12

Geometric & Poisson Distributions and More

Computer Science Mentors 70

Nov 23 - Nov 27
Fall 2020

Prepared by: Aishani Sil, Austin Lei, Agnibho Roy, Debayan Bandyopadhyay, Abinav Routhu
Roast us here: *https://tinyurl.com/csm70-feedback20*

I Geometric Distributions

1. **Intro**
You are rolling a fair, six-sided die.

(a) Compute the probability that it takes you 6 tries to roll a 6.

Solution: Let's define a random variable X to represent the number of tries it takes to roll a 6. $X \sim Geom(\frac{1}{6})$, and the probability of taking 6 tries to roll a 6 is $(\frac{5}{6})^5(\frac{1}{6})$.

(b) Compute the expected number of times it takes you to roll a prime number.

Solution: Let's define a random variable Y to represent this quantity. $Y \sim Geom(\frac{1}{2})$, since $\frac{3}{6}$ of the numbers on the die are prime. The expected value of this distribution is $\frac{1}{(\frac{1}{2})} = 2$.

2. **Free Food!**
A cafe has a free sandwich menu, and each day gives you a random one of the 10 sandwiches on their menu for free. You want to try all of the available sandwiches, but you have two restrictions: you're allergic to fish, so you cannot have the 2 sandwiches that contain that ingredient, and you need all of your sandwiches toasted. The server doesn't care enough though, so you will still get a totally random sandwich that has a probability of $\frac{1}{2}$ of being toasted; you throw out the order if you cannot eat it. What is the expected number of days it takes you to try the 8 sandwiches you're able to have?

Solution: This is a variation of the coupon collector problem, with some added restrictions. Let us again define S to be the number of days it takes to try the 8 sandwiches. We know that

$$S = S_1 + S_2 + \ldots S_8$$

where S_i is the number of days it takes to get the i^{th} sandwich, starting immediately after we successfully get the $(i-1)^{th}$ sandwich.

What is S_1? For the first sandwich, there are $\frac{8}{10}$ choices that are valid, and we must multiply by an additional $\frac{1}{2}$ because you discard the sandwich if it is not toasted. So the probability that you get to try your first free sandwich on any given day is $\frac{8}{10} \cdot \frac{1}{2}$. S_1 is a geometric distribution with parameter $\frac{2}{5}$, so the EV is $\frac{5}{2}$. We follow a similar approach for finding S_2, but we only have 7 sandwiches left to eat, so $S_2 \sim Geom(\frac{7}{10} \cdot \frac{1}{2})$. In fact, for any S_i, the probability of eating the i^{th} sandwich given that we've eaten all $i-1$ previous ones is

$$\frac{8-(i-1)}{10} \cdot \frac{1}{2}.$$

It follows that

$$E[S_k] = \frac{20}{8-k+1}.$$

A key takeaway is that each S_i follows a geometric distribution. Going back to our expression for S earlier, we can now conclude that

$$E[S] = \sum_{i=1}^{8} \frac{20}{8-i+1} = 20 \sum_{i=1}^{8} \frac{1}{8-i+1}.$$

3. **Minimum of Geometric Random Variables**

In this question, we will explore how the minimum, Z, of two geometric random variables, X and Y, is distributed geometrically as well.

(a) Find the CDF of Z in terms of F_X and F_Y, the CDFs of X and Y, respectively.

Solution:

$$P(Z \leq z) = 1 - P(Z > z) = 1 - P(\min\{X, Y\} > z) = 1 - P(X > z)P(Y > z) = 1 - (1 - F_X)(1 - F_Y)$$

(b) Let's assume that the probability parameters of X and Y are p_x and p_y, respectively. Using the previous part, explicitly calculate the CDF of Z.

Solution: We first calculate the CDF of a geometric random variable, which is going to be the probability that we have less than k failures until a success. This can also be thought of as the complement of at least k failures, which can be represented as $1 - (1 - p)^k$ for a general geometric RV. Thus, from the previous part, we have:

$$P(Z \leq z) = 1 - (1 - (1 - (1 - p_x)^z))(1 - (1 - (1 - p_y)^z)) = 1 - (1 - p_x)^z(1 - p_y)^z$$

(c) Calculate the PDF of Z, show that it is geometrically distributed, and find its parameter, p_z.

Solution: We know that we can find the PDF of a discrete random variable by doing the following:

$$P(Z = z) = P(Z \leq z) - P(Z \leq z - 1)$$

Applying this to the previous part, where we have derived the CDF of Z, we have:

$$\begin{aligned}
P(Z = z) &= 1 - (1 - p_x)^z(1 - p_y)^z - (1 - (1 - p_x)^{z-1}(1 - p_y)^{z-1}) \\
&= (1 - p_x)^{z-1}(1 - p_y)^{z-1} - (1 - p_x)^{z-1}(1 - p_y)^{z-1} \\
&= ((1 - p_x)(1 - p_y))^{z-1}(1 - (1 - p_x)(1 - p_y)) \\
&= \left(1 - (p_x + p_y - p_x p_y)^{z-1}\right)(p_x + p_y - p_x p_y)
\end{aligned}$$

This is the Geometric distribution with $p_z = p_x + p_y - p_x p_y$.

4. **Throwing Darts**

Alex and Bob are playing darts. They take turns throwing the dart and the first one to hit the center wins. Each turn, Alex hits the center with probability p and Bob hits the center with probability q. Alex gets to go first. We will explore two ways of finding out who is more likely to win.

(a) What is the probability that Alex wins on the k^{th} turn?

Solution: Both Alex and Bob must miss for the first $k - 1$ turns, then Alex must hit the center on the k^{th} turn. Thus, the probability he wins on the k^{th} turn is $((1 - p)(1 - q))^{k-1}p$

(b) Using the law of total probability, what is the probability that Alex wins?

Solution:

$$\sum_{k=1}^{\infty}((1-p)(1-q))^{k-1}p = p\sum_{k=0}^{\infty}((1-p)(1-q))^{k} = \frac{p}{1-(1-p)(1-q)} = \frac{p}{p+q-pq}$$

(c) We say that the game ended on turn k if either Alex or Bob hit the center on turn k. Let X be a random variable corresponding to the turn that the game ended on. What is the distribution of X (hint: what section are we on right now)?

Solution: The probability that the game ends on each turn is equalled to the probability that either Alex or Bob hits the center on that turn, and each turn is independent. By the inclusion-exclusion principal, the probability that either Alex or Bob hits the center on a given turn is $p + q - pq$. Thus $X \sim Geom(p + q - pq)$. Note that from discussion, this is the minimum of two geometric distributions, which makes sense because the number of turns taken in the game is equalled to the minimum number of turns taken by either of the two players.

(d) Using ONLY parts a and c, what is the probability that Alex wins given that the game ended on turn k.

Solution: Let A_k denote the event that Alex won on turn k. Then we have

$$\begin{aligned}
Pr[A_k|X = k] &= \frac{Pr[A_k \cap X = k]}{Pr[X = k]} \\
&= \frac{Pr[A_k]}{Pr[X = k]} \\
&= \frac{((1-p)(1-q))^{k-1}p}{((1-p)(1-q))^{k-1}(p+q-pq)} \\
&= \frac{p}{p+q-pq}
\end{aligned}$$

Where we note that $(1-p)(1-p) = (1-p-q+pq)$

(e) What does this tell us about the chances of Alex winning the game?

Solution: It's equal because the previous expression is independent of k. The intuition here is that regardless of when the game ended, it must end on at least one of the two people hitting the center so Alex's chance of winning is equalled to his chance of hitting center over the the chance that at least one person hits center.

II Poisson Distributions

5. **Pancake Flips**

Leanne makes X pancakes, where X is a Poisson(λ) random variable. She flips each pancake one-handed, and each pancake has a probability p of being flipped properly, and probability $1 - p$ of flying off the pan, and the probability of each pancake getting flipped properly is independent of the other pancakes. Show that the number of pancakes that get flipped properly follows a Poisson($p\lambda$) distribution.

Solution: Let Y be a Poisson distribution. Notice that if $X = n$ pancakes are made, the number of pancakes Y that are flipped properly follows a Binomial(n, p) distribution; in other words, the distribution of Y given $X = n$ is Binomial(n, p).

Thus, we have that

$$
\begin{aligned}
P(Y=k) &= \sum_{n=k}^{\infty} P(Y=k \cap X=n) \\
&= \sum_{n=k}^{\infty} P(X=n)\cdot P(Y=k|X=n) \\
&= \sum_{n=k}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!}\cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{n=k}^{\infty} \frac{\lambda^k \lambda^{n-k} e^{-p\lambda} e^{-(1-p)\lambda}}{n!}\cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{n=k}^{\infty} \frac{(p\lambda)^k e^{-p\lambda}}{k!}\cdot \frac{((1-p)\lambda)^{n-k} e^{-(1-p)\lambda}}{(n-k)!} \\
&= \frac{(p\lambda)^k e^{-p\lambda}}{k!} \sum_{n=k}^{\infty} \frac{((1-p)\lambda)^{n-k} e^{-(1-p)\lambda}}{(n-k)!} \\
&= \frac{(p\lambda)^k e^{-p\lambda}}{k!} \sum_{i=0}^{\infty} \frac{((1-p)\lambda)^{i} e^{-(1-p)\lambda}}{i!}
\end{aligned}
$$

Let X' be a Poisson$((1-p)\lambda)$ random variable. Then $\sum_{i=0}^{\infty} \frac{((1-p)\lambda)^i e^{-(1-p)\lambda}}{i!} = \sum_{i=0}^{\infty} P(X'=i)$, which by the law of total probability is just 1. Thus, $P(Y=k) = \frac{(p\lambda)^k e^{-p\lambda}}{k!}$, and Y is a Poisson$(p\lambda)$ distribution.

6. **Poisson as Limit of Binomial**

Let's explore merging 2 Poisson random variables, $A \sim Pois(\lambda_1)$ and $B \sim Pois(\lambda_2)$; that is, finding $E[A+B]$. It's totally possible to do this with tedious calculations, but let's do something different.

(a) The Poisson distribution is really a niche variant of the Binomial distribution for exceedingly rare events that have many, many chances of occurring (like car accidents, heart attacks, and sunny days in London). How can the Binomial distribution be used to model such an event? What would the expectation of this "coincidence" model be?

Solution: For this case, n, the number of trials, is high and p, the probability of success, is low. Now, since the trials are independent, the expectation of the Binomial RV is simply np.

If we think of the Poisson distribution as this special binomial trial where $n \longrightarrow \infty$ and $p \longrightarrow 0$, then as long as np is held constant so it's well-defined, the expectation of this random variable is $np = \lambda$.

(b) Suppose that A corresponds to the number of seasons Berkeley beats Stanford in a football game and B corresponds to the number of seasons Berkeley beats USC. Using the "coincidence" model above, what is the probability of Berkeley beating Stanford in any one season? What about the probability of beating USC? And lastly, the probability of beating either?

Solution: Recall that $\lambda = np$. So the probability of Berkeley beating Stanford in any one season/trial is $p_1 = \frac{\lambda_1}{n}$. Similarly, for USC, $p_2 = \frac{\lambda_2}{n}$. By inclusion-exclusion, beating either school in any 1 season has probability $p = \frac{\lambda_1}{n} + \frac{\lambda_2}{n} - \frac{\lambda_1\lambda_2}{n^2}$. Note that a Poisson process cannot have two events occur at the same time.

(c) Using your solution from part (b), what can you say about $E[A+B]$, the number of seasons Berkeley beats Stanford or

USC?

Solution: There are n independent trials or seasons each with probability $p = \frac{\lambda_1}{n} + \frac{\lambda_2}{n} - \frac{\lambda_1\lambda_2}{n^2}$. This itself is our "coincidence model"! In fact $X = A + B$ is a Poisson distribution with expectation $\lambda = np = n \cdot (\frac{\lambda_1}{n} + \frac{\lambda_2}{n} - \frac{\lambda_1\lambda_2}{n^2})$. Remember that n approaches infinity in our model, so $\lambda = \lim_{n\to\infty} n \cdot (\frac{\lambda_1}{n} + \frac{\lambda_2}{n} - \frac{\lambda_1\lambda_2}{n^2}) = \lambda_1 + \lambda_2$.

III Union Bound, Hashing, and Load Balancing

7. **Paper Triangle**
On a piece of paper, 70% of the paper is black and 30% is white. Show that an equilateral triangle can be drawn on the paper such that all of the vertices of the triangle are black. (Hint: Use the union bound to show that the probability of all the vertices being black is nonzero.)

Solution: Let E_1, E_2, E_3 be represent the events that vertex 1, vertex 2, and vertex 3 are black, respectively. We want to show that $P(E_1 \cap E_2 \cap E_3) > 0$, or taking the complement, $P(E_1^c \cup E_2^c \cup E_3^c) < 1$. By the union bound, we have that

$$P(E_1^c \cup E_2^c \cup E_3^c) \leq P(E_1^c) + P(E_2^c) + P(E_3^c) = 0.3 + 0.3 + 0.3 = 0.9 < 1,$$

since E_i^c represents the event that vertex i is white. Thus, the probability that all three vertices are black is nonzero, so there must be some triangle such that all the vertices of the triangle are black.

8. **Hashing and Load Balancing Questions**
Imagine that we have m tasks and we are trying to distribute it amongst n identical processors. Imagine the task being distributed uniformly at random amongst the processors. We will discuss some metrics that we may be interested in.

(a) Using a balls and bins model, what is the probability that one of the processors does not end up with any tasks? What about the expected number of processors with no tasks (for this part, assume $n = m$ and approximate for a large n)? *Hint:* $\left(1 - \frac{1}{n}\right)^n \longrightarrow e$ *as* $n \longrightarrow \infty$

Solution: The probability that one of the processors, say processor i, does not end up with any tasks is the probability that all tasks gets assigned to one of the other $n-1$ processor. A single task being assigned to one the other $n-1$ processors is $\frac{n-1}{n} = 1 - \frac{1}{n}$. Thus, for all m tasks, it would be:

$$\left(1 - \frac{1}{n}\right)^m$$

The expected number of processors with no tasks can be find by setting an indicator X_i to the probability that a processor i does not end up with any tasks, which we just computed. Using linearity of expectation:

$$E\left[\text{Processors with no tasks}\right] = \left[\sum_{i=1}^{n} X_i\right] = nE\left[X_1\right] = n\left(1 - \frac{1}{n}\right)^m$$

If we take $n = m$ and $n \longrightarrow \infty$ then we have:

$$n\left(1 - \frac{1}{n}\right)^n \approx \frac{n}{\epsilon}$$

(b) We want to limit the chance that one or more of our processors become overloaded. What is the probability that any one processor gets exactly k tasks?

Solution: Exactly k tasks must be assigned to processor i and the rest must be assigned to one of the other $n-1$ processors, which happens with probability $\left(\frac{1}{n}\right)^k \left(1-\frac{1}{n}\right)^{n-k}$. Since there are $\binom{n}{k}$ ways to choose a subset of k balls, we have:

$$P(\text{Exactly k tasks}) = \binom{n}{k}\left(\frac{1}{n}\right)^k \left(1-\frac{1}{n}\right)^{n-k}$$

(c) What is an upper bound on the probability that any one processor gets *at least* k tasks?

Solution: Lets enumerate all groups of k tasks from $1 \ldots \binom{n}{k}$ Let A_i be the probability that at least the specific subset i of k tasks gets assigned to any one processor. $P(A_i) = \left(\frac{1}{n}\right)^k$, because we don't care where the rest of the $n-k$ balls go. Using the union bound we can find that:

$$P\left(\bigcup_{i=1}^{\binom{n}{k}} A_i\right) \leq \sum_{i=1}^{\binom{n}{k}} A_i = \binom{n}{k} A_1 = \binom{n}{k}\left(\frac{1}{n}\right)^k$$