Chris Piech
CS109

Section #3
Oct 18, 2023

Section #3: Random Variables

1 Gender Composition of Sections

A massive online Stanford class has sections with 10 students each. Each student in our population has a 50% chance of identifying as female, 47% chance of identifying as male and 3% chance of identifying as non-binary. Even though students are assigned randomly to sections, a few sections end up having a very uneven distribution just by chance. You should assume that the population of students is so large that the percentages of students who identify as male / female / non-binary are unchanged, even if you select students without replacement.

a. Define a random variable for the number of people in a section who identify as female.

Let X denote the number of people in a section who identify as female.
$X \sim \text{Bin}(n = 10, p = 0.5)$

b. What is the expectation and standard deviation of number of students who identify as female in a single section?

$E[X] = n \cdot p = 10 \cdot 0.5 = 5$
$\text{Std}(X) = \sqrt{\text{Var}(X)} = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{10 \cdot 0.5 \cdot 0.5} \approx 1.6$

c. Write an expression for the exact probability that a section is skewed. We defined skewed to be that the section has 0, 1, 9 or 10 people who identify as female.

Recall that $p = 0.5$

$$P(\text{skewed}) = P(X=0) + P(X=1) + P(X=9) + P(X=10)$$
$$= \binom{10}{0}(1-p)^{10} + \binom{10}{1}p(1-p)^9 + \binom{10}{9}p^9(1-p) + \binom{10}{10}p^{10} \quad \approx 0.021$$

```
import scipy.stats as st
import numpy as np
st.binom(10, 0.5).pmf(np.array([0,1,9,10])).sum()
0.02148437500000002
```

d. The course has 1,200 sections. Approximate the probability that 5 or more sections will be skewed.

The exact probability of number of skewed sections is $S \sim \text{Bin}(n = 1200, p = 0.021)$. This will require excessive calculations to reason about. Instead, we can approximate the number of skewed sections using a Poisson approximation. Let Y be the Poisson approximation of S.
$Y \sim \text{Poi}(\lambda = 25.2)$ since $n \cdot p = 1200 \cdot 0.021 = 25.2$

$$\begin{aligned} P(Y \geq 5) &= 1 - P(Y < 5) \\ &= 1 - \Big(P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 4)\Big) \\ &> 0.9999 \end{aligned}$$

2 Better Evaluation of Eye Disease

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 patient's eye and count how many ”cells” they see.

There is uncertainty in these counts. If the true average number of cells for a given patient's 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.

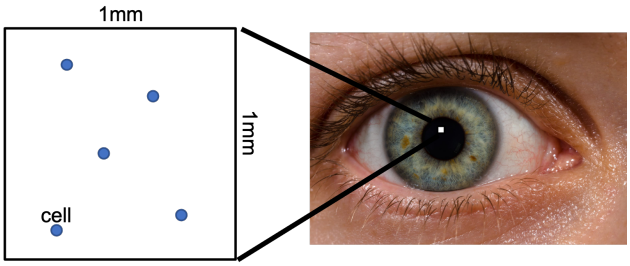

Figure 1: *A 1x1mm sample used for inflammation grading. Inflammation is graded by counting cells in a randomly chosen 1mm by 1mm square. This sample has 5 cells.*

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 person's eye, the presence of a cell in a location is independent of the presence of a cell in another location.

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. Therefore, 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 doesn't 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 1×1 sample. See Figure 2.

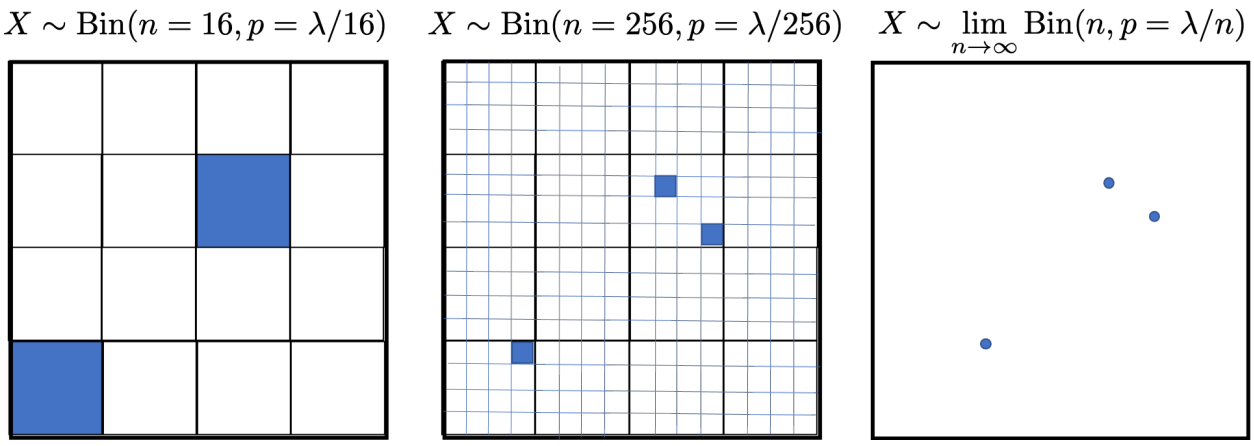

Figure 2: *X is counts of events in discrete buckets. In the limit, as n (number of buckets)* $\to \infty$*, X becomes a Poisson.*

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?

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

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

3 Website Visits

If this problem doesn't convince you that the Poisson and Exponential RVs are coupled, then I'm not sure what will!

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.

a. Using the random variable X, defined as the length of time a user stays on your website, what is the probability that a user stays longer than 10 mins?

The problem tells us that length of stay, X, is exponentially distributed. To define λ, we need to choose a unit time. If we choose 1 minute, $X \sim Exp(\lambda = \frac{1}{5})$:

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

We could also choose other units of time, such as 10 minutes. Then, $X \sim Exp(\lambda = 2)$:

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

b. Using the random variable Y, defined as the number of users who leave your website over a 10-minute interval, what is the probability that a user stays longer than 10 mins?

Y is the number of users leaving the website in the next 10 minutes. If one user leaves every 5 minutes on average, then the average number of users leaving every 10 minutes is 2. $Y \sim \text{Poi}(\lambda = 2)$.

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

In this case, we had only one choice for the unit time to define λ for, because only a unit time of 10 minutes allowed us to then use the PMF to ask for the probability of 0 users leaving in a 10 minute time window.

4 Continuous Random Variables

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

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

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

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

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

b. What is $P(X < 0.75)$? What is $P(X < x)$?

$$
\begin{aligned}
P(X < 0.75) &= \int_0^{0.75} c(e^{x-1} + e^{-x})dx \\
&= c\left[e^{x-1} - e^{-x}\right]_{x=0}^{0.75} \\
&= c\left((e^{0.75-1} - e^{-0.75}) - (e^{0-1} - e^{-0})\right)
\end{aligned}
$$

$$
\begin{aligned}
P(X < x) &= \int_0^{x} c(e^{y-1} + e^{-y})dy \\
&= c\left[e^{y-1} - e^{-y}\right]_{x=0}^{x} \\
&= c\left((e^{x-1} - e^{-x}) - (e^{0-1} - e^{-0})\right)
\end{aligned}
$$