📚 Math Review

DA Learning & Development

Ravi Dayabhai

January 20, 2021

This is a brief primer (or refresher, depending on your background) of the core concepts from calculus that will help ease the transition from "Pebble World" (and the discrete distributions that we've encountered in our tour of the Distribution Zoo so far) to the next exhibit: the *continuous distributions*. You've likely encountered a few of these in your mathematical career already (e.g., the Normal). Adding these distributions to our toolkits is essential to make sense of bedrock ideas in statistics (e.g., the Central Limit Theorem), but before we get ahead of ourselves, let's make sure we feel comfortable with the basic underpinnings that enable us to reckon with these new and exotic distributions.

1 Functions

We've already touched on sets (see Lesson 2 material in our repo), so we'll start by taking a closer look at **functions**.

Definition 1.1 (Function)**.** A function is a relation that associates each element x of a set X, called the *domain* of the function, to a single element y of another set Y (possibly the same set), the *codomain* of the function. We can write this as

$$f : X \mapsto Y.$$

Intuitively, a function is a deterministic rule that "maps" an element from one set to an element in another set. Another interpretation is that functions are "machines," taking inputs and producing outputs. (It's easy to see why these objects are central to computer science as well.) Importantly, different x's can map to the same y, but each x only maps to one y (e.g., the "vertical line test" from grade school).

Remark. We should take care to distinguish f from $f(x)$. The former is the function itself, the rule; the latter is a number for each number x. As in programming, a function (treated as an object) is not the same thing as what the function returns when called with a particular set of arguments.

Certain table relations can be thought of functions: "one-to-one" or "many-to-one" are examples of this. A "one-to-many" table relation violates the definition of a function, even if its *inverse* is a valid function. In this way, we don't need to restrict ourselves to numbers. In fact, we've already seen functions in the context of probability when we defined random variables[1].

[1] **random variable** ≔ a function that maps the sample space to real numbers or vectors

1.1 Injective, Bijective, and Surjective Functions

Let $f\colon A \mapsto B$. We can describe or characterize this function by the relationship between its domain ($\forall a \in A$) and codomain ($\forall b \in B$).

Definition 1.2 (Injective Function)**.** A function is said to be injective or "one-to-one" if $f(a_1) \neq f(a_2)$ whenever $a_1 \neq a_2$. Any two distinct inputs to the function get mapped to two distinct outputs. Said another way: for every b there can be at most one a that maps to it.

Definition 1.3 (Surjective Function)**.** A function is said to be surjective or "onto" if $\forall b \in B, \exists a \in A, f(a) = b$. Said another way: every b has a corresponding a, but the converse may not necessarily be true.

Definition 1.4 (Bijective Function)**.** A function is said to be bijective or "one-to-one correspondence" if it is both injective and surjective.

A key distinction to make is between "one-to-one" and "one-to-one correspondence": the latter describes a function as being both injective and surjective, whereas the former describes it as being injective only. The terminology may seem clunky at first, but with a little practice, it makes talking about functions much easier.

1.2 Increasing and Decreasing Functions

Using the same function f from above, we can fashion two straightforward definitions for increasing and decreasing functions:

Definition 1.5 (Increasing Function)**.** If $a_1 \leq a_2 \implies f(a_1) \leq f(a_2)$, then f is said to be *increasing* over $[a_1, a_2]$.

Definition 1.6 (Decreasing Function)**.** If $a_1 \leq a_2 \implies f(a_1) \geq f(a_2)$, then f is said to be *decreasing* over $[a_1, a_2]$.

Note that these definitions allow for flat regions. *Strictly* increasing or decreasing restricts these definitions a bit more (read: replace the $\leq, \geq$ with $<, >$, respectively in the definitions above). Another word to describe increasing or decreasing functions is *monotone* (e.g., letting $f(x) = x^3$, f is *monotonically* increasing).

Remark. Any strictly monotone function is one-to-one (i.e., injective).

1.3 Even and Odd Functions

Let $f\colon \mathbb{R} \mapsto \mathbb{R}$. This simply means the function f maps *real* numbers to *real* numbers in one dimension. (Without digressing too much, a function $g\colon \mathbb{R}^2 \mapsto \mathbb{R}^3$ means g is a *vector*-valued function from two dimensions to three.)

Definition 1.7 (Even Function)**.** If, $\forall x$ in the domain of f, $f(x) = f(-x)$, then f is an *even* function.

Definition 1.8 (Odd Function)**.** If, $\forall x$ in the domain of f, $-f(x) = f(-x)$, then f is an *odd* function.

Remark. A function that exhibits neither property is neither even nor odd.

Even and odd functions exhibit symmetries: even functions can be reflected across the y-axis (in $\mathbb{R}^2$) and odd functions have rotational symmetry around the origin. We can leverage these properties when we need to take integrals. Even functions have the property that, for any a,

$$\int_{-a}^{a} f(x)\,dx = 2\int_{0}^{a} f(x)\,dx,$$

and odd functions have the property that, for any a,

$$\int_{-a}^{a} f(x)\,dx = 0.$$

(**Hint:** Draw a picture to see why this is true.)

2 Calculus

In "Pebble World", we dealt with *countably* finite (or infinite) supports[2], which made it easy to relate pebbles to numbers. Now, we're leaving the comfort of counting with integers behind and venturing into the *uncountable*.

2.1 Real Numbers

Earlier, we talked about **real** numbers, but let's remind ourselves of what they are.

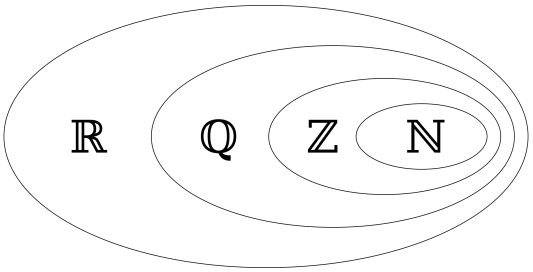

Figure 1: $\mathbb{Q}$ are the *rationals*; $\mathbb{R}$ are the *reals*

Definition 2.1 (Rational Numbers)**.** A rational number is a number such as $\frac{-3}{7}$ that can be expressed as the quotient or fraction $\frac{p}{q}$ of two integers, a numerator p and a non-zero denominator q.

Definition 2.2 (Irrational Numbers)**.** An irrational number is a number that cannot be expressed as the quotient or fraction $\frac{p}{q}$ of two integers, a numerator p and a non-zero denominator q. Examples include π, e, $\sqrt[3]{2}$, $\sqrt{2}$, and a bunch of others, to say the least (in fact, probability helps to intuit that there are "more" irrationals than rationals).

[2] **support** := all values a random variable can take; if X is a discrete random variable, then the finite or countably infinite set of values x such that $P(X = x) > 0$ is called the support of X

Definition 2.3 (Real Numbers)**.** The set of real numbers is the union of the rational and irrational numbers. Real numbers can be thought of as points on an infinitely long line (called the "number line" or "real line"), where the points corresponding to integers are equally-spaced.

There are infinitely many integers, and they are *countable*. There are infinitely many real numbers, but they are *uncountable*. In fact, the "smoothness" of a continuous function over a given interval arises because the function is defined for all, infinite reals in that interval (read: no "gaps" anywhere). The main idea to transplant to your mental model for probability should be that we no longer can rely on counting or "weighing" individual "pebbles"!

2.2 Limits

Let's now return to what it means to "take the limit" of a function at a certain point, a, and how this notion allows us to make the logical leaps necessary to work on a continuum in probability.

Definition 2.4 (Limits)**.** The limit of $f(x)$ as x approaches a is a number L, written

$$\lim_{x \to a} f(x) = L$$

if the value of $f(x)$ is as close as one wishes to L for all x sufficiently close, but not equal to a.

The best case scenario is when $f(a) = L$ (this means f is *continuous* at a when combined with Definition 2.4), but this need not be necessarily true (e.g., removable discontinuities) for the limit to exist. Luckily, for our purposes, we won't worry about discontinuities (in general) because the distributions we will be studying are called *continuous* distributions for a reason!

The notion of a limit is, however, important when we think about "adding up" (think: integrating) under a *probability density function* (PDF) whose domain is over $(-\infty, \infty)$ – the area under a PDF must be 1 in the same way the total mass of pebbles in "Pebble World" sums to 1. This is discussed more in Section 2.3 below.

2.3 Differentiation & Integration

We won't burden ourselves with a rigorous treatment of these topics, but we should call out the big ideas. First, a useful theorem to segue us from Section 2.2:

Theorem 2.1. If a function $f(x)$ is differentiable in some interval, then it is also continuous in that interval. The converse may not be true.

If you can take the derivative of a function over some interval then you know the function is continuous over that interval. Later, we'll see this means we're permitted to differentiate a continuous random variable's CDF to yield its PDF.

Remember, we can fully describe a discrete distribution with either the CDF or the PMF (or the parameters and class of distribution, e.g., $\text{Bin}(n, p)$). Discrete CDFs have jumps, and these jumps correspond to values of the PMF, as seen in Fig. 2. We could also start with the PMF and construct the CDF. This sensation (of going back and forth between

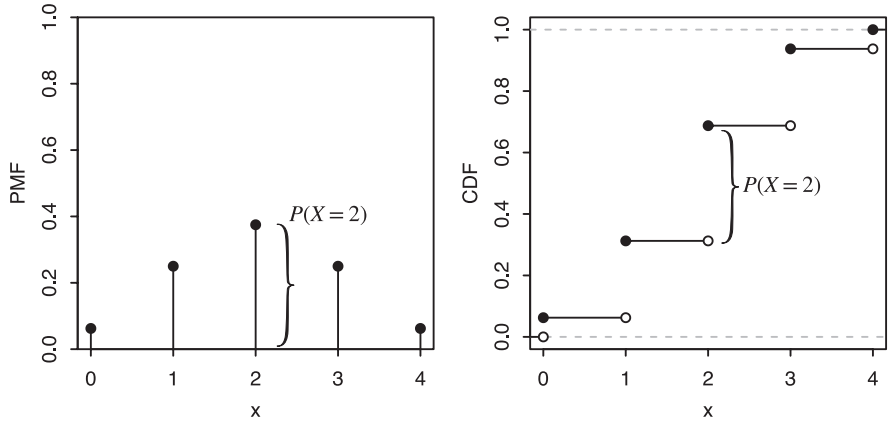

Figure 2: Even in the discrete setting, the PMF "feels" like the derivative of the CDF.

two related functions) is the most fundamental idea of calculus – so fundamental, in fact, that it's called the **Fundamental Theorem of Calculus** (FTC). The two variants of the FTC demonstrate how F and f characterize each other.

Theorem 2.2 (Fundamental Theorem of Calculus: Part 1 (FTC 1))**.** If f is continuous and $F'(x) = f(x)$, then

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

Theorem 2.3 (Fundamental Theorem of Calculus: Part 2 (FTC 2))**.** If f is continuous and $G(x) = \int_a^x f(t)\,dt$, then $G'(x) = f(x)$. Note that $G(x) = F(x) + C$, where C is a constant. The exact value for C depends on the choice of a. This can be confirmed by Theorem 2.2.

Stare at this for a while if some time has passed since you last acquainted yourself with calculus – it's not that bad, I promise! In short, these two theorems establish the following, key insight:

> **Integration and differentiation are inverse operations**, in the same way addition and subtraction are inverse operations of each other.
>
> $$x^2 \xrightarrow{\int} \frac{x^3}{3}\,(+C)\,\xrightarrow{\frac{d}{dx}} x^2$$

Another way to reason about integration ($\int$) is that it is the continuous analog of summation ($\sum$). We can think of the [signed] area under the graph of $f(x)$ as approximating the sum of all of the rectangles formed by height $f(x)$ and width Δx, which is just some small, non-zero length that evenly divides (a, b). This approximation gets better as we let $\Delta x \to 0$ because this means we're now generating more (albeit thinner) rectangles, per Fig. 3. (Grant Sanderson does a much better job of making this idea come to life than I can here.)

Remark (Notation). Leibniz's notation (e.g., $\frac{dy}{dx}$) is a great reminder of what's going on: $f(x)$, the derivative of $F(x)$, tells us how much a "little nudge" to x (Δx) affects $F(x)$.

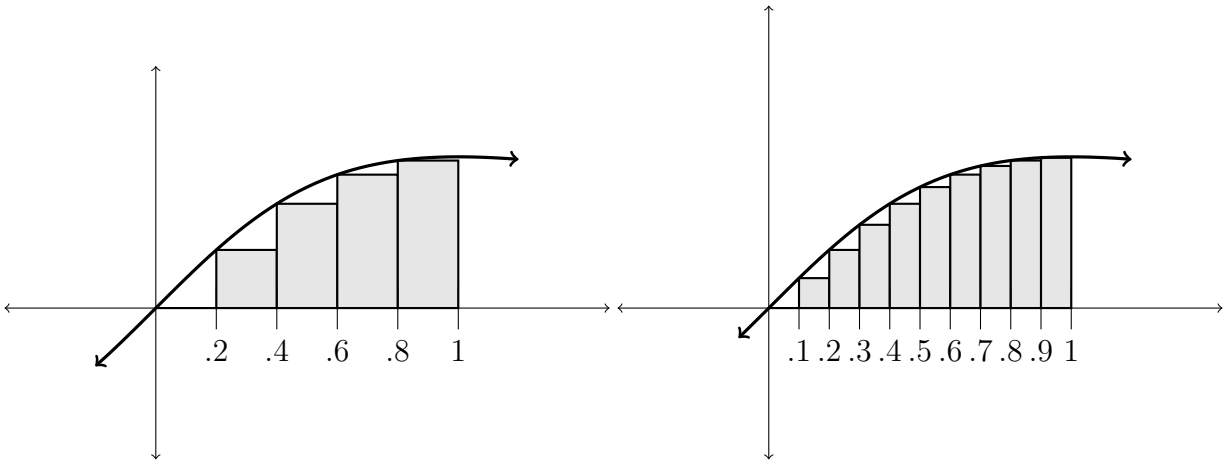

Figure 3: Smaller choices of $\Delta x \implies$ more rectangles $\implies$ better approximation of the area under the curve

Making this nudge smaller and smaller (read: $\Delta x \to 0$ over all x's) is precisely the idea captured by the notation – the *limit* as the size of the nudges approaches 0 is dx and the effect on $F(x)$ of these nudges (in the limit) is what we call $dF(x)$ (or dy when $y = F(x)$). The ratio of these two quantities in the limit (i.e., $\frac{d}{dx}F(x)$) is what the derivative is!

But what does summing up the areas of ever-thinner rectangles (i.e., "areas under curves") have to do with the differences of two quantities, namely, $F(b) - F(a)$? (See Theorem 2.2.) It's pretty remarkable that we can describe a certain area under $f(x)$ just by looking at two points on the graph of F!

First, consider what information is encoded by f: per the definitions above,

$$F'(x) = \frac{d}{dx}F(x) = f(x).$$

So, if integrating f from a to b is just doing some sort of "fancy sum", and f describes vanishingly tiny nudges to F (due to vanishingly tiny nudges to x, i.e., dx), then it stands to reason that integrating "brings together" all of the accumulated nudges of F from a to b . . . but this is precisely what $F(b) - F(a)$ describes!

We can apply this logic to the two primary functions that govern continuous random variables, namely, the CDF $F_X(x)$ and PDF $f_X(x)$ (for a continuous random variable X). Again, the notation suggests what our intuition leads us to believe: integrating the PDF yields the CDF and differentiating the CDF yields the PDF. We'll have more to say about this when we visit these distributions in the Distribution Zoo, but Fig. 4 below illustrates this notion.

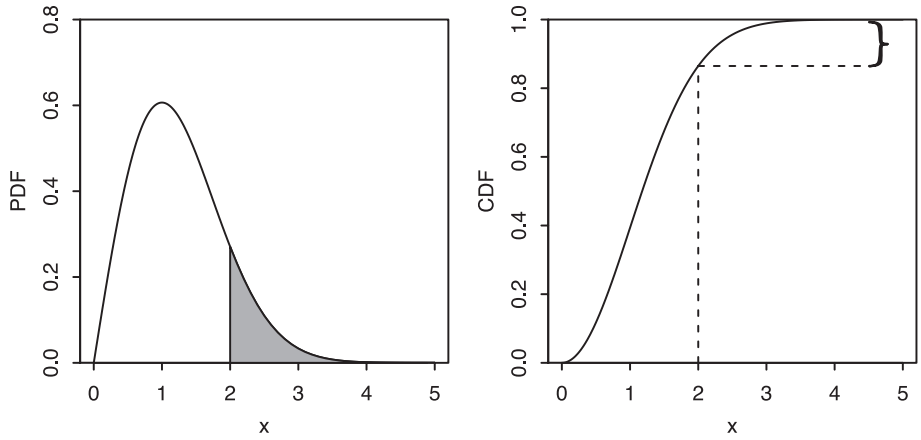

Figure 4: The area of the shaded region (left) under the PDF is the same thing as the difference between the CDF for the corresponding values of x (right).

2.4 Taylor Series

Taylor series help us to approximate a function about a point using polynomials.

Definition 2.5 (Taylor Series)**.** Let $f(x)$ be a real-valued function that is infinitely differentiable at $x = x_0$. The Taylor series expansion for the function $f(x)$ centered around the point $x = x_0$ is given by

$$\sum_{n=0}^{\infty} f^{(n)}(x_0)\frac{(x - x_0)^n}{n!}.$$

Note that $f^{(n)}(x_0)$ represents the n^{th} derivative of $f(x)$ at $x = x_0$. If we choose $x_0 = 0$, we get a special case of this approximation method, which is called the **Maclaurin series**.

This is an extremely powerful approximation technique that leverages the fact that polynomials are relatively easy to deal with (e.g., taking repeated derivatives means applying the power rule over and over again) and can be applied to functions that would be hard to directly compute. The reason we get better approximations as we add more terms is because the added terms help to further characterize how f changes in the neighborhood around x_0. Again, see this 3Blue1Brown video for the visual intuition behind this statement (also see Fig. 5).

For our purposes, being able to identify patterns in infinite series that suggest they are approximations for other, more well-known functions is sufficient. See Table 1 below for some of the more common examples (for $x_0 = 0$) that will show up in the upcoming lessons.

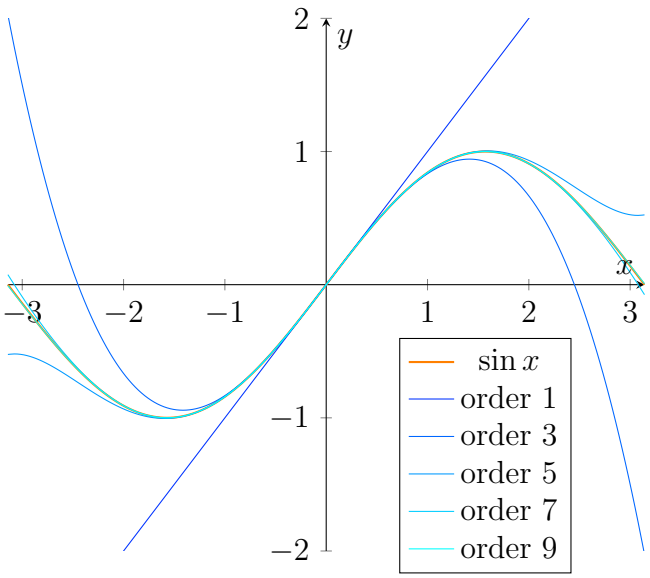

Figure 5: The plot of the original function is barely visible because we have a pretty good approximation using just polynomials.

Function	Maclaurin Series	Interval of Convergence
$\frac{1}{1-x}$	$\sum_{n=0}^{\infty} x^n$	$-1 < x < 1$
e^x	$\sum_{n=0}^{\infty} \frac{x^n}{n!}$	$-\infty < x < \infty$
$\ln(1+x)$	$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}x^n}{n}$	$-1 < x \leq 1$
$\sin(x)$	$\sum_{n=0}^{\infty} \frac{(-1)^{n+1}x^{2n+1}}{(2n+1)!}$	$-\infty < x < \infty$
$\cos(x)$	$\sum_{n=0}^{\infty} \frac{(-1)^{n}x^{2n}}{(2n)!}$	$-\infty < x < \infty$

Table 1: The interval of convergence governs where this approximation works.