Continuous Probability Notes

CMS 380 Simulation and Stochastic Modeling

Probability Density Functions

Discrete to Continuous

In the previous unit, we defined the *probability mass function* of a discrete random variable X:

$$p(x) = P(X = x)$$

The pmf was defined as the probability that random variable X took on value x (note: capitalized X is the random variable and little x is a particular value in its sample space.

Now, we're ready to consider *continuous* distributions, where X can take values over all or part of the real line.

The continuous counterpart of the pmf is the *probability density function* (PDF), denoted $f_X(x)$. It denotes the *relative likelihood* that random variable X takes on value x. All questions about the behavior of a continuous random variable can be answered using its PDF.

Example: the Normal Distribution

Figure 1 shows the PDF of most well-known continuous distribution, the grand-daddy-of-them-all: the *standard normal distribution*, which has the famous bell-shaped curve that you've probably seen before.

The PDF shows that this random variable is *most likely* to take values close to zero. The relative likelihood drops away quickly as you move away from zero: values outside the range $[-3.5, 3.5]$ have negligible probability.

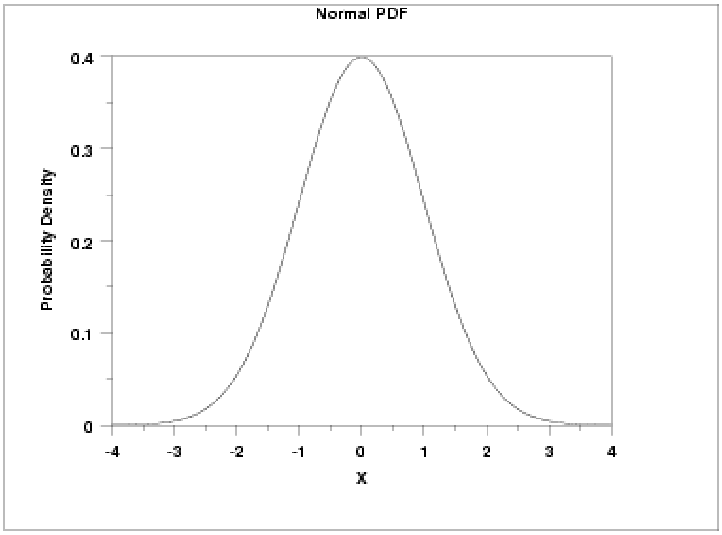

Figure 1: The PDF of a normally distributed random variable (NIST image)

The Weirdness of Continuous Probability

Consider a lightbulb with a lifetime that can be modeled as a continuous random variable. What is the probability that the bulb survives for exactly 200 days and then fails?

Here's the first weird thing about continuous probability: you can't reason about the probability of one single point in the sample space. If X is continuous, there are *infinitely many* possible values that it can take, so the probability of any one exact value must be 0. Therefore, the probability that our lightbulb works for *exactly* 200.0000000... days must be 0.

It does make sense, though, to think about X taking a value that lies in a certain range. For example, we might consider the probability that the lightbulb's life is in the interval $[200.0, 200.0001]$. Formally, the probability that X lies in an interval $[a, b]$ is defined by *integrating* over the PDF:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

Calculus is not a prerequisite for this class, so I won't require you to calculate any integrals.

CDFs and CCDFs

The *cumulative distribution function* (CDF) of X, denoted $F_X(x)$, is the probability that X is less than x. Formally,

$$F_X(x) = P(X \leq x)$$

The CDF is related to the PDF by integration:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

If you want even more calculus in your life, you can note that the PDF is therefore the derivative of the CDF:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Again, I'm not going to make you calculate integrals or derivatives in this class, so this is mainly background information.

What if you want to reason about the probability X is greater than some value x? This is *complementary cumulative distribution function* (CCDF):

$$\overline{F}_X(x) = P(X > x)$$

In terms of the CDF:

$$\overline{F}_X(x) = 1 - F_X(x)$$

The next section will show examples of using the CDF and CCDF to solve problems.

The Exponential Distribution

The exponential distribution is the most important probability distribution in system analysis and queueing theory. Its PDF has one parameter, $\lambda > 0$, and is given by

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{if } x \geq 0$$

$f_X(x) = 0$ if $x < 0$, so the exponential random variable can only take on non-negative values.

The CDF can be calculated by integrating, as described above.

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \int_0^x \lambda e^{-\lambda u} du \\ &= 1 - e^{-\lambda x} \end{aligned}$$

The CCDF is therefore

$$\begin{aligned} \overline{F}_X(x) &= 1 - F_X(x) \\ &= e^{-\lambda x} \end{aligned}$$

Expected Values

Recall that the expected value of a discrete random variable was the weighted sum of its possible values.

$$E[X] = \sum_(x : p(x) > 0) x p(x)$$

The same definition applies to continuous random variables, with the appropriate conversion of discrete summation to integration:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

The variance of a continuous random variable has the same definition as a discrete random variable:

$$\sigma_X^2 = E[X^2] - E[X]^2$$

From the definition, the expected value of an exponential variable is

$$E[X] = \int_0^{\infty} x \lambda e^{\lambda x} dx$$

If you want to integrate this, you have to use integration by parts ($u = x$ and $v = e^{-\lambda x} dx$)[1]. Performing the necessary calculations will show that

$$E[X] = \frac{1}{\lambda}$$

[1] Yuck.

The expected value of an exponentially distributed random variable is the inverse of its parameter λ.

Treadmill lifetime. *The lifetime of a treadmill is exponentially distributed with parameter $\lambda = .25$ years. What is the expected lifetime of the treadmill in years?*

If you know λ, it's easy to calculate the expected value.

$$E[X] = \frac{1}{.25} = 4 \text{ years}$$

Parameter estimation. *The time between customer arrivals to a system is exponentially distributed. The average time between arrivals is 10 minutes. What is the parameter of the exponential interarrival distribution?*

This is a common case. You measure a system and observe an average value over time, then want to find the parameter of the appropriate exponential distribution. Use the definition of expected value again:

$$\lambda = \frac{1}{E[X]} = .10$$

Example Problems

The exponential distribution shows up frequently in reliability and lifetime modeling. It is used to represent devices or components that can file at *random* times.

Bulb. *The lifetime of an industrial lightbulb is exponentially distributed with mean 3000 hours. What is the probability that a bulb survives for no more than 1000 hours?*

Let X be the exponential random variable representing the lifetime of the bulb. First, determine the parameter of the distribution:

$$\lambda = \frac{1}{3000} = .000333\ldots$$

We would like to know $P(X \leq 1000)$. Use the exponential CDF:

$$
\begin{aligned}
F_X(1000) &= 1 - e^{-\frac{1}{3000}1000} \\
&\approx .28347
\end{aligned}
$$

About 28% of bulbs will live for less than 1000 hours.

What is the probability that the bulb survives for more than 5000 hours?

A similar problem. Use the CCDF to get $P(X > 5000)$.

$$
\begin{aligned}
\overline{F}_X(5000) &= e^{-\frac{1}{3000}5000} \\
&\approx .18887
\end{aligned}
$$

Almost 19% of bulbs survive more than 5000 hours.

What is the probability that the bulb survives between 1000 and 2000 hours?

The goal is to find $P(1000 \leq X \leq 2000)$. You can define this probability in terms of the CDF:

$$P(1000 \leq X \leq 2000) = P(X \leq 2000) - P(X \leq 1000)$$

If the lifetime is *between* 1000 and 2000 it must be *less than 2000* (this is the first term), but *not less than 1000* (this is the second term).

The Memoryless Property

Consider the following questions:

- Lifetimes of jobs in a computer system are exponentially distributed with mean fifteen minutes. A job has already been running for one hour. What is the probability that its total running time will exceed two hours?
- The time between customer arrivals at a system is exponentially distributed with mean five minutes. We have been waiting five minutes with no arrival. What is the probability that the total time until the next arrival exceeds eight minutes?

- The lifetime of an industrial printing press is exponentially distributed with mean ten years. If the press has been in operation for five years, what is the probability that its lifetime exceeds twelve years?

All of these questions are the same from a modeling perspective: we have an exponentially distributed lifetime that we have already observed for a certain amount of time.

Let t denote the already-observed time that has passed. For example, for the computer job lifetimes, $t = 1$ hour. For the customer arrivals, $t = 5$ minutes.

We'd like to reason about the probability that the lifetime exceeds some value $t+s$. For the computer jobs, $s = 1$ additional hour. For the customer arrivals, $s = 3$ additional minutes. For the printing press, $s = 7$ additional years.

Mathematically, this is a conditional probability problem:

$$P(X > t + s | X > t)$$

In words: what is the probability that the system survives beyond time $t + s$, given that it's already survived up to time t?

Let's attack this problem using the definition of conditional probability and the exponential CCDF to see what happens.

$$P(X > t + s | X > t) = \frac{P(X > t + s \text{ and } X > t)}{P(X > t)}$$

Consider the numerator: if $X > t + s$, then $X > t$ is automatically implied, so the *and* is unnecessary.

$$P(X > t + s | X > t) = \frac{P(X > t + s)}{P(X > t)}$$

Using the exponential CCDF:

$$P(X > t + s | X > t) = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}}$$

Using the rules of exponents:

$$P(X > t + s | X > t) = \frac{e^{-\lambda t} e^{-\lambda s}}{e^{-\lambda t}}$$

The $e^{-\lambda t}$ terms cancel, leaving the result:

$$P(X > t + s | X > t) = e^{-\lambda s}$$

This result seems wonky and technical, but it turns out to be surprising. Notice:

- The left-hand side defines a probability conditioned on t. We're interested in reasoning about the future lifetime of a system that we've already observed for a period of length t, so conditioning on t would seem to be rather important.
- The right-hand side **does not contain** t.

The conclusion therefore is this: *the probability of surviving until time* $t + s$ ***depends only on*** s *and* ***does not depend on*** t.

This result is called **the memoryless property of the exponential distribution**. The exponential is the only continuous distribution that has this property.

The memoryless property is tricky to understand. It tells you that the probability of continuing for at least s *additional* units of time *is the same* as the probability that a brand-new job runs for at least s.

- In the computer program example, P(job runs for more than one additional hour) is equal to P(a new job runs for more than one hour). The amount of time the job has already been running does not affect this probability.
- In the printing press example, P(press runs for at least 7 additional years) is equal to P(a new press runs for at least seven years). The amount of time the press has already been in service does not affect this probability.

Therefore, the memoryless property makes it easy to reason about the future behavior of exponentially distributed lifetimes. The amount of time the system has been operating has no effect on its future lifetime, so we don't need to use conditional probability. This behavior is counterintuitive.

- Suppose you're waiting for a bus, and that buses run on a very tight schedule so that one bus arrives every ten minutes on average with little variation. If you've already been waiting for nine minutes, you would expect the next bus to be along very soon. This makes sense if the distribution of waiting times has low variability: the longer you observe the system, the more likely it is that the next event will happen soon.

- Human lifetimes have the opposite property. In general, the longer person has been alive, the more likely they are to stay alive. This is less of a concern nowadays, but in traditional societies, a person who made it out of childhood had a high probability of making it to at least young adulthood, and a person who survived being a young adult had a good chance of living to a relatively advanced age.

- In the memoryless property, failure is *random*. The amount of time that a system has been in operation gives you no help in predicting its future behavior.

Residual service times. *Consider a queue with exponentially distributed service times. The average service time is measured to be* $\overline{s} = 10$ *minutes. You arrive to this queue and find that a customer has already been receiving service for 10 minutes. How much longer would you expect to wait before the customer in service finishes and departs? What if the customer in service had only been there for one minute?*

The service times in this system are exponential, so they are memoryless. Therefore, the amount of time the customer has already been in service has no effect on its future behavior. In both cases, the expected future service time is the same as the expected service time for a new customer, so you would expect to wait 10 minute in both cases.

More Bulbs. *Suppose that an industrial lightbulb has an exponentially distributed lifetime with an expected value of 5000 hours. Suppose that a bulb has already been in service for 4000 hours. What is the probability that it expires before a total of 6000 hours have elapsed?*

Hint: we're interested in the probability that the additional lifetime is less than 2000 hours.