4.1 Probability density function

Unit 4: Continuous Random Variables

Adapted from Blitzstein-Hwang Chapter 5.

So far we have been working with *discrete* random variables, whose possible values can be written down as a list. In this unit we will introduce *continuous* r.v.s, which can take on any real value in an interval (possibly of infinite length, such as $(0, \infty)$ or the entire real line). First we'll look at properties of continuous r.v.s in general. Then we'll introduce three famous continuous distributions---the Uniform, Normal, and Exponential---which, in addition to having important stories in their own right, serve as building blocks for many other useful continuous distributions.

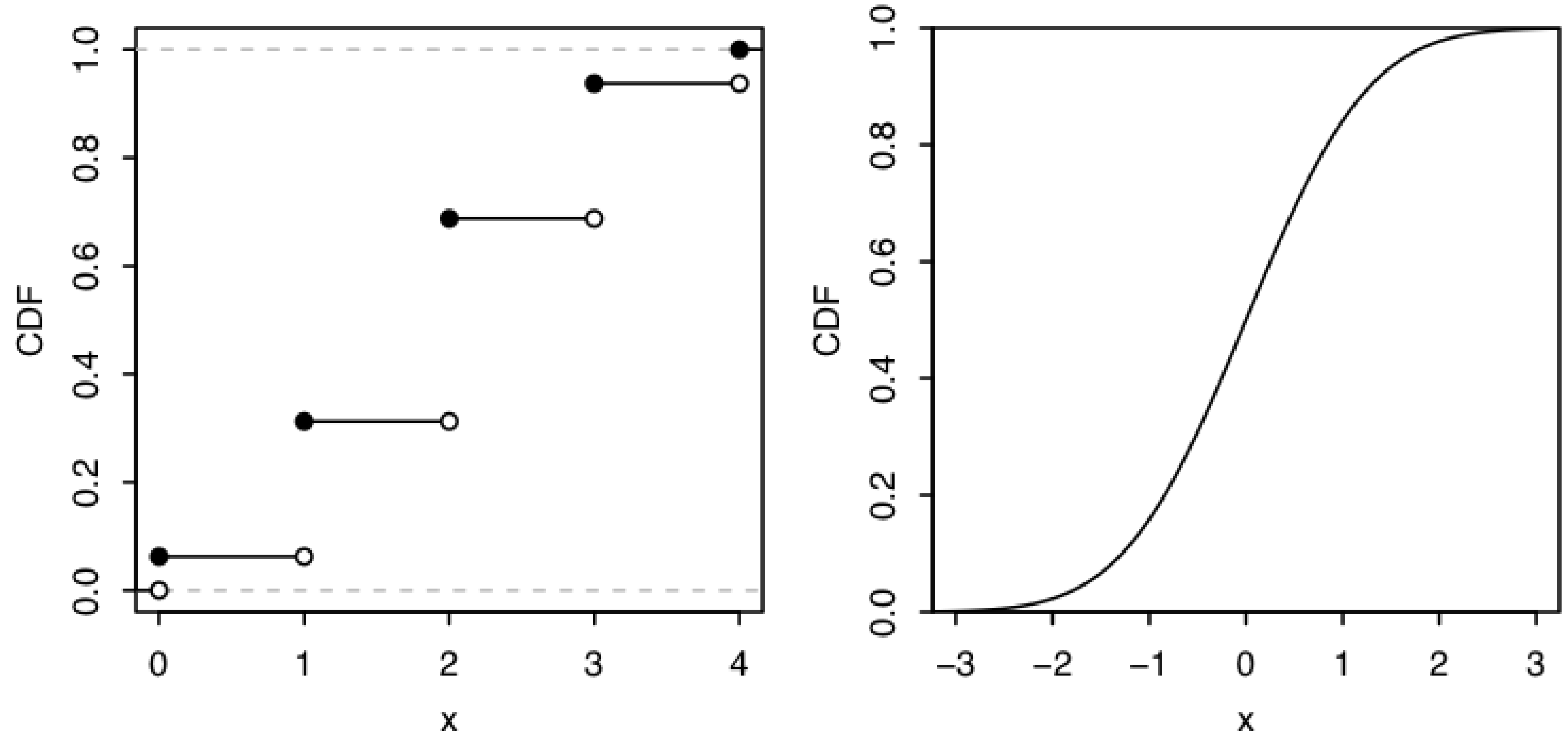

Figure 4.1.1: Discrete vs. continuous r.v.s. Left: The CDF of a discrete r.v. has jumps at each point in the support. Right: The CDF of a continuous r.v. increases smoothly.

View Larger Image
Image Description

Recall that for a discrete r.v., the CDF jumps at every point in the support, and is flat everywhere else. In contrast, for a continuous r.v. the CDF increases smoothly; see Figure 4.1.1 for a comparison of discrete vs. continuous CDFs.

DEFINITION 4.1.2 (CONTINUOUS R.V.).

An r.v. has a *continuous distribution* if its CDF is differentiable. We also allow there to be endpoints (or finitely many points) where the CDF is continuous but not differentiable, as long as the CDF is differentiable everywhere else. A *continuous random variable* is a random variable with a continuous distribution.

DEFINITION 4.1.3 (PROBABILITY DENSITY FUNCTION).

For a continuous r.v. X with CDF F, the *probability density function* (PDF) of X is the derivative f of the CDF, given by $f(x) = F'(x)$. The *support* of X, and of its distribution, is the set of all x where $f(x) > 0$.

An important way in which continuous r.v.s differ from discrete r.v.s is that for a continuous r.v. X, $P(X = x) = 0$ for all x. This is because $P(X = x)$ is the height of a jump in the CDF at x, but the CDF of X has no jumps! Since the PMF of a continuous r.v. would just be 0 everywhere, we work with a PDF instead.

The PDF is analogous to the PMF in many ways, but there is a key difference: for a PDF f, the quantity $f(x)$ is *not* a probability, and in fact it is possible to have $f(x) > 1$ for some values of x. To obtain a probability, we need to *integrate* the PDF. The fundamental theorem of calculus tells us how to get from the PDF back to the CDF.

Proposition 4.1.4 (PDF to CDF).

Let X be a continuous r.v. with PDF f. Then the CDF of X is given by

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Proof

By the definition of PDF, F is an antiderivative of f. So by the fundamental theorem of calculus,

$$\int_{-\infty}^{x} f(t)dt = F(x) - F(-\infty) = F(x).$$

The above result is analogous to how we obtained the value of a discrete CDF at x by summing the PMF over all values less than or equal to x; here we *integrate* the PDF over all values up to x, so the CDF is the *accumulated area* under the PDF. Since we can freely convert between the PDF and the CDF using the inverse operations of integration and differentiation, both the PDF and CDF carry complete information about the distribution of a continuous r.v. Since the PDF determines the distribution, we should be able to use it to find the probability of X falling into an interval (a, b). A handy fact is that we can include or exclude the endpoints as we wish without altering the probability, since the endpoints have probability 0: $P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b)$.

☣ **Warning 4.1.5 (Including or excluding endpoints).**

We can be carefree about including or excluding endpoints as above for continuous r.v.s, but we must not be careless about this for discrete r.v.s.

By the definition of CDF and the fundamental theorem of calculus,

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx.$$

Therefore, to find the probability of X falling in the interval $(a, b]$ (or (a, b), $[a, b)$, or $[a, b]$) using the PDF, we simply integrate the PDF from a to b. Just as a valid PMF must be nonnegative and sum to 1, a valid PDF must be nonnegative and integrate to 1.

Theorem 4.1.6 (Valid PDFs).

The PDF f of a continuous r.v. must satisfy the following two criteria:

- Nonnegative: $f(x) \geq 0$;
- Integrates to 1: $\int_{-\infty}^{\infty} f(x)dx = 1$.

Proof

The first criterion is true because probability is nonnegative; if $f(x_0)$ were negative, then we could integrate over a tiny region around x_0 and get a negative probability. Alternatively, note that the PDF at x_0 is the slope of the CDF at x_0, so $f(x_0) < 0$ would imply that the CDF is *decreasing* at x_0, which is not allowed. The second criterion is true since $\int_{-\infty}^{\infty} f(x)dx$ is the probability of X falling somewhere on the real line, which is 1.

For practice, let's now look at a specific example of a PDF.

Example 4.1.7 (Logistic).

The Logistic distribution has CDF

$$F(x) = \frac{e^x}{1 + e^x}, \quad x \in \mathbb{R}.$$

To get the PDF, we differentiate the CDF, which gives

$$f(x) = \frac{e^x}{(1+e^x)^2}, \quad x \in \mathbb{R}.$$

Let $X \sim \textbf{Logistic}$. To find $P(-2 < X < 2)$, we need to integrate the PDF from -2 to 2.

$$P(-2 < X < 2) = \int_{-2}^{2} \frac{e^x}{(1+e^x)^2} dx = F(2) - F(-2) \approx 0.76.$$

The integral was easy to evaluate since we already knew that F was an antiderivative for f, and we had a nice expression for F. Otherwise, we could have made the substitution $u = 1 + e^x$, so $du = e^x dx$, giving

$$\int_{-2}^{2} \frac{e^x}{(1+e^x)^2} dx = \int_{1+e^{-2}}^{1+e^2} \frac{1}{u^2} du = \left(-\frac{1}{u}\right)\Big|_{1+e^{-2}}^{1+e^2} \approx 0.76.$$

Figure 4.1.8 shows the Logistic PDF (left) and CDF (right). On the PDF, $P(-2 < X < 2)$ is represented by the shaded area; on the CDF, it is represented by the height of the curly brace. You can check that the properties of a valid PDF and CDF are satisfied.

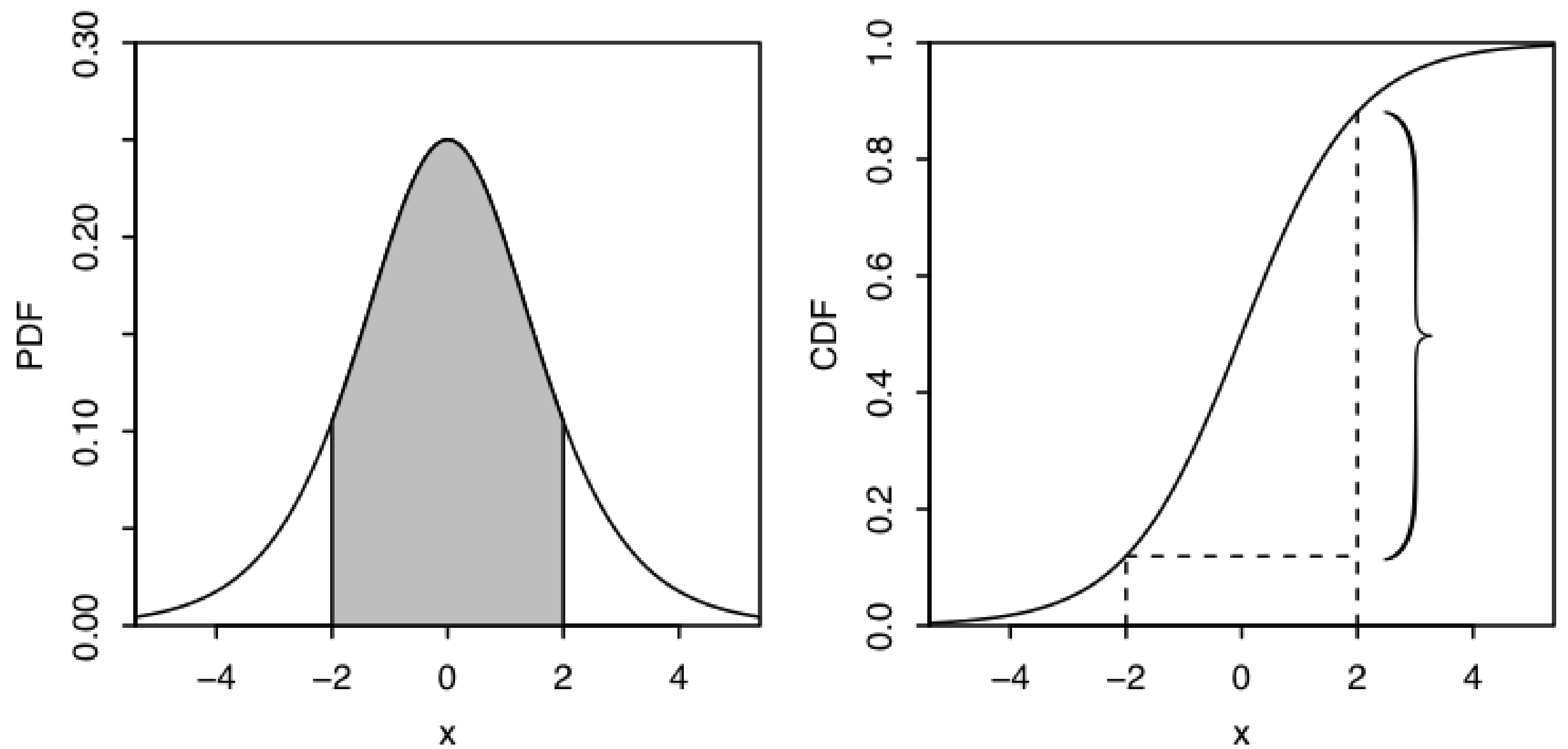

Figure 4.1.8: Logistic PDF and CDF. The probability $P(-2 < X < 2)$ is indicated by the shaded area under the PDF and the height of the curly brace on the CDF.

View Larger Image
Image Description

Although the height of a PDF at x does not represent a probability, it is closely related to the probability of falling into a tiny interval around x, as the following intuition explains.

Intuition 4.1.9.

Let F be the CDF and f be the PDF of a continuous r.v. X. As mentioned earlier, $f(x)$ is *not* a probability; for example, we could have $f(3) > 1$, and we know $P(X = 3) = 0$. But thinking about the probability of X being *very close* to 3 gives us a way to interpret $f(3)$. Specifically, the probability of X being in a tiny interval of length ϵ, centered at 3, will essentially be $f(3)\epsilon$. This is because

$$P(3 - \epsilon/2 < X < 3 + \epsilon/2) = \int_{3-\epsilon/2}^{3+\epsilon/2} f(x)dx \approx f(3)\epsilon,$$

if the interval is so tiny that f is approximately the constant $f(3)$ on that interval. In general, we can think of $f(x)dx$ as the probability of X being in an infinitesimally small interval containing x, of length dx.

In practice, X often has *units* in some system of measurement, such as units of distance, time, area, or mass. Thinking about the units is not only important in applied problems, but also it often helps in checking that answers make sense. Suppose for concreteness that X is a length, measured in centimeters (cm). Then $f(x) = dF(x)/dx$ is the probability per cm at x, which explains why $f(x)$ is a probability *density*. Probability is a dimensionless quantity (a number without physical units), so the units of $f(x)$ are cm^{-1}. Therefore, to be able to get a probability again, we need to multiply $f(x)$ by a length. When we do an integral such as $\int_0^5 f(x)dx$, this is achieved by the often-forgotten dx.