6. Continous Distributions

Chris Piech and Mehran Sahami

Oct 2017

So far, all random variables we have seen have been *discrete*. In all the cases we have seen in CS109 this meant that our RVs could only take on integer values. Now it's time for *continuous* random variables which can take on values in the real number domain ($\mathbb{R}$). Continuous random variables can be used to represent measurements with arbitrary precision (eg height, weight, time).

1 From Discrete to Continuous

To make our transition from thinking about discrete random variable, to thinking about continuous random variables, lets start with a thought experiment: Imagine you are running to catch the bus. You know that you will arrive at 2:15pm but you don't know exactly when the bus will arrive, and want to think of the arrival time in minutes past 2pm as a random variable T so that you can calculate the probability that you will have to wait more than five minutes $P(15 < T < 20)$.

We immediately face a problem. For discrete distributions we would describe the probability that a random variable takes on exact values. This doesn't make sense for continuous values, like the time the bus arrives. As an example, what is the probability that the bus arrives at exactly 2:17pm and 12.12333911102389234 seconds? Similarly, if I were to ask you: what is the probability of a child being born with weight **exactly** $= 3.523112342234$ kilos, you might recognize that question as ridiculous. No child will have precisely that weight. Real values can have infinite precision and as such it is a bit mind boggling to think about the probability that a random variable takes on a specific value.

Instead, let's start by discretizing time, our continuous variable, by breakint it into 5 minute chunks. We can now think about something like, the probability that the bus arrives between 2:00p and 2:05 as an event with some probability (see figure 1, left). Five minute chunks seem a bit coarse. You could imagine that instead, we could have discretized time into 2.5minute chunks (figure 1, center). In this case the probability that the bus shows up between 15 mins and 20 mins after 2pm is the sum of two chunks, shown in orange. Why stop there? In the limit we could keep breaking time down into smaller and smaller pieces. Eventually we will be left with a derivative of probability at each moment of time, where the probability that $P(15 < T < 20)$ is the integral of that derivative between 15 and 20 (figure 1, right).

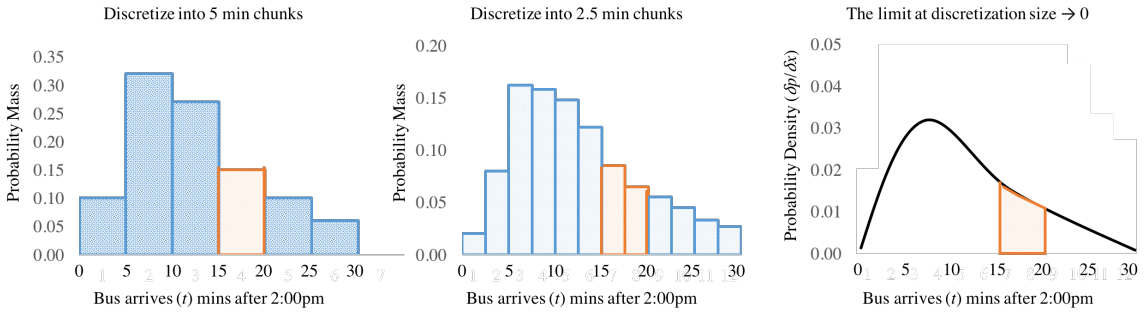


Figure 1: Bus thought process: Discrete to Continuous

2 Probability Density Functions

In the world of discrete random variables, the most important property of a random variable was its probability mass function (PMF) that would tell you the probability of the random variable taking on any value. When we move to the world of continuous random variables, we are going to need to rethink this basic concept.

In the continuous world, every random variable instead has a Probability *Density* Function (PDF) which defines the relative likelihood it is that a random variable takes on a particular value. Like in the bus example, the PDF is the derivative of probability at all points of the random variable. This means that the PDF has the important property that you can integrate over it to find the probability that the random variable takes on values within a range (a,b).

X is a Continuous Random Variable if there is a Probability Density Function (PDF) $f(x)$ for $-\infty \leq x \leq \infty$ such that:

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

The following properties must also hold. These preserve the axiom that $P(a \leq X \leq b)$ is a probability:

$$0 \leq P(a \leq X \leq b) \leq 1$$
$$P(-\infty < X < \infty) = 1$$

A common misconception is to think of $f(x)$ as a probability. It is instead what we call a probability density. It represents probability/unit of X. Generally this is only meaningful when we either take an integral over the PDF **or** we compare probability densities. As we mentioned when motivating probability densities, the probability that a continuous random variable takes on a specific value (to infinite precision) is 0.

$$P(X = a) = \int_a^a f(x)dx = 0$$

That is pretty different than in the discrete world where we often talked about the probability of a random variable taking on a particular value.

3 Cumulative Distribution Function

Having a probability density is great, but it means we are going to have to solve an integral every single time we want to calculate a probability. To avoid this unfortunate fate, we are going to use a standard called a cumulative distribution function (CDF). The CDF is a function which takes in a number and returns the probability that a random variable takes on a value less than that number. It has the pleasant property that, if we have a CDF for a random variable, we don't need to integrate to answer probability questions!

For a continuous random variable X the Cumulative Distribution Function, written $F(a)$ is:

$$F_X(a) = P(X \leq a) = \int_{-\infty}^a f(x)dx$$

Why is the CDF the probability that a random variable takes on a value **less than** the input value as opposed to greater than? It is a matter of convention. But it is a useful convention. Most probability questions can be

solved simply by knowing the CDF (and taking advantage of the fact that the integral over the range $-\infty$ to ∞ is 1. Here are a few examples of how you can answer probability questions by just using a CDF:

Probability Query	Solution	Explanation
$P(X < a)$	$F(a)$	That is the definition of the CDF
$P(X \leq a)$	$F(a)$	Trick question. $P(X = a) = 0$
$P(X > a)$	$1 - F(a)$	$P(X < a) + P(X > a) = 1$
$P(a < X < b)$	$F(b) - F(a)$	$F(a) + P(a < X < b) = F(b)$

The continuous distribution also exists for discrete random variables, but there is less utility to a CDF in the discrete world as none of our discrete random variables had "closed form" (eg without any summation) functions for the CDF:

$$F_X(a) = \sum_{i=1}^{a} P(X = i)$$

Example 1

Let X be a continuous random variable (CRV) with PDF:

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{when } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

In this function, C is a constant. What value is C? Since we know that the PDF must sum to 1:

$$\int_0^2 C(4x - 2x^2)dx = 1$$
$$C\left(2x^2 - \frac{2x^3}{3}\right)\Big|_0^2 = 1$$
$$C\left(\left(8 - \frac{16}{3}\right) - 0\right) = 1$$

And if you solve the equation for C you find that $C = 3/8$.

What is $P(X > 1)$

$$\int_1^\infty f(x)dx = \int_1^2 \frac{3}{8}(4x - 2x^2)dx = \frac{3}{8}\left(2x^2 - \frac{2x^3}{3}\right)\Big|_1^2 = \frac{3}{8}\left[\left(8 - \frac{16}{3}\right) - \left(2 - \frac{2}{3}\right)\right] = \frac{1}{2}$$

Example 2

Let X be a random variable which represents the number of days of use before your disk crashes with PDF:

$$f(x) = \begin{cases} \lambda e^{-x/100} & \text{when } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

First, determine λ. Recall that $\int e^u du = e^u$:

$$\int \lambda e^{-x/100}dx = 1 \Rightarrow -100\lambda \int \frac{-1}{100} e^{-x/100}dx = 1$$
$$-100\lambda e^{-x/100}\Big|_0^\infty = 1 \Rightarrow 100\lambda = 1 \Rightarrow \lambda = \frac{1}{100}$$

What is the $P(X < 10)$?

$$F(10) = \int_0^{10} \frac{1}{100} e^{-x/100} dx = -e^{-x/100}\Big|_0^{10} = -e^{-1/10} + 1 \approx 0.095$$

For a discrete random variable X, the Cumulative Distribution Function (CDF) is defined as:

$$F(a) = P(X \leq a) \text{ where } -\infty < a < \infty$$

4 Expectation and Variance

For continuous RV X:

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

For both continuous and discrete RVs:

$$E[aX + b] = aE[X] + b$$
$$\text{Var}(X) = E[(X - \mu)^2] = E[X^2] - (E[X])^2$$
$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

5 Uniform Random Variable

The most basic of all the continuous random variables is the uniform random variable, which is equally likely to take on any value in its range (α, β).

X is a Uniform Random Variable $X \sim Uni(\alpha, \beta)$ if it has PDF:

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{when } \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

Notice how the density $(1/(\beta - \alpha)$ is exactly the same regardless of the value for x. That makes the density uniform. So why is the PDF $(1/(\beta - \alpha)$ and not 1? That is the constant that makes it such that the integral over all possible inputs evaluates to 1.

The key properties of this RV are:

$$P(a \leq X \leq b) = \int_a^b f(x) dx = \frac{b - a}{\beta - \alpha} \text{ (for } \alpha \leq a \leq b \leq \beta)$$
$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha} dx = \frac{x^2}{2(\beta - \alpha)}\Big|_{\alpha}^{\beta} = \frac{\alpha + \beta}{2}$$
$$Var(X) = \frac{(\beta - \alpha)^2}{12}$$

6 Normal Random Variable

The single most important random variable type is the Normal (aka Gaussian) random variable, parametrized by a mean (μ) and variance (σ^2). If X is a normal variable we write $X \sim N(\mu, \sigma^2)$. The normal is important

for many reasons: it is generated from the summation of independent random variables and as a result it occurs often in nature. Many things in the world are not distributed normally but data scientists and computer scientists model them as Normal distributions anyways. Why? Because it is the most entropic (conservative) modelling decision that we can make for a random variable while still matching a particular expectation (average value) and variance (spread).

The Probability Density Function (PDF) for a Normal $X \sim N(\mu, \sigma^2)$ is:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Notice the x in the exponent of the PDF function. When x is equal to the mean (μ) then e is raised to the power of 0 and the PDF is maximized.

By definition a Normal has $E[X] = \mu$ and $Var(X) = \sigma^2$.

There is no closed form for the integral of the Normal PDF, and as such there is no closed form CDF. However we can use a transformation of any normal to a normal with a precomputed CDF. The result of this mathematical gymnastics is that the CDF for a Normal $X \sim N(\mu, \sigma^2)$ is:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Where Φ is a precomputed function that represents that CDF of the Standard Normal.

Linear Transform

If X is a Normal such that $X \sim N(\mu, \sigma^2)$ and Y is a linear transform of X such that $Y = aX + b$ then Y is also a Normal where:

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

Projection to Standard Normal

For any Normal X we can find a linear transform from X to the standard normal $\sim N(0,1)$. For any normal, if you subtract the mean (μ) of the normal and divide by the standard deviation (σ) the result is always the standard normal. We can prove this mathematically. Let $W = \frac{X-\mu}{\sigma}$:

$W = \frac{X-\mu}{\sigma}$	Transform X: Subtract by μ and diving by σ
$= \frac{1}{\sigma}X - \frac{\mu}{\sigma}$	Use algebra to rewrite the equation
$= aX + b$	Where $a = \frac{1}{\mu}$, $b = -\frac{\mu}{\sigma}$
$\sim N(a\mu + b, a^2\sigma^2)$	The linear transform of a Normal is another Normal
$\sim N(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2})$	Substituting values in for a and b
$\sim N(0,1)$	The standard normal

Using this transform we can express $F_X(x)$, the CDF of X, in terms of the known CDF of Z, $F_Z(x)$. Since the CDF of Z is so common it gets its own Greek symbol: $\Phi(x)$

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

The values of $\Phi(x)$ can be looked up in a table. We also have an online calculator.

Example 1

Let $X \sim \mathcal{N}(3,16)$, what is $P(X > 0)$?

$$\begin{aligned} P(X>0) &= P\left(\frac{X-3}{4} > \frac{0-3}{4}\right) = P\left(Z > -\frac{3}{4}\right) = 1 - P\left(Z \leq -\frac{3}{4}\right) \\ &= 1 - \Phi(-\frac{3}{4}) = 1 - (1 - \Phi(\frac{3}{4})) = \Phi(\frac{3}{4}) = 0.7734 \end{aligned}$$

What is $P(2 < X < 5)$?

$$\begin{aligned} P(2<X<5) &= P\left(\frac{2-3}{4} < \frac{X-3}{4} < \frac{5-3}{4}\right) = P\left(-\frac{1}{4} < Z < \frac{2}{4}\right) \\ &= \Phi(\frac{2}{4}) - \Phi(-\frac{1}{4}) = \Phi(\frac{1}{2}) - (1 - \Phi(\frac{1}{4})) = 0.2902 \end{aligned}$$

Example 2

You send voltage of 2 or -2 on a wire to denote 1 or 0. Let X = voltage sent and let R = voltage received. $R = X + Y$, where $Y \sim \mathcal{N}(0,1)$ is noise. When decoding, if $R \geq 0.5$ we interpret the voltage as 1, else 0. What is $P(\text{error after decoding} | \text{original bit} = 1)$?

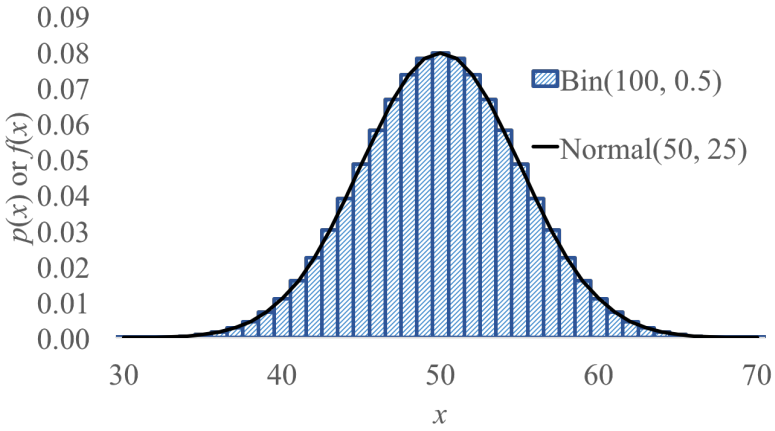
$$P(X+Y<0.5) == P(2+Y<0.5) = P(Y<-1.5) = \Phi(-1.5) = 1 - \Phi(1.5) \approx 0.0668$$

Binomial Approximation

Imagine this terrible scenario. You need to solve a probability question on a binomial random variable (see the chapter on discrete distributions) with a large value for n (the number of experiments). You quickly realize that it is way too hard to compute by hand. Recall that the binomial probability mass function has an $n!$ term. You decide to turn to you computer, but after a few iterations you realize that this is too hard even for your GPU boosted mega computer (or your laptop).

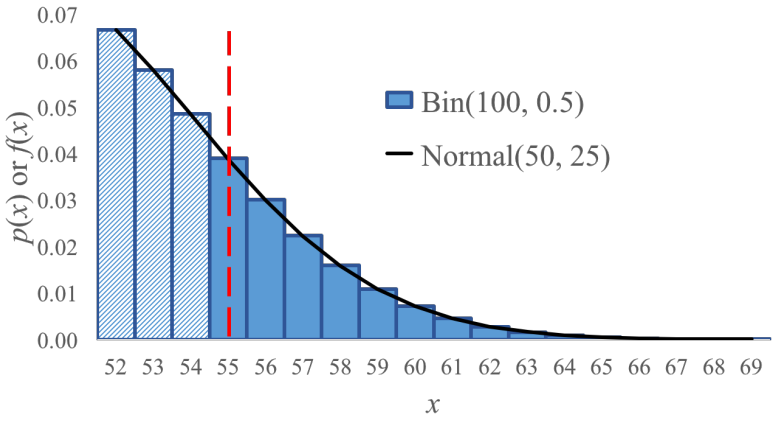
As a concrete example, imagine that in an election each person in a country with 10 million people votes in an election. Each person in the country votes for candidate A, with independent probability 0.53. You want to know the probability that candidate A gets more than 5 million votes. Yikes!

Don't panic (unless you are candidate B, then sorry, this election is not for you). Did you notice how similar a normal distribution's PDF and a binomial distributions PMF look? Lets take a side by side view:



Lets say our binomial is a random variable $X \sim \text{Bin}(100, 0.5)$ and we want to calculate $P(X \geq 55)$. We could cheat by using the closest fit normal (in this case $Y \sim N(50, 25)$). How did we chose that particular Normal? I simply selected one with a mean and variance that matches the Binomial expectation and variance. The binomial expectation is $np = 100 \cdot 0.5 = 50$. The Binomial variance is $np(1-p) = 100 \cdot 0.5 \cdot 0.5 = 25$.

Since $Y \approx X$ then $P(X \geq 55)$ seems like it should be $\approx P(Y \geq 55)$. That is almost true. It turns out that there is a formal mathematical reason why the normal is a good approximation of the binomial as long as the Binomial parameter p is reasonable (eg in the range [0.3 to 0.7]) and n is large enough. However! There was an oversight in our logic. Let's look a bit closer at the binomial we are approximating.



Since we want to approximate $P(X \geq 55)$, our goal is to calculate the sum of all of the columns in the Binomial PMF from 55 and up (all those dark columns). If we calculate the probability that the approximating Normal random variable takes on a value greater than 55 $P(Y \geq 55)$ we will get the integral starting at the vertical dashed line. Hey! That's not where the columns start. Really we want the area under the curve starting half way between 54 and 55. The correct approximation would be to calculate $P(X \geq 54.5)$.

Yep, that adds an annoying layer of complexity. The simple idea is that when you approximate a discrete distribution with a continuous one, if you are not careful your approximating integral will only include half of one of your boundary values. In this case we were only adding half of the column for $P(X = 55)$). The correction is called the continuity correction.

You can use a Normal distribution to approximate a Binomial $X \sim Bin(n, p)$. To do so define a normal $Y \sim (E[X], Var(X))$. Using the Binomial formulas for expectation and variance, $Y \sim (np, np(1-p))$. This approximation holds for large n and moderate p. Since a Normal is continuous and Binomial is discrete we

have to use a continuity correction to discretize the Normal.

$$P(X=k) \sim P\left(k-\frac{1}{2} < Y < k+\frac{1}{2}\right) = \Phi\left(\frac{k-np+0.5}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-np-0.5}{\sqrt{np(1-p)}}\right)$$

You should get comfortable deciding what continuity correction to use. Here are a few examples of discrete probability questions and the continuity correction:

Discrete (Binomial) probability question	Equivalent continuous probability question
$P(X=6)$	$P(0.5 < X < 6.5)$
$P(X \geq 6)$	$P(X > 5.5)$
$P(X > 6)$	$P(X > 6.5)$
$P(X < 6)$	$P(X < 5.5)$
$P(X \leq 6)$	$P(X < 6.5)$

Example 3

100 visitors to your website are given a new design. Let X = # of people who were given the new design and spend more time on your website. Your CEO will endorse the new design if $X \geq 65$. What is $P(\text{CEO endorses change}|\text{it has no effect})$?

$E[X] = np = 50$. $Var(X) = np(1-p) = 25$. $\sigma = \sqrt{(Var(X))} = 5$. We can thus use a Normal approximation: $Y \sim \mathcal{N}(50, 25)$.

$$P(X \geq 65) \approx P(Y > 64.5) = P\left(\frac{Y-50}{5} > \frac{64.5-50}{5}\right) = 1 - \Phi(2.9) = 0.0019$$

Example 4

Stanford accepts 2480 students and each student has a 68% chance of attending. Let X = # students who will attend. $X \sim Bin(2480, 0.68)$. What is $P(X > 1745)$?

$E[X] = np = 1686.4$. $Var(X) = np(1-p) = 539.7$. $\sigma = \sqrt{(Var(X))} = 23.23$. We can thus use a Normal approximation: $Y \sim \mathcal{N}(1686.4, 539.7)$.

$$P(X > 1745) \approx P(Y > 1745.5) = P\left(\frac{Y-1686.4}{23.23} > \frac{1745.5-1686.4}{23.23}\right) = 1 - \Phi(2.54) = 0.0055$$

7 Exponential Random Variable

An Exponential Random Variable $X \sim Exp(\lambda)$ represents the time until an event occurs. It is parametrized by $\lambda > 0$, the rate at which the event occurs. This is the same λ as in the Poisson distribution.

Properties

The Probability Density Function (PDF) for an Exponential is:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

The expectation is $E[X] = \frac{1}{\lambda}$ and the variance is $Var(X) = \frac{1}{\lambda^2}$

There is a closed form for the Cumulative distribution function (CDF):

$$F(x) = 1 - e^{-\lambda x} \text{ where } x \geq 0$$

Example 1

Let X be a random variable that represents the number of minutes until a visitor leaves your website. You have calculated that on average a visitor leaves your site after 5 minutes and you decide that an Exponential function is appropriate to model how long until a person leaves your site. What is the $P(X > 10)$?

We can compute $\lambda = \frac{1}{5}$ either using the definition of $E[X]$ or by thinking of how much of a person leaves every minute (one fifth of a person). Thus $X \sim Exp(1/5)$.

$$\begin{aligned} P(X > 10) &= 1 - F(10) \\ &= 1 - (1 - e^{-\lambda 10}) \\ &= e^{-2} \approx 0.1353 \end{aligned}$$

Example 2

Let X be the # hours of use until your laptop dies. On average laptops die after 5000 hours of use. If you use your laptop for 7300 hours during your undergraduate career (assuming usage = 5 hours/day and four years of university), what is the probability that your laptop lasts all four years?

We can compute $\lambda = \frac{1}{5000}$ either using the definition of $E[X]$. Thus $X \sim Exp(1/5000)$.

$$\begin{aligned} P(X > 7300) &= 1 - F(7300) \\ &= 1 - (1 - e^{-7300/5000}) \\ &= e^{-1.46} \approx 0.2322 \end{aligned}$$