CS 70 Discrete Mathematics and Probability Theory
Summer 2023 Huang, Suzani, and Tausik

DIS 7B

1 It's Raining Fish

Note 19
Note 21

A hurricane just blew across the coast and flung a school of fish onto the road nearby the beach. The road starts at your house and is infinitely long. We will label a point on the road by its distance from your house (in miles). For each $n \in \mathbb{N}$, the number of fish that land on the segment of the road $[n, n+1]$ is independently Poisson(λ) and each fish that is flung into that segment of the road lands uniformly at random within the segment. Keep in mind that you can cite any result from lecture or discussion without proof.

(a) What is the distribution of the number of fish arriving in segment $[0, n]$ of the road, for some $n \in \mathbb{N}$?

(b) Let $[a, b]$ be an interval in $[0, 1]$. What is the distribution of the number of fish that lands in the segment $[a, b]$ of the road?

(c) Let $[a, b]$ be any interval such that $a \geq 0$. What is the distribution of the number of fish that land in $[a, b]$?

(d) Suppose you take a stroll down the road. What is the distribution of the distance you walk (in miles) until you encounter the first fish?

(e) Suppose you encounter a fish at distance x. What is the distribution of the distance you walk until you encounter the next fish?

Solution:

(a) From lecture, we learned that if X and Y are independent, and $X \sim$ Poisson(λ) and $Y \sim$ Poisson(μ), then $X + Y \sim$ Poisson($\lambda + \mu$). We know the number of fish to land in the segment $[0, n]$ is the sum of the number of fish to land in $[i, i+1]$ for each $i \in [0, 1, \ldots, n-1]$. Thus the number of fish is in $[0, n]$ is Poisson($n\lambda$).

(b) Intuitively, the probability that a particular fish lands in the interval $[a, b]$ is $b - a$ since its location is uniformly distributed within $[0, 1]$. Thus, the distribution is Poisson($(b-a)\lambda$).

More rigorously, suppose we want the probability that k fish landed in the interval $[a, b]$. This can only happen if we had a total of some $i \geq k$ fish and exactly k of the i fish landed in the subinterval.

Generally, we have a probability $\frac{\lambda^i}{i!} e^{-\lambda}$ of getting i fish in the interval $[0, 1]$, as this is distributed with Poisson(λ). The probability that exactly k of these fish landed in $[a, b]$ is binomial, distributed with Binomial$(i, b-a)$, since there are a total of i fish, each with probability $b-a$ of landing in the interval (as the fish are uniformly distributed).

The probability that k fish land in the interval $[a,b]$ is then the summation over all possible $i \geq k$:

$$\begin{aligned}\mathbb{P}[Y=k] &= \sum_{i=k}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} \binom{i}{k} (b-a)^k (1-b+a)^{i-k} \\ &= (b-a)^k e^{-\lambda} \sum_{i=k}^{\infty} \frac{\lambda^i}{i!} \frac{i!}{k!(i-k)!} (1-b+a)^{i-k} \\ &= \frac{(b-a)^k}{k!\, e^{\lambda}} \sum_{i=k}^{\infty} \frac{\lambda^i (1-b+a)^{i-k}}{(i-k)!}\end{aligned}$$

Changing bounds and substituting $j = i-k$ or $i = j+k$, we have

$$\begin{aligned}\mathbb{P}[Y=k] &= \frac{(b-a)^k}{k!\, e^{\lambda}} \sum_{j=0}^{\infty} \frac{\lambda^{(j+k)} (1-b+a)^j}{j!} \\ &= \frac{(b-a)^k}{k!\, e^{\lambda}} \sum_{j=0}^{\infty} \frac{\lambda^k (\lambda(1-b+a))^j}{j!} \\ &= \frac{(b-a)^k \lambda^k}{k!\, e^{\lambda}} e^{\lambda(1-b+a)} \quad \text{(by the Taylor expansion of } e^u\text{)} \\ &= \frac{(\lambda(b-a))^k}{k!} e^{-\lambda(b-a)}\end{aligned}$$

Which is the PMF of a Poisson distribution with mean $\lambda(b-a)$, or Poisson$(\lambda(b-a))$.

(c) The answer is still Poisson$((b-a)\lambda)$. Clearly, this is true if $[a,b]$ is contained within some interval $[n, n+1]$. If it's not, then let i be the smallest integer such that $i \geq a$ and let j be the largest integer such that such that $j \leq b$. Then the distribution is

$$\text{Poisson}((i-a)\lambda) + \text{Poisson}((j-i)\lambda) + \text{Poisson}((b-j)\lambda) = \text{Poisson}((b-a)\lambda).$$

(d) The distance is Exp(λ). To prove this, it suffices to show that the cdf matches the exponential cdf. Let X be the distance of the first fish from the house. Note that $\mathbb{P}[X \geq t] = \mathbb{P}[\text{no fish in } [0,t]]$. By the previous parts, we know that the number of fish in $[0,t]$ is Poisson(λt), which is equal to 0 with probability

$$\frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}.$$

Thus, we have that $\mathbb{P}[X < t] = 1 - e^{-\lambda t}$ which is exactly the exponential cdf.

(e) Still Exp(λ). Using the same logic as the previous part, we have that $\mathbb{P}[X \geq t]$ is equal to the probability that no fish lands in the segment $[x, x+t]$. This in turn is equal to $e^{-\lambda t}$, because the number of fish in that segment is distributed as Poisson$(-\lambda t)$.

2 Practical Confidence Intervals

(a) It's New Year's Eve, and you're re-evaluating your finances for the next year. Based on previous spending patterns, you know that you spend \$1500 per month on average, with a standard deviation of \$500, and each month's expenditure is independently and identically distributed. As a college student, you also don't have any income. How much should you have in your bank account if you don't want to run out of money this year, with probability at least 95%?

(b) As a UC Berkeley CS student, you're always thinking about ways to become the next billionaire in Silicon Valley. After hours of brainstorming, you've finally cut your list of ideas down to 10, all of which you want to implement at the same time. A venture capitalist has agreed to back all 10 ideas, as long as your net return from implementing the ideas is positive with at least 95% probability.

Suppose that implementing an idea requires 50 thousand dollars, and your start-up then succeeds with probability p, generating 150 thousand dollars in revenue (for a net gain of 100 thousand dollars), or fails with probability $1-p$ (for a net loss of 50 thousand dollars). The success of each idea is independent of every other. What is the condition on p that you need to satisfy to secure the venture capitalist's funding?

(c) One of your start-ups uses error-correcting codes, which can recover the original message as long as at least 1000 packets are received (not erased). Each packet gets erased independently with probability 0.8. How many packets should you send such that you can recover the message with probability at least 99%?

Solution:

(a) Let T be the random variable representing the amount of money we spend in the year.

We have $T = \sum_{i=1}^{12} X_i$, where X_i represents the spending in the i-th month. So, $\mathbb{E}[T] = 12 \cdot \mathbb{E}[E_1] = 18000$.

And, since the X_is are independent, $\text{Var}(T) = 12 \cdot \text{Var}(X_1) = 12 \cdot 500^2 = 3{,}000{,}000$.

We want to have enough money in our bank account so that we don't finish the year in debt with 95% confidence. So, we want to keep some money ε more than the mean expenditure such that the probability of deviating above the mean by more than ε is less than 0.05.

Let's use Chebyshev's inequality here to express this.

$$\mathbb{P}[|T - \mathbb{E}[T]| \geq \varepsilon] \leq \frac{\text{Var}(T)}{\varepsilon^2} \leq 0.05$$

This gives us $\varepsilon^2 \geq \dfrac{3{,}000{,}000}{0.05}$. So, $\varepsilon \geq 7746$. This means that we want to have a balance of $\geq \mathbb{E}[T] + \varepsilon = 25746$.

Observe that here, while we wanted to estimate $\mathbb{P}[T-\mathbb{E}[T] \geq \varepsilon]$, Chebyshev's inequality only gives us information about $\mathbb{P}[|T-\mathbb{E}[T]| \geq \varepsilon]$. But since

$$\mathbb{P}[|T-\mathbb{E}[T]| \geq \varepsilon] \geq \mathbb{P}[T-\mathbb{E}[T] \geq \varepsilon],$$

this is fine. We just get a more conservative estimate.

(b) For this question, to keep the numbers from exploding, let's work in thousands of dollars. Let X_i be the profit made from idea i, and T be the total profit made. We have $T = \sum_{i=1}^{10} X_i$.

Here, $\mathbb{E}[X_1] = 100p - 50(1-p) = 150p - 50$.

And $\mathrm{Var}(X_1) = 150^2 p(1-p)$ as the distribution of X_1 is a shifted and scaled Bernoulli distribution. Using $\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2$ yields the same answer.

We have, $\mathbb{E}[T] = 10 \cdot \mathbb{E}[X_1]$. Similarly, $\mathrm{Var}(T) = 10 \cdot \mathrm{Var}(X_1)$.

Now, we want to bound the probability of T going below 0 by 0.05. In other words, we want $\mathbb{P}[T<0] \leq 0.05$.

But, in order to apply Chebyshev's inequality, we need to look at deviation from the mean. We use the assumption that to get our funding we obviously need $\mathbb{E}[T] > 0$. Then:

$$\mathbb{P}[T<0] \leq \mathbb{P}[T \leq 0 \cup T \geq 2\mathbb{E}[T]] = \mathbb{P}[|T-\mathbb{E}[T]| \geq \mathbb{E}[T]] \leq \frac{\mathrm{Var}(T)}{\mathbb{E}[T]^2} \leq 0.05$$

Looking at just the last inequality, we have:

$$\frac{\mathrm{Var}(T)}{\mathbb{E}[T]^2} = \frac{10 \cdot \mathrm{Var}(X_1)}{100 \cdot \mathbb{E}[X_1]^2} = \frac{\mathrm{Var}(X_1)}{10 \cdot \mathbb{E}[X_1]^2} \leq 0.05$$

$$\therefore \frac{\mathrm{Var}(X_1)}{\mathbb{E}[X_1]^2} \leq 0.5$$

Now, substituting what we have for variance and expectation, we get the following:

$$-22500p^2 + 22500p \leq 0.5(150p-50)^2$$

which gives us the quadratic:

$$33750p^2 - 30000p + 1250 \geq 0$$

The solutions for p are $p \geq \frac{1}{9}(4+\sqrt{13})$ and $p \leq \frac{1}{9}(4-\sqrt{13})$. So $p \geq 0.845$ or ≤ 0.0438.

The relevant solution here is to pick $p \geq 0.845$, since the other solution yields negative expectation (contradicting the earlier assumption of positive expectation).

(c) We want $k = 1000$ packets to get across without being erased. Say we send n packets. Let X_i be the indicator random variable representing whether the ith packet got across or not.

Let the total number of unerased packets sent across be T. We have $T = \sum_{i=1}^{n} X_i$ and we want $T \geq 1000$.

We want $\mathbb{P}[T < 1000] \leq 0.01$. Now, let's try to get this in a form so that we can use Chebyshev's inequality. We know that $\mathbb{E}[T] > 1000$, so we can say that

$$\begin{aligned}\mathbb{P}[T < 1000] &\leq \mathbb{P}[T \leq 1000 \cup T \geq \mathbb{E}[T] + (\mathbb{E}[T] - 1000)] \\ &= \mathbb{P}[|T - \mathbb{E}[T]| \geq (\mathbb{E}[T] - 1000)] \leq \frac{\mathrm{Var}(T)}{(\mathbb{E}[T] - 1000)^2} \leq 0.01.\end{aligned}$$

What is $\mathbb{E}[T]$? $\mathbb{E}[T] = n\mathbb{E}[X_1] = n(1-p) = 0.2n$.

Next, what is $\mathrm{Var}(T)$? $\mathrm{Var}(T) = n\,\mathrm{Var}(X_1) = np(1-p) = 0.16n$.

Now, $\frac{\mathrm{Var}(T)}{(\mathbb{E}[T]-k)^2} \leq 0.01 \implies 16n \leq (0.2n - 1000)^2$. This gives us the quadratic:

$$0.04n^2 - 416n + 1000000 \geq 0$$

Solving the last quadratic, we get $n \geq 6629$ or $n \leq 3774$. Since the second inequality doesn't make sense for our situation, our answer is $n \geq 6629$.

3 Waiting For the Bus

Note 21

Edward and Jerry are waiting at the bus stop outside of Soda Hall.

Like many bus systems, buses arrive in periodic intervals. However, the Berkeley bus system is unreliable, so the length of these intervals are random, and follow Exponential distributions.

Edward is waiting for the 51B, which arrives according to an Exponential distribution with parameter λ. That is, if we let the random variable X_i correspond to the difference between the arrival time ith and $(i-1)$st bus (also known as the inter-arrival time) of the 51B, $X_i \sim \mathrm{Exp}(\lambda)$.

Jerry is waiting for the 79, whose inter-arrival times also follows Exponential distributions with parameter μ. That is, if we let Y_i denote the inter-arrival time of the 79, $Y_i \sim \mathrm{Exp}(\mu)$. Assume that all inter-arrival times are independent.

(a) What is the probability that Jerry's bus arrives before Edward's bus?

(b) After 20 minutes, the 79 arrives, and Jerry rides the bus. However, the 51B still hasn't arrived yet. Let D be the additional amount of time Edward needs to wait for the 51B to arrive. What is the distribution of D?

(c) Lavanya isn't picky, so she will wait until either the 51B or the 79 bus arrives. Find the distribution of Z, the amount of time Lavanya will wait before catching her bus.

(d) Khalil doesn't feel like riding the bus with Edward. He decides that he will wait for the second arrival of the 51B to ride the bus. Find the distribution of $T = X_1 + X_2$, the amount of time that Khalil will wait to ride the bus.

Solution:

(a) Let f_{Y_i} be the pdf of Y_i. By total probability,

$$\begin{aligned}
\mathbb{P}[X_i > Y_i] &= \int_{t=0}^{\infty} f_{Y_i}(t) \cdot \mathbb{P}[X_i > Y_i \mid Y_i = t] \, \mathrm{d}t \\
&= \int_{t=0}^{\infty} f_{Y_i}(t) \cdot \mathbb{P}[X_i > t] \, \mathrm{d}t \\
&= \int_{t=0}^{\infty} f_{Y_i}(t) \cdot (1 - F_{X_i}(t)) \, \mathrm{d}t \\
&= \int_{t=0}^{\infty} \mu e^{-\mu t} (e^{-\lambda t}) \, \mathrm{d}t \\
&= \mu \int_{t=0}^{\infty} e^{-(\lambda+\mu)t} \, \mathrm{d}t \\
&= \frac{\mu}{\lambda+\mu} \int_{t=0}^{\infty} (\lambda+\mu) e^{-(\lambda+\mu)t} \, \mathrm{d}t \\
&= \frac{\mu}{\lambda+\mu},
\end{aligned}$$

where the integral in the second-to-last line evaluates to 1, since it is the total integral of the Exponential$(\lambda+\mu)$ density.

(b) We observe that $\mathbb{P}[D > d] = \mathbb{P}[X > 20 + d \mid X \geq 20]$. Then, we apply Bayes Rule:

$$\begin{aligned}
\mathbb{P}[X > 20 + d \mid X \geq 20] &= \frac{\mathbb{P}[X > 20 + d]}{\mathbb{P}[X \geq 20]} \\
&= \frac{1 - F_X(20 + d)}{1 - F_X(20)} \\
&= \frac{e^{-\lambda(20+d)}}{e^{-20\lambda}} \\
&= e^{-\lambda d}
\end{aligned}$$

Thus, the CDF of D is given by $\mathbb{P}[D \leq d] = 1 - \mathbb{P}[D > d] = 1 - e^{-\lambda d}$. This is the CDF of an exponential, so D is exponentially distributed with parameter λ.

One can also directly apply the memoryless property of the exponential distribution to arrive at this answer.

(c) Lavanya's waiting time is the minimum of the time it takes for the 51B and the time it takes for the 79 to arrive. Thus, $Z = \min(X,Y)$.

$$\begin{aligned}
\mathbb{P}[Z > t] &= \mathbb{P}[X > t \cap Y > t] \\
&= \mathbb{P}[X > t] \cdot \mathbb{P}[Y > t] \\
&= (1 - F_X(t))(1 - F_Y(t)) \\
&= (1 - (1 - e^{-\mu t}))(1 - (1 - e^{-\lambda t})) \\
&= e^{-\mu t} e^{-\lambda t} \\
&= e^{-(\mu+\lambda)t}
\end{aligned}$$

It follows that the CDF is Z, $\mathbb{P}[Z \le t] = 1 - e^{-(\mu+\lambda)t}$. Thus, Z is exponentially distributed with parameter $\mu + \lambda$.

(d) Let $t > 0$. By total probability,

$$\begin{aligned}
\mathbb{P}[T \le t] &= \mathbb{P}[X_1 + X_2 \le t] \\
&= \int_0^\infty \mathbb{P}[X_1 + X_2 \le t \mid X_1 \in \mathrm{d}x] \cdot \mathbb{P}[X_1 \in \mathrm{d}x] \\
&= \int_0^t \mathbb{P}[X_1 + X_2 \le t \mid X_1 \in \mathrm{d}x] \cdot \mathbb{P}[X_1 \in \mathrm{d}x] + \int_t^\infty 0 \cdot \mathbb{P}[X_1 \in \mathrm{d}x] \\
&= \int_0^t \mathbb{P}[X_2 \le t - X_1 \mid X_1 \in \mathrm{d}x] \cdot \mathbb{P}[X_1 \in \mathrm{d}x] + 0 \\
&= \int_0^t \mathbb{P}[X_2 \le t - x] \cdot \mathbb{P}[X_1 \in \mathrm{d}x] \\
&= \int_0^t F_{X_2}(t - x) \cdot f_{X_1}(x)\,\mathrm{d}x \\
&= \int_0^t \left(1 - e^{-\lambda(t-x)}\right) \cdot \lambda e^{-\lambda x}\,\mathrm{d}x \\
&= \int_0^t \lambda e^{-\lambda x} - \lambda e^{-\lambda t}\,\mathrm{d}x \\
&= \int_0^t \lambda e^{-\lambda x} - \lambda e^{-\lambda t} \int_0^t \mathrm{d}x \\
&= F_{X_1}(t) - \lambda e^{-\lambda t} \cdot t \\
&= 1 - e^{-\lambda t} - \lambda t e^{-\lambda t}
\end{aligned}$$

Upon differentiating the CDF, we have

$$\begin{aligned}
f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}[T \le t] &= \lambda e^{-\lambda t} - \lambda e^{-\lambda t} + \lambda^2 t e^{-\lambda t} \\
&= \lambda^2 t e^{-\lambda t}, \quad \text{for } t > 0.
\end{aligned}$$

4 Student Life

Note 19

In an attempt to avoid having to do laundry often, Marcus comes up with a system. Every night, he designates one of his shirts as his dirtiest shirt. In the morning, he randomly picks one of his shirts to wear. If he picked the dirtiest one, he puts it in a dirty pile at the end of the day (a shirt in the dirty pile is not used again until it is cleaned).

When Marcus puts his last shirt into the dirty pile, he finally does his laundry, and again designates one of his shirts as his dirtiest shirt (laundry isn't perfect) before going to bed. This process then repeats.

(a) If Marcus has n shirts, what is the expected number of days that transpire between laundry events? Your answer should be a function of n involving no summations.

(b) Say he gets even lazier, and instead of organizing his shirts in his dresser every night, he throws his shirts randomly onto one of n different locations in his room (one shirt per location), designates one of his shirts as his dirtiest shirt, and one location as the dirtiest location.

In the morning, if he happens to pick the dirtiest shirt, *and* the dirtiest shirt was in the dirtiest location, then he puts the shirt into the dirty pile at the end of the day and does not throw any future shirts into that location and also does not consider it as a candidate for future dirtiest locations (it is too dirty).

What is the expected number of days that transpire between laundry events now? Again, your answer should be a function of n involving no summations.

Solution:

(a) The number of days that it takes for him to throw a shirt into the dirty pile can be represented as a geometric RV. For the first shirt, this is the geometric RV with $p = 1/n$. We can see this by noticing that every day up to the day he picks the dirtiest shirt, the probability of getting the dirtiest shirt remains $1/n$.

We'll call X_i the number of days that go until he throws the ith shirt into the dirty pile. Since on the ith shirt, there are $n-i+1$ shirts left, we get that $X_i \sim \text{Geometric}(1/(n-i+1))$. The number of days until he does his laundry is a sum of these variables. Therefore, we can get the following result:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} (n-i+1) = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

(b) For this part we can use a similar approach but the probability for X_i becomes $1/(n-i+1)^2$. This is because the dirtiest shirt falls into the dirtiest spot with probability $1/(n-i+1)$ and we pick it after that with probability $1/(n-i+1)$, so the probability of picking the dirtiest shirt from the dirtiest spot for the ith shirt is $1/(n-i+1)^2$. Using the same approach, we get the following sum:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} (n-i+1)^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

5 Playing Blackjack

You are playing a game of Blackjack where you start with \$100. You are a particularly risk-loving player who does not believe in leaving the table until you either make \$400, or lose all your money. At each turn you either win \$100 with probability p, or you lose \$100 with probability $1-p$.

(a) Formulate this problem as a Markov chain; i.e. define your state space, transition probabilities, and determine your starting state.

(b) Compute the probability that you end the game with \$400.

Solution:

(a) Since it is only possible for us to either win or lose \$100, we define the following state space $\chi = \{0, 100, 200, 300, 400\}$. The following are the transition probabilities:

$$\mathbb{P}[X_j = 0 | X_{j-1} = 0] = \mathbb{P}[X_j = 400 | X_{j-1} = 400] = 1$$
$$\mathbb{P}[X_j = i + 100 | X_{j-1} = i] = p \text{ for } i \in \{100, 200, 300\}$$
$$\mathbb{P}[X_j = i - 100 | X_{j-1} = i] = 1 - p \text{ for } i \in \{100, 200, 300\}$$

(b) We want to find the probability that we are "absorbed" by state 400 before we are absorbed by state 0. We can calculate this probability by leveraging the memoryless property of Markov Chains. Define a_i as the probability of reaching state 400 before 0 starting at state i.

We also know that for $i \in \{100, 200, 300\}$, we have the following relation:

$$a_i = (1-p)a_{i-100} + pa_{i+100} \text{ for } i \in \{100, 200, 300\}$$

We also know that $a_0 = 0$, since if you are at state 0, then there is no chance that you end up at state 400. We also have $a_{400} = 1$ since if we are at state 400, then we have already succeeded in our goal to reach 400.

We have three unknowns $(a_{100}, a_{200}, a_{300})$ and three equations, and we can now solve this system of equations for a_{100}.

$$
\begin{gathered}
a_0 = 0, a_{400} = 1 \\
\implies a_i = (1-p)a_{i-100} + pa_{i+100} \text{ for } i \in \{100, 200, 300\} \\
a_{100} = pa_{200} \\
a_{200} = (1-p)a_{100} + pa_{300} \implies a_{200}[1-p(1-p)] = pa_{300} \\
\implies a_{200} = \frac{pa_{300}}{1-p(1-p)} \\
a_{300} = (1-p)a_{200} + p \implies a_{300} = \frac{(1-p)pa_{300}}{1-p(1-p)} + p \\
\implies a_{300} = \frac{p(1-p(1-p))}{1-2p(1-p)} \\
\implies a_{200} = \frac{p^2}{1-2p(1-p)} \\
\implies a_{100} = \frac{p^3}{1-2p(1-p)}
\end{gathered}
$$

This problem is called Gambler's Ruin, where it is used to show that even if p is decently large, after playing a large number of games without stopping, you will end up at 0 dollars with high probability.