CS 109: Probability for Computer Scientists
Problem Set#4

Adonis Pugh

February 21, 2020

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?

Answer.

a. From the rate given for users signing up for Lookbook, this situation can be modeled by a Poisson distribution with λ = 7.5.

$$L \sim Poi(7.5)$$

$$\boxed{P(L > 10) = 1 - P(L \leq 10) = 1 - \sum_{x=0}^{10} \frac{7.5^x}{x!} e^{-7.5} \approx 0.1378}$$

b. From the rate given for users signing up for Quickgram, this situation can be modeled by a Poisson distribution with λ = 11, twice the rate given because we are looking for a two minute probability.

$$Q \sim Poi(11)$$

$$\boxed{P(Q > 13) = 1 - P(Q \leq 13) = 1 - \sum_{x=0}^{13} \frac{11^x}{x!} e^{-11} \approx 0.2187}$$

c. Combining both rates and multiplying by two to account for a two-minute probability, we can model the combined scenario as a Poisson.

$$L + Q \sim Poi(2 \cdot 7.5 + 2 \cdot 5.5) = Poi(26)$$

$$\boxed{P(L + Q > 40) = 1 - P(Q \leq 40) = 1 - \sum_{x=0}^{40} \frac{26^x}{x!} e^{-26} \approx 0.0039418}$$

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

Answer.

a.

$$m = \frac{a+b}{2}$$

b.

$$m = \mu$$

c.

$$\begin{aligned}
&1 - e^{-\lambda m} = 0.5 \\
&e^{-\lambda m} = \frac{1}{2} \\
&-\lambda m = \ln(0.5) \\
&\boxed{m = \frac{-\ln(0.5)}{\lambda} = \frac{\ln(2)}{\lambda}}
\end{aligned}$$

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.

Answer.

a.

$$A = X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

b.

$$B = 4X + 3 \sim \mathcal{N}(4\mu_1 + 3, 16\sigma_1^2)$$

c.

$$C = aX - b^2Y + cZ \sim \mathcal{N}(a\mu_1, a^2\sigma_1^2) + \mathcal{N}(-b^2\mu_2, b^4\sigma_2^2) + \mathcal{N}(c\mu_3, c^2\sigma_3^3)$$
$$C = aX - b^2Y + cZ \sim \mathcal{N}(a\mu_1 - b^2\mu_2 + c\mu_3, a^2\sigma_1^2 + b^4\sigma_2^2 + c^2\sigma_3^2)$$

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?

Answer. We can think about this problem as being a binomial distribution of n chips with probability of lifetimes less than 1.9×10^6 hours given by the normal distribution of their lifetimes.

$$X \sim \mathcal{N}(1.5 \times 10^6, (9 \times 10^5)^2) = \mathcal{N}(1.5 \times 10^6, 8.1 \times 10^{11})$$

$$P(X < 1.9 \times 10^6) = F(1.9 \times 10^6) = \Phi\left(\frac{1.9 \times 10^6 - 1.5 \times 10^6}{9 \times 10^5}\right)$$

$$P(X < 1.9 \times 10^6) = \Phi\left(\frac{4 \times 10^5}{9 \times 10^5}\right) = \Phi\left(\frac{4}{9}\right) \approx 0.67$$

$$Y \sim Bin(100, 0.67)$$

$$\boxed{P(Y \geq 65) = \sum_{x=65}^{100} \binom{100}{x} (0.67)^x (1 - 0.67)^{100-x} \approx 0.7054}$$

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.

Answer. Modeling a dice roll as a multinomial distribution, we can compute the ratio of the relative probabilities.

$$\boxed{\frac{\binom{6}{1,1,1,1,1,1}\left(\frac{1}{6}\right)^6}{\binom{6}{6}\left(\frac{1}{6}\right)^6} = 720}$$

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).

Answer.

a. We can model the bug manifestation as a binomial distribution with a "trial" each hour the program is run with probability of occurrence $1/200$ per hour.

$$X \sim Bin(3, 1/200)$$
$$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{3}{0}\left(\frac{1}{200}\right)^0\left(1 - \frac{1}{200}\right)^{3-0}$$
$$\boxed{P(X \geq 1) = 1 - \left(1 - \frac{1}{200}\right)^3 \approx 0.01493}$$

b. Using the probability that a single user encounters the bug from part (a), we can estimate the probability that more than 10 thousand users out of 1 million experience by approximating a binomial distribution as a normal distribution.

$$X \sim Bin(10^6, 0.01493)$$
$$Y \sim \mathcal{N}(10^6 \times 0.01493, (10^6 \times 0.01493)(1 - 0.01493)) = \mathcal{N}(14925, 14702)$$
$$P(Y > 10^4) = 1 - P(Y < 10^4) = 1 - F(10^4) = 1 - \Phi\left(\frac{10^4 - 14925}{\sqrt{14702}}\right)$$
$$\boxed{P(Y > 10^4) = 1 - (1 - \Phi(40.623)) = \Phi(40.623) \approx 1.00}$$

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?

Answer. This situation can described using a multinomial distribution.

$$\boxed{\binom{15}{6, 4, 2, 3}(0.4)^6(0.3)^4(0.1)^2(0.2)^3 \approx 0.01674}$$

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 $E[X]$?

Answer.

a. The integral across the full bounds of the function must equal a probability of 1, so we compute the integral as follows:

$$c\int_0^1\int_0^x \frac{y}{x}dydx = 1 \to c\int_0^1 \frac{1}{x}\left[\frac{1}{2}y^2\right]_0^x dx = 1 \to c\int_0^1 \frac{x}{2}dx = 1$$

$$c\left[\frac{1}{4}x^2\right]_0^1 = 1 \to \frac{1}{4}c = 1 \to \boxed{c = 4}$$

b. X and Y are NOT independent, because the domain of X depends on Y. Mathematically, $P(Y = y \mid X = x) \neq P(Y = y)$. That is, knowing X changes the values that Y can take on.

c. To get the marginal density function of X, we must integrate over all values of Y.

$$f_X(x) = \int_0^x 4\frac{y}{x}dy = \frac{4}{x}\int_0^x ydy = \frac{4}{x}\left[\frac{1}{2}y^2\right]_0^x = \frac{4}{x}\left(\frac{x^2}{2}\right) = \boxed{2x}$$

d. To get the marginal density function of Y, we must integrate over all values of X.

$$f_X(x) = \int_y^1 4\frac{y}{x}dx = 4y\int_y^1 \frac{1}{x}ydy = 4y\ln(x)|_1^y = \boxed{-4y\ln(y)}$$

e. The expectation of X is obtained by integrating over all values of the marginal density function of X multiplied by the value of X.

$$E[X] = \int_0^1 2x^2dx = \left[\frac{2}{3}x^3\right]_0^1 = \boxed{\frac{2}{3}}$$

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 $E[D]$.

Answer. Let X and Y be uniform distributions, and let D be the sum of their absolute values. Since the expectation of a sum is the sum of expectations, we can equivalently say $E[D] = E[X] + E[Y]$, where X and Y are the absolute values of their distributions.

$$D = |x| + |y|$$
$$X = |x| \quad Y = |y|$$
$$E[D] = E[X + Y] = E[X] + E[Y]$$
$$E[X] = E[Y] = \frac{5 + 0}{2} = 2.5$$
$$\boxed{E[D] = E[X] + E[Y] = 2.5 + 2.5 = 5}$$

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.

Answer.

a. A table can be created to completely describe the joint probability mass function of X and Y, after which an equation can be derived describing the trend, where X is vertical and Y is horizontal.

$1/5$	0	0	0	0
$1/10$	$1/10$	0	0	0
$1/15$	$1/15$	$1/15$	0	0
$1/20$	$1/20$	$1/20$	$1/20$	0
$1/25$	$1/25$	$1/25$	$1/25$	$1/25$

From the tabular form of the joint PMF, we can see that the formula for the joint PMF can be described by $\boxed{P(X = x, Y = y) = \frac{1}{5x}, 1 \leq x \leq 5}$.

b. The values in tabular form of a condition mass function of X given Y can be computed using the relation $(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$. The intersection of X and Y are simply the values in table above and the probability of Y is described by the marginal PMF of Y, which is obtained by summing all the X values along each column of Y. The conditional PMF of X given Y is obtained as piece-wise function by carrying out algebra using the formula for the joint PMF from part (a).

$60/137$	0	0	0	0
$30/137$	$30/77$	0	0	0
$20/137$	$20/77$	$20/47$	0	0
$15/137$	$15/77$	$15/47$	$15/27$	0
$12/137$	$12/77$	$12/47$	$12/27$	1

$$\begin{cases} 60/137x & y = 1, 1 \leq x \leq 5 \\ 60/77x & y = 2, 2 \leq x \leq 5 \\ 60/47x & y = 3, 3 \leq x \leq 5 \\ 60/27x & y = 4, 4 \leq x \leq 5 \\ 1 & y = 5, x = 5 \\ 0 & otherwise \end{cases}$$

c. X and Y are NOT independent because via simple examination of the tables and formulas, $P_{X|Y}(x \mid y) \neq P_X(x)$. For example, $P_{X|Y}(1 \mid 1) = {}^{60}/_{137} \neq P_X(1) = {}^{1}/_{5}$. Therefore, X and Y are not independent.

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.

Answer. This problem can be solves using two known probability facts, including (1) $Cov(X, X) = Var(X)$ and (2) $E[XY] = E[X]E[Y]$ if X and Y are independent.

$$Cov(X, Y) = E[XY] - E[Y]E[X]$$
$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad Y_n = X_n + X_{n+1}$$

$$Cov(Y_n, Y_n) = Var(Y_n) = Var(X_n + X_{n+1}) = 2Var(X_n) = 2\sigma^2$$
$$\boxed{Cov(Y_n, Y_n) = 2\sigma^2}$$
$$Cov(Y_n, Y_{n+1}) = Cov(X_n + X_{n+1}, X_{n+1} + X_{n+2}) = Cov(X_{n+1}, X_{n+1}) = Var(X_{n+1}) = \sigma^2$$
$$\boxed{Cov(Y_n, Y_{n+1}) = \sigma^2}$$
$$Cov(Y_n, Y_{n+1}) = Cov(X_n + X_{n+1}, X_{n+2} + X_{n+3}) = Cov(X_n, X_{n+2})$$
$$Cov(Y_n, Y_{n+2}) = E[X_n, X_{n+2}] - E[X_n]E[X_{n+2}] = E[X_n]E[X_{n+2}] - E[X_n]E[X_{n+2}]$$
$$\boxed{Cov(Y_n, Y_{n+2}) = 0}$$

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.

Answer.

a. The probability that N strings are hashed to any bucket other than bucket 0 can be modeled as a geometric distribution where we want the first success to occur after N trials.

$$N \sim Geo(1 - p_0)$$
$$P(N = n) = (1 - (1 - p_0))^{n-1}(1 - p_0)$$
$$\boxed{P(N = n) = (p_0)^{n-1}(1 - p_0)}$$

b. The probability of any specific non-zero bucket is the probability of that bucket being hashed (event space) divided by all other non-zero buckets (sample space).

$$\boxed{P(X = j) = \frac{p_j}{1 - p_0}}$$

c. N and X are independent if the probability of their intersection is the product of their probabilities, which can be proven mathematically. Intuitively, the probability of the intersection of N and X is the probability of $n-1$ strings being hashed to bucket zero followed by a string being hashed to bucket j, which happens to be the product of the two individual probabilities previously calculated. Therefore, X and N must be independent.

$$P(N = n, X = j) = (p_0)^{n-1}(p_j)$$
$$P(N = n)P(X = j) = \frac{(p_0)^{n-1}(1 - p_0)(p_j)}{1 - p_0} = (p_0)^{n-1}(p_j)$$
$$\boxed{P(N = n, X = j) = P(N = n)P(X = j)}$$

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 $E[Y]$?

Answer.

a. If X = the value returned by `roll()`, then $E[X]$ can be computed using the law of total expectation, with special recursive cases to consider when the roll is less than 3. This results in an equation of one variable, $E[X]$ which can be solved to yield the expected value returned by the function.

$$E[X] = \sum_{i=1}^{6} (E[X] \mid roll = i) P(roll = i)$$
$$E[X] = \left(\frac{1}{6}\right) 6 + \left(\frac{1}{6}\right) 5 + \left(\frac{1}{6}\right) 4 + \left(\frac{1}{6}\right) 3 + \left(\frac{1}{6}\right) (2 + E[X]) + \left(\frac{1}{6}\right) (1 + E[X])$$
$$E[X] = 3 + \left(\frac{1}{6}\right) (2 + E[X] + 1 + E[X]) = 3 + \left(\frac{1}{6}\right) (3 + 2E[X])$$
$$E[X] - \frac{1}{3}E[X] = \frac{7}{2} \rightarrow \frac{2}{3}E[X] = \frac{7}{2}$$
$$\boxed{E[X] = \frac{21}{4} = 5.25}$$

b. This situation can be modeled as a geometric distribution with expected probability of success being the probability that a value greater than 3 is rolled.

$$Y \sim Geo(4/6) = Geo(2/3)$$

$$\boxed{E[Y] = \left(\frac{2}{3}\right)^{-1} = \frac{3}{2}}$$

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()`?

Answer.

The expected return value of `numInfected()` can be computed using the law of total expectation, taking into account special recursive cases. This results in an equation of one variable, which can be solved to yield the expected return value of the function.

$$E[Y] = E[Y \mid X = 0]P(X = 0) + E[Y \mid X = 1]P(X = 1)$$
$$E[Y] = 0(0.99) + (25E[Y] + 1)(0.01) = 0.25E[Y] + 0.01$$
$$0.75E[Y] = 0.01$$
$$\boxed{E[Y] = \frac{1}{75}}$$

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.

Answer.

a. CODE

b. CODE

c. The parameters for a normal distribution are the mean μ and the variance σ^2. The mean is given by expected time of a keystroke for a user $(E[X])$ and variance is calculated via the formula $Var(X) = E[X^2] - E[X]^2$. These values can all be computed using the functions in part (a) and part (b).

$$\mu_X = E[X] = 7.4067 \quad \mu_Y = E[Y] = 8.0308$$
$$E[X]^2 = 54.8596 \quad E[Y]^2 = 64.4933$$
$$E[X^2] = 58.8402 \quad E[Y^2] = 68.1769$$
$$\sigma_X^2 = E[X^2] - E[X]^2 = 58.8402 - 54.8596 = 3.9806$$
$$\sigma_Y^2 = E[Y^2] - E[Y]^2 = 68.1769 - 64.4933 = 3.6835$$
$$\boxed{X \sim \mathcal{N}(7.4067, 3.9806) \quad Y \sim \mathcal{N}(8.0308, 3.6835)}$$

d. Using the functions in part (a) and part (b), we have generated normal distributions for the keystroke times of user A and B. To calculate the probability that user A wrote the email over the probability that user B wrote the email, we can use the formula $\frac{f_X(z)}{f_Y(z)}$, where f is the probability density function of a normal distribution and Z is the expected keystroke time for the email writer computed using the function in part (a).

$$\frac{f_X(z)}{f_Y(z)} = \frac{f_X(7.94417)}{f_Y(7.94417)} = \frac{0.19283}{0.20765} \approx \boxed{0.9286}$$

The function in part (a) computes the expected keystroke time by summing over all the times (subtracting the current time from the last time to get the keystroke time) and multiplying by the probability of that time. In this case, we can do this by treating the probability of a keystroke as $\frac{1}{N}$ where N is the total number of keystroke times. The function in part (b) does the same as part (a) but computes the expected squared keystroke time by squaring each keystroke time within the summation.