Jerry Cain
CS109

Section #3
April 19, 2022

Section #3: Discrete and Continuous Random Variables

1 Warmups

1.1 Website Visits

You have a website where only one visitor can be on the site at a time, but there is an infinite queue of visitors, so that immediately after a visitor leaves, a new visitor will come onto the website. On average, visitors leave your website after 5 minutes. Assume that the length of stay is exponentially distributed. What is the probability that a user stays more than 10 minutes, if we calculate this probability:

a. using the random variable X, defined as the length of stay of the user?

b. using the random variable Y, defined as the number of users who leave your website over a 10-minute interval?

If this problem doesn't convince you that the Poisson and Exponential RVc are coupled, then I'm not sure will! As defined above, $X \sim \text{Exp}(\lambda = \frac{1}{5})$.

$$P(X > 10) = 1 - F_X(10) = 1 - (1 - e^{-10\lambda}) = e^{-2} \approx 0.1353$$

Alternatively, we have that Y is the number of users leaving on the website in the next 10 minutes. The average number of users leaving is 2 users per 10 minutes. $Y \sim \text{Poi}(\lambda = 2)$.

$$\begin{aligned} P(Y = 0) &= \frac{2^0 e^{-2}}{0!} \\ &= e^{-2} \approx 0.1353 \end{aligned}$$

1.2 Continuous Random Variables

Let X be a continuous random variable with the following probability density function:

$$f_X(x) = \begin{cases} c(e^{x-1} + e^{-x}) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

a. Find the value of c that makes f_X a valid probability distribution.

b. What is $P(X > 0.75)$?

a. We need $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 c(e^{x-1} + e^{-x})dx$$
$$1 = c\left[e^{x-1} - e^{-x}\right]_{x=0}^{1}$$
$$1 = c(e^{1-1} - e^{-1} - (e^{0-1} - e^{-0}))$$
$$c = \frac{1}{1 - e^{-1} - (e^{-1} - 1)} = \frac{1}{2 - \frac{2}{e}}$$

b.

$$P(X > 0.75) = \int_{0.75}^{1} c(e^{x-1} + e^{-x})dx$$
$$= c\left[e^{x-1} - e^{-x}\right]_{x=0.75}^{1}$$
$$= c\left(e^{1-1} - e^{-1} - (e^{0.75-1} - e^{-0.75})\right)$$
$$= c\left(1 - e^{-1} - e^{-0.25} + e^{-0.75}\right) = \frac{1 - e^{-1} - e^{-0.25} + e^{-0.75}}{2 - \frac{2}{e}}$$

2 Problems

2.1 More Bit Strings

Once again, we're sending sending bit strings across potentially noisy communication channels, just like last week. However, this week we're identifying bit string corruptions in a slightly different way. Now, whenever we want to send n bits of information, we send an extra as the $n + 1^{st}$ bit. Specifically, if the sum of the n data bits is even, the extra $n + 1^{st}$ bit sent is set to 0. If the sum of the n data bits is odd, then $n + 1^{st}$ bit appended is set to 1. If the recipient of the bit string adds all bits and gets an odd number, that recipient knows there's a problem and can request a repeat transmission. We'll assume that each bit is erroneously inverted with probability nonzero $p \leq 0.5$, and that all bit corruptions are independent of one another.

a. Assuming that n = 4 and p = 0.1, what is the probability the transmitted message has errors that go undetected?

b. For arbitrary n and p, what is the probability that a bit string has errors that go undetected? You may leave it as a sum of O(n) terms.

c. Simplify your answer from part b by letting $a = \sum_{\text{odd}k} \binom{n+1}{k} p^k (1-p)^{n+1-k}$ and $b = \sum_{\text{even}k} \binom{n+1}{k} p^k (1-p)^{n+1-k}$ and then considering what $a + b$ and $a - b$ equal. Leverage the fact that, in general, $(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$.

a. Note that a bit string with an odd number of errors will be flagged as erroneous, but those bit strings with an even number of errors, regardless of n, will be taken as correct. That means the probability errors will go undetected is:

$$\binom{5}{2}(0.1)^2(0.9)^3 + \binom{5}{4}(0.1)^4(0.9)^1 = 0.07335 \tag{1}$$

b. The above generalizes to arbitrary n, so that the probability of interest is:

$$\sum_{\text{even } k \geq 2} \binom{n+1}{k} p^k (1-p)^{n+1-k} \tag{2}$$

c. If a and b are defined that way, then:

$$a + b = \sum_{\text{all} k} \binom{n+1}{k} p^k (1-p)^{n+1-k} = 1 \tag{3}$$

and

$$a - b = \sum_{\text{all} k} \binom{n+1}{k} (-p)^k (1-p)^{n+1-k} = (1-2p)^{n+1} \tag{4}$$

Solving for a, we arrive at:

$$a = \sum_{\text{even} k} \binom{n+1}{k} (-p)^k (1-p)^{n+1-k} = \frac{1 + (1-2p)^{n+1}}{2} \tag{5}$$

Now, a includes a term for no errors, so we need to subtract that one term off so it doesn't contribute. That leaves us with our final answer for general n and p, which is:

$$\frac{1 + (1-2p)^{n+1}}{2} - (1-p)^{n+1} \tag{6}$$

2.2 Air Quality

Throughout the United States, the Environmental Protection Agency monitors levels of PM2.5, a type of dangerous air pollution. These PM2.5 measurements can be approximately modeled by a normal distribution.

a. Let us model PM2.5 measurements with a normal distribution that has a mean of 8. If three-quarters of all measurements fall below 11.4, what is the standard deviation? Round to the nearest integer.

b. PM2.5 values above 12 can pose some health risks, especially to sensitive populations. Using the standard deviation found above, what is the probability that a randomly selected PM2.5 measurement is over 12?

c. What is the probability that a randomly selected PM2.5 measurement is between 7 and 8?

a. $\Phi(\frac{x-\mu}{\sigma}) = \Phi(\frac{11.4-8}{\sigma}) = 0.75 \implies \frac{3.4}{\sigma} \approx .68 \implies \sigma \approx 5.$

b. $P(q > 12) = 1 - P(q < 12) = 1 - \Phi(\frac{12-8}{5}) = 1 - \Phi(.8) = 1 - 0.7881 = 0.2119.$

c. $P(7 < h < 8) = P(h < 8) - P(h < 7) = \Phi(\frac{8-8}{5}) - \Phi(\frac{7-8}{5})$
$= \Phi(\frac{8-8}{5}) - \Phi(\frac{-1}{5})) = \Phi(\frac{8-8}{5}) - (1 - \Phi(\frac{1}{5}))$
$= \Phi(0) - (1 - \Phi(0.2)) = 0.5 - (1 - 0.5793) = 0.0793$

3 Previous Exam Questions

3.1 Winter 2021: Quiz 2

When a patient has eye inflammation, eye doctors "grade" the inflammation. When "grading" inflammation they randomly look at a single 1 millimeter by 1 millimeter square in the patients eye and count how many "cells" they see.

There is uncertainty in these counts. If the true average number of cells for a given patients eye is 6, the doctor could get a different count (say 4, or 5, or 7) just by chance. As of 2021, modern eye medicine does not have a sense of uncertainty for their inflammation grades! In this problem we are going to change that. At the same time we are going to learn about poisson distributions over space.

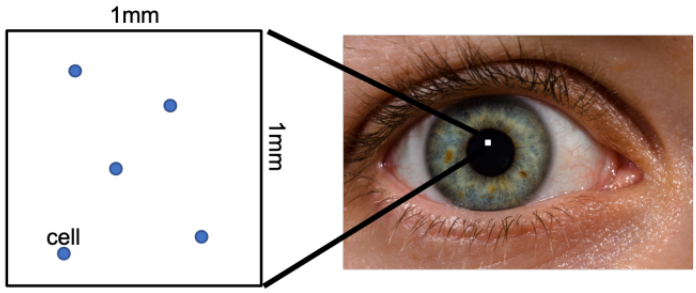

Figure 1: A 1x1mm sample used for inflammation grading. Inflammation is graded by counting cells in a randomly chosen 1mm by 1mm square. This sample has 5 cells

a. Explain, as if teaching, why the number of cells observed in a 1x1 square is governed by a poisson process. Make sure to explain how a binomial distribution could approximate the count of cells. Explain what λ means in this context. Note: for a given persons eye, the presence of a cell in a location is independent of the presence of a cell in another location. 100 word limit. Pictures not necessary, but allowed.

b. For a given patient the true average rate of cells is 5 cells per 1x1 sample. What is the probability that in a single 1x1 sample the doctor counts 4 cells? In addition to providing an expression, please compute a numeric answer.

c. For a given patient the true average rate of cells is 5 cells per 1mm by 1mm sample. In an attempt to be more precise, the doctor counts cells in **two** different, larger **2mm by 2mm** samples. Assume that the occurrences of cells in one 2mm by 2mm samples are independent of the occurrences in any other 2mm by 2mm samples. What is the probability that she counts 20 cells in the first samples and 20 cells in the second? In addition to providing an expression, please compute a numeric answer.

a. We can approximate a distribution for the count by discretizing the square into a fixed number of equal sized buckets. Each bucket either has a cell or not. There-

fore, the count of cells in the 1x1 square is a sum of Bernoulli random variables with equal p, and as such can be modeled as a binomial random variable. This is an approximation because it doesnt allow for two cells in one bucket. Just like with time, if we make the size of each bucket infinitely small, this limitation goes away and we converge on the true distribution of counts. The binomial in the limit, i.e. a binomial as $n \to \infty$, is truly represented by a Poisson random variable. In this context, λ represents the average number of cells per 1x1 sample. See Figure 2 below.

b. Let X denote the number of cells in the 1x1 sample. We note that $X \sim Poi(5)$. We want to find $P(X = 4)$.

$$P(X = 4) = \frac{5^4 e^{-5}}{4!} \approx 0.175$$

c. Let Y_1 and Y_2 denote the number of cells in each of the 2x2 samples. Since there are 5 cells in a 1x1 sample, there are 20 samples in a 2x2 sample since the area quadrupled, so we have that $Y_1 \sim Poi(20)$ and $Y_2 \sim Poi(20)$. We want to find $P(Y_1 = 20 \wedge Y_2 = 20)$. Since the number of cells in the two samples are independent, this is equivalent to finding $P(Y_1 = 20)P(Y_2 = 20)$.

$$P(Y_1 = 20 \wedge Y_2 = 20) = P(Y_1 = 20)P(Y_2 = 20) = \left(\frac{20^{20} e^{-20}}{20!}\right)^2 = 0.00789$$

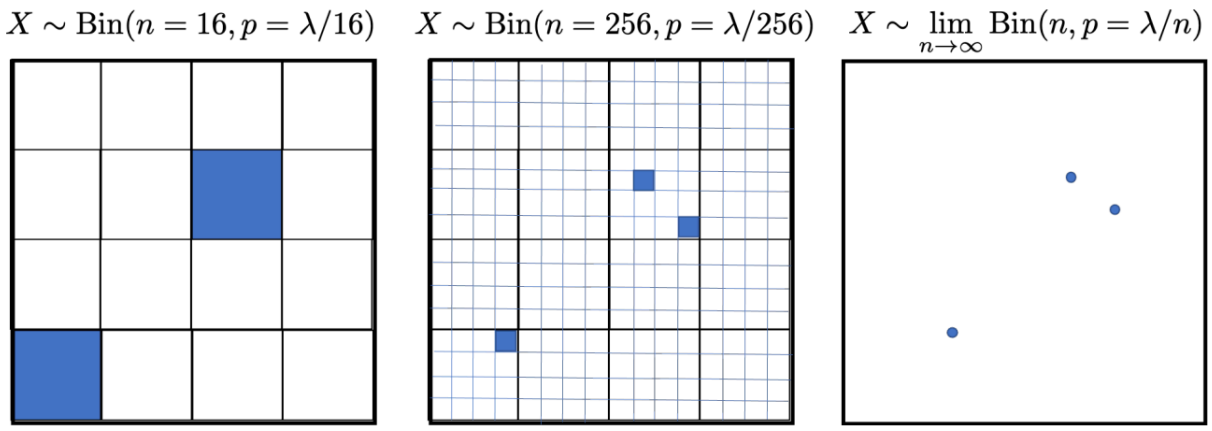

Figure 2: X is counts of events in discrete buckets. In the limit, as n (number of buckets)$\to \infty$, X becomes a Poisson.