CHEM 116 March 7, 2018

Unit 5, Lecture 2

Numerical Methods and Statistics

Companion Reading

Bulmer Chapter 6

1 Working With Distributions

1.1 Computing the Probability of a Sample

You are on Facebook. You see a post. The probability you like a post is 0.10. What kind of distribution is this? Bernoulli because there are two outcomes and fixed probability for each. [Draw a Picture]

You scroll down Facebook and decide to stop once you like a single post. Again you like a post with probability 0.1. What is the probability you liked a single post and stopped after seeing 10 posts? This is a geometric distribution because we stop as soon as we succeed and the number of trials is unbounded. [Draw a Picture]

$$P(n = 10) = (1 - 0.1)^{10-1}(0.1) = 0.039$$

You no longer stop after liking a post, you just scroll through a fixed number. What is the probability you liked 3 posts after seeing 10 posts? This is a binomial distribution because it has a fixed number of trials and the order does not matter. [Draw a Picture]

$$P(n = 3) = \binom{10}{3}(1 - 0.1)^{10-3}(0.1)^3 = 0.057$$

You've become very cynical and only like 2% of posts now. You scroll through 250 posts. What is the probability you liked 4 posts? This is a Poisson distribution because it is the same as set-up as a binomial but the probability is low and the trial number is high. First, to get the parameter for the Poisson, μ, we must compute it. It's the expected number of successes $\mu = Np = 250 \times 0.02 = 5.0$ [Draw a Picture]

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

1.2 Computing the Probability of an Interval

All of these examples have small probabilities because we're looking at single points. What about intervals.

Return to the geometric example. What's the probability that you liked a post and stopped scrolling in less than 10 posts? [Draw a picture]

$$P(n = 1) + P(n = 2) + ... = \sum_{i=1}^{9} P(n = i)$$

Using Python...

$$P(n < 10) = 0.724$$

What about the opposite? What's the probability it took more than or 10 posts?

$$P(n >= 10) = 1 - P(n < 10) = 0.276$$

[Draw a picture]. An example of the NOT rule.

We can use the same equation, $\sum_{i=1}^{N} P(i)$, to compute these intervals with discrete probability mass functions.

1.3 Computing the Probability of an Interval for Continuous Distributions

Let's assume Facebook posts arrive according to an exponential distribution. This makes sense because time is continuous, cannot be negative, and exponential distributions describe the time between random events. If on average 10 posts occur per hour, λ the exponential parameter is 1/6 inverse minutes. What is the probability of a new post within the 5 minutes? [Draw a picture]

$$\mathrm{P}(t < 5) = \int_0^5 \lambda e^{-\lambda t}\, dt = -\frac{\lambda}{\lambda}\, e^{-\lambda t}\Big]_0^5 = 0.63$$

What is the probability a new post arrives after 2 minutes but before 10 minutes?

$$\mathrm{P}(2 < t < 10) = \int_2^{10} \lambda e^{-\lambda t}\, dt = 0.53$$

[Draw a picture]

What is the probability the new post comes after 4 minutes?

$$\mathrm{P}(t > 4) = \int_4^{\infty} \lambda e^{-\lambda t}\, dt = \, -e^{-\lambda t}\Big]_4^{\infty} = 0.51$$

[Draw a picture]

You decide to stop playing on the Internet so much, but you only like going outside if the temperature is greater than 65 degrees and less then 80 degrees. Assume the temperature has a mean of 45 degrees in Rochester and the standard deviation is 15 degrees. Assume you measure the temperature once per day. What's the probability you'll go outside?

$$\mathrm{P}(65 < T < 80) = \int_{65}^{80} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = 0.15$$

Note, the above equation must be evaluated numerically. [Draw a picture]

1.4 Finding Prediction Intervals

Let's return to the geometric distribution. What if we're interested in solving this equation? [Draw a Picture]

$$\operatorname*{argmin}_x P(1 \geq n \leq x) = 0.9$$

What is the number of trials for which there is a 90% probability we succeeded? This called a *Prediction Interval*. This generally must be solved with Python. The answer is 15. Compare that with the expected value, which is 10. The answer to the question: "How long will it take to find a post you like?" is about 10. A prediction interval allows us to give a worst-case value, where we are wrong about it 10% of the time. We state it as "With a prediction level of 10%, the number of posts I will see before liking one is 15".

The interval should also exceed or be equal to the desired probability. The reason for the argmin_x is that we could trivially choose x to be ∞. When we work with other distributions, the lower point can change. For example, in a binomial we would have:

$$\operatorname*{argmin}_x P(0 \leq n \leq x) = 0.95$$

for a 95% prediction interval. When the lower-bound is the lowest value in the sample space (i.e., 0 for binomial or 1 for geometric), this is called a “lower” prediction interval. We can swap where x goes to create an “upper” prediction interval. For example, again with geometric:

$$\operatorname*{argmax}_x P(x \leq n \leq \infty) = 0.80$$

Notice we now have an argmax_x. In words, we’re asking at what value will 80% of samples fall above.