CHEM 116 February 14, 2017

Unit 5, Lecture 1

Numerical Methods and Statistics

1 Distributions

Some random variables/sample spaces are so common, we have a name and equations that describe them. For example, consider the simplest probability mass function/distribution:

1.1 Bernoulli Distribution

The Bernoulli distribution has a sample space of size 2 (e.g., 0, 1) with probability:

$$\mathrm{P}(X=0)=p,\ \mathrm{P}(X=1)=1-p \tag{1}$$

The expected value is p and the variance is $p(1-p)$. An example of a process that follows a Bernoulli distribution is flipping a coin once.

There are three important things to recall about a distribution: its **sample space type** (continuous/discrete), its **support** or **sample space**, and what it describes. You should remember these things for many common distributions and then look up the details such as their variance in your textbook or online. Support (the probability is greater than 0) is

$$x \in \mathrm{P}(X=x)>0 \tag{2}$$

Support is important when the sample space contains zero probability elements. For example, if the probability of success in a binomial is 0, we essentially have removed that element from the sample space.

1.2 Geometric Distribution

A geometric distribution is discrete and has a sample space of $[1,\infty)$. It is the number of trials required until success. Specifically, we have a Bernoulli process where the probability of success is a constant p and n is the number of failures until a success is achieved:

$$\mathrm{P}(X=n)=(1-p)^{n-1}p \tag{3}$$

The expected value is $1/p$ and the variance is:

$$\frac{1-p}{p^2} \tag{4}$$

An example is the number of times until a heads is seen while flipping a coin. Notice that the Geometric is unbounded in sample space. This makes it simple to distinguish between the other discrete probability distributions. For example, a Bernoulli distribution could be if you like a Facebook post or not. There are two outcomes. A Geometric distribution could be the number of Facebook posts you view before you like one. Depending on how cynical you are, there is no upper bound on the number you view.

1.3 Binomial Distribution

A binomial distribution is discrete and has a sample space of $[0, N]$. It is the number of successes in N trials. It is an important distribution for considering permutations. Its equation is:

$$\mathrm{P}(X = n) = \binom{N}{n} p^n (1-p)^{N-n} \tag{5}$$

and the expected value is Np. The variance is $Np(1-p)$. An example is the number of boys in a family of eight children.

An easy way to distinguish a Bernoulli from a Binomial is to think about if the sample space has more than 2 elements. For example, I may retweet a tweet or not. That's a Bernoulli. If I examine 5 tweets and I may retweet any of them, the number of retweets I make is a Binomial, since it can vary from 0 to 5 (6 possibilities).

1.4 Poisson Distribution

A Poisson distribution is discrete and has a sample space of $[0, \infty)$. It can be thought of an approximation to the Binomial distribution in the case of very low chances of success ($p << 0.5$), where $\mu = Np$. Its equation is

$$\mathrm{P}(X = x) = \frac{e^{-\mu}\mu^x}{x!} \tag{6}$$

and the expected value is μ. The variance is also μ. An example is the number of deaths by horse kick in the 14 corps of the Prussian Army. It can also be viewed as the number of events in a fixed time interval. For example, if we are watching a live stream of an event, the number of comments could be modeled by a Poisson distribution.

1.5 Exponential Distribution

The exponential distribution is a continuous distribution with a support of $[0, \infty)$. It generally describes time between events, especially events whose occurrence follow the Poisson distribution (rare events). Its distribution is

$$\mathrm{P}(X = x) = \lambda e^{-\lambda x} \tag{7}$$

and the expected value is $1/\lambda$. You may also think of it in terms of a residence time $\tau = 1/\lambda$. The variance is λ^{-2} or τ^2. Note that we should imagine the process in terms of its cumulative probability (see drawing) and not instantaneous pdf.

Following our previous example of a live stream, the time between comments will follow an exponential distribution if the comments follow a Poisson distribution. Even though the Poisson is discrete, the time between comments is continuous since time is a continuous value.

1.6 Normal Distribution

The normal distribution is a continuous distribution with a support of $(-\infty, \infty)$. We will see why later in the semester, but a normal distribution is almost always a good guess for the description of a process if it is continuous. Its equation is:

$$\mathrm{P}(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{8}$$

2 Working With Distributions

2.1 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]

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

2.2 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.

2.3 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?

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

[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]

2.4 Finding Prediction Intervals

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

$$P(n < x) = 0.9$$

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".