you write, but also by how you write? Coursera uses Biometric Keystroke signatures for plagiarism detection. If you cannot write a sentence with the same statistical distribution of key press timings as in your previous work, they assume that you are not the person sitting behind the computer. In this problem we provide you with three files:

    - `personKeyTimingA.txt` has keystroke timing information for a user A writing a passage. The first column is the time in milliseconds (since the start of writing) when the user hit each key. The second column is the key that the user hit.
    - `personKeyTimingB.txt` has keystroke timing information for a second user (user B) writing the same passage as the user A. Even though the content of the passage is the same the timing of how the second user wrote the passage is different.
    - `email.txt` has keystroke timing information for an unknown user. We would like to know if the author of the email was user A or user B.

    Let X and Y be random variables for the duration of time, in milliseconds, for users A and B (respectively) to type a key. Assume that each keystroke from a user has a duration that is an independent random variable with the same distribution.

    a. **[Coding]** Complete the function `part_a` provided in the starter code. This function takes in a `filename` (which is either `personKeyTimingA.txt` or `personKeyTimingB.txt`) and should return $E[X]$ or $E[Y]$, respectively.

    b. **[Coding]** Complete the function `part_b` provided in the starter code. This function should return $E[X^2]$ or $E[Y^2]$, depending on which file is supplied as `filename`.

    c. **[Written]** Use your answers to part (a) and (b) and approximate X and Y as Normal random variables with mean and variance that match their biometric data. Report both distributions.

    d. **[Written]** Calculate the ratio of the probability that user A wrote the email over the probability that user B wrote the email. You do not need to submit code, but you should include the formula that you attempted to calculate and a short description (a few sentences) of how your code works.