Lecture 17: Continuous Random Variables

Anup Rao

May 10, 2019

We discuss continuous random variables.

So far, we have been considering random random variables that only take on discrete values. However, many real-world processes involve random variables that take *continuous* values. For example, in Buffon's needle experiment, we counted the number of times that the needle hit a line, which is a discrete random variable. However, we might have wanted to model the angle of the needle after it falls. That would be a continuous random variable. Another example has to do with the Poisson process—we could be interested in the time that the first request comes in, rather than just counting the number of server requests.

There is a basic issue with defining continuous random variables—our old definition of what counts as a distribution no longer makes sense. Recall that we said that a distribution $p(x)$ must satisfy:

- For all x, $p(x) \geq 0$.
- $\sum_x p(x) = 1$.

If wanted to define a random variable that takes on a uniformly random value between 0 and 1, what should $p(x)$ be for $0 \leq x \leq 1$? Since the distribution is uniform, $p(x) = p(y)$ for all x, y in the interval. But then if we set $p(x) = \epsilon > 0$, we have

$$\sum_{0 \leq x \leq 1} p(x) = \sum_{0 \leq x \leq 1} \epsilon > 1.$$

So, we cannot set $p(x)$ to be positive to satisfy the definition.

Continuous random variables require a new definition for what a distribution is. Basically, we need to change sums to integrals.

In fact, discrete random variables can also be modeled using integrals, though we do not spend too much time doing this in this course.

There are two ways to define a distribution on real numbers. The first is using its *probability density function*, or pdf. The pdf is a function $f : \mathbb{R} \to \mathbb{R}$ with

- For all x, $f(x) \geq 0$.
- $\int\limits_{-\infty}^{\infty} f(x)dx = 1$.

Given this description of the distribution, the probability that the random variable takes a value in a set S is just

$$p(X \in S) = \int\limits_S f(x)dx.$$

If the pdf of the distribution is $f(x)$, intuitively this means that the probability that the random variable will take a value in an interval of length ϵ around x converges to $\epsilon \cdot f(x)$, as $\epsilon \to 0$.

Another way to describe the same distribution is using the *cumulative distribution function* or cdf. The cdf of a variable is a function $g : \mathbb{R} \to \mathbb{R}$ with $g(a) = p(X \leq a)$. If the variable has pdf f, we must have

$$g(a) = p(X \leq a) = \int_{-\infty}^{a} f(x)dx.$$

This means that the pdf is the derivate of the cdf.

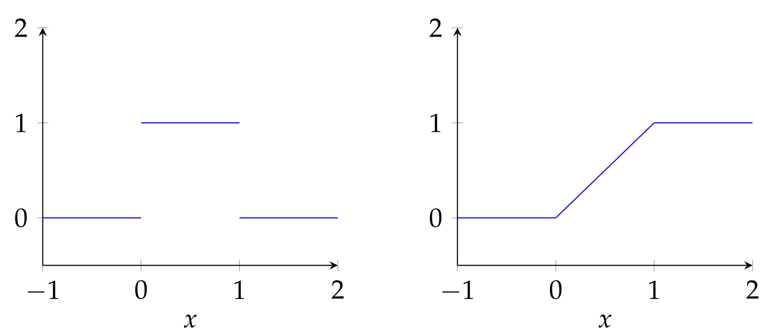
Example: The uniform distribution

The uniform distribution on numbers between 0 and 1 has the pdf

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The cdf is given by

$$g(a) = \int_{-\infty}^{a} f(x)dx = \begin{cases} 0 & \text{if } a \leq 0, \\ a & \text{if } 0 \leq a \leq 1, \\ 1 & \text{if } 1 \leq a. \end{cases}$$



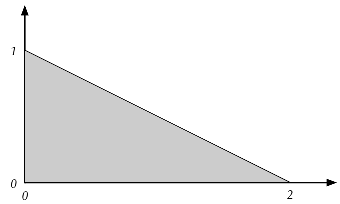
(a) pdf

(b) cdf

Figure 1: The pdf and cdf of a uniformly random number in between 0 and 1.

There are a few important things to note here. First of all, the pdf of a continuous variable can actually take on values *larger* than 1. For example, the pdf of variable that is a uniformly random number in between 0 and 1/2 is the function that is 2 in this interval, and 0 everywhere else. However, the cdf actually computes a probability, so it is always a number in between 0 and 1.

Random point from a triangle



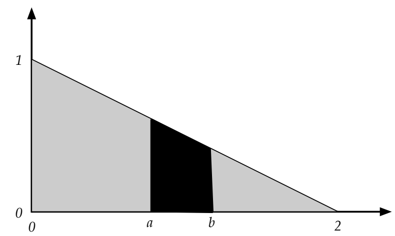


Figure 3: The probability that X lies in the black region is the area of the black region divided by the area of the whole triangle.

Suppose we sample a uniformly random point in the interior of the triangle shown in Figure 2. Let X be the x-coordinate of the resulting point. What are the pdf and the cdf of this distribution?

In this case, the pdf is particularly easy to guess. We want a function $f(x)$ so that for any $a < b$, $p(a \le X \le b) = \int_a^b f(x)dx$. Now, when a, b are in between 0 and 2, $p(a \le X \le b)$ is simply the area of the black region shown in Figure 3, divided by the area of the whole triangle. The area of the whole triangle in this example is 1. So, we want a function where the area under the curve should be equal to exactly the area of the black region. This is very easy... the function should look exactly like the triangle!

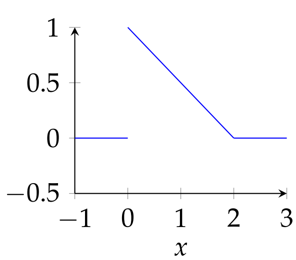
It turns out that this means we should set $f(x) = 1 - x/2$. To compute the cdf, we need a function $g(x)$ so that

$$g(a) = \int_{-\infty}^{a} f(x)dx.$$

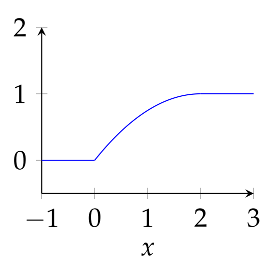
Computing this gives $g(a) = 0$ for $a < 0$,

$$g(a) = x - x^2/4\Big|_0^a = a - a^2/4$$

for $0 \le a \le 1$, and $g(a) = 1$ for $a > 1$.



(a) pdf

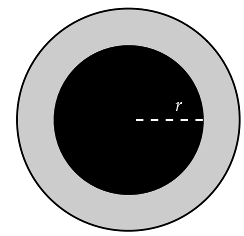


(b) cdf

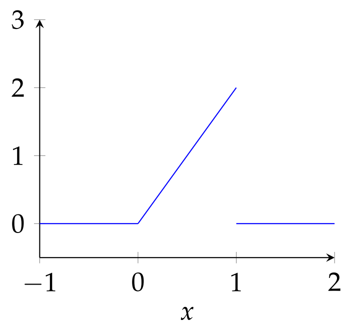
Figure 4: The pdf and cdf of the x-coordinate of a point in the triangle

Example: Throwing a dart

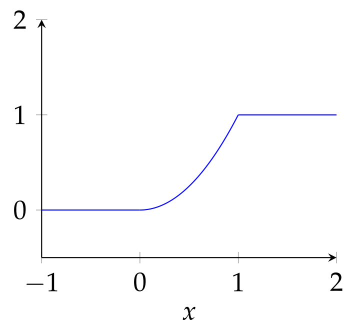
Suppose we throw a dart at a dartboard of radius 1, and the dart lands on a uniformly random point in the circle. What is the pdf and



cdf of the distance of the dart from the center? Let R be the distance of the dart from the center of the circle. This time, let us compute the cdf first. The cdf $g(r)$ has to satisfy that $g(r) = p(R \leq r)$. When $r < 0$, this probability is certainly 0. When $r \geq 1$, this probability is 1. When r is in between $0, 1$, the probability is the area of the black region shown on the right, divided by the area of the whole circle. So, we have $g(r) = 2\pi r^2/(2\pi) = r^2$. The pdf of R is then the derivative of this function $f(r) = 2r$.

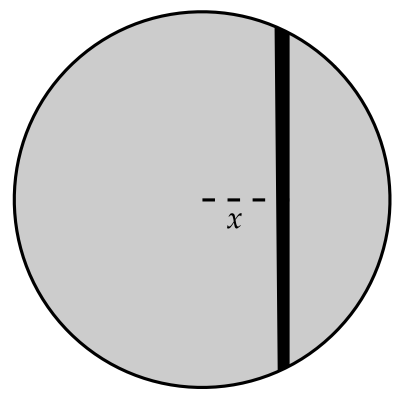


(a) pdf



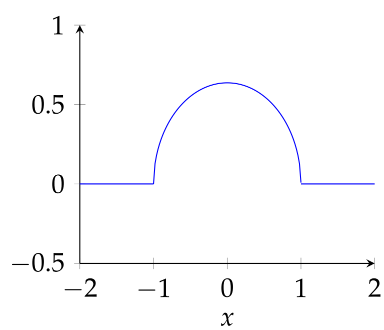
(b) cdf

Figure 5: The pdf and cdf of the distance of the point from $0, 0$.

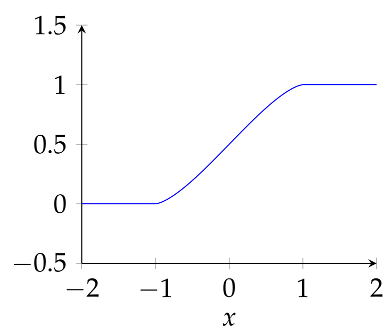


What about the the distribution of the x-coordinate of the dart? To calculate the pdf of this distribution, we need to figure out the area of the black region shown on the right.

If the strip is of width ϵ, its height is $2\sqrt{1-x^2}$, and its area is $2\sqrt{1-x^2}$. So, the probability of landing in the strip is about $\epsilon \cdot 2\sqrt{1-x^2}/\pi$, since the area of the whole circle is π. Thus the pdf should be $f(x) = 2\sqrt{1-x^2}/\pi$ in the interesting region. The cdf is the integral of this function, which turns out to be $(1/\pi)(\arcsin(x) + x\sqrt{1-x^2} + \pi/2)$.



(a) pdf



(b) cdf

Figure 6: The pdf and cdf of the distance of the x-coordinate of the dart.