Continuous random variables

Up to now we have focused exclusively on discrete random variables, which take on only a finite (or countably infinite) number of values.

But in real life many quantities that we wish to model probabilistically are continuous-valued; examples include the position of a particle in a box, the time at which a certain incident happens, or the direction of travel of a meteorite. In this lecture, we discuss how to extend the concepts we’ve seen for discrete random variables to continuous ones. As we shall see, everything translates in a natural way once we have set up the right framework. The framework involves some elementary calculus.

Uniform continuous random variables

Suppose we spin a “wheel of fortune” and record the position $X$ of the pointer on the outer circumference of the wheel. Assuming that the circumference is of length $\ell$ and that the wheel is unbiased, the position $X$ is presumably equally likely to take on any value in the real interval $[0, \ell]$. How do we model the distribution of $X$?

Consider for a moment the (almost) analogous discrete setting, where the pointer can stop only at a finite number $m$ of positions distributed evenly around the wheel. (If $m$ is very large, then presumably this is in some sense similar to the continuous setting.) Then we would model this situation using a discrete random variable $X$, with a pmf given by

$$P_X(\frac{k}{m} \ell) = P(X = \frac{k}{m} \ell) = \frac{1}{m}, \quad k = 0, 1, \ldots, m - 1,$$

and a sample space $\Omega = \{0, \frac{1}{m} \ell, \ldots, \frac{m-1}{m} \ell\}$.

In the continuous world, however, we get into trouble if we try the same approach. If we let $a$ range over all real numbers in $[0, \ell]$, what value should we assign to each $P(X = a)$? By uniformity this probability should be the same for all $x$, but then if we assign to it any positive value, the sum of all probabilities $P(X = a)$ will be $\infty$! Thus $P(X = a)$ must be zero for all $a \in [0, \ell]$. However, this carries no information about the distribution of $X$.

To rescue this situation, consider instead any non-empty interval $[a, b] \subseteq [0, \ell]$. Can we assign a non-zero probability to the event that $X$ lies in this interval? Since the probability that $X \in [0, \ell]$ must be 1, and since we want our probability to be uniform, the natural assignment of probability to the interval $[a, b]$ is

$$P(X \in [a, b]) = \frac{\text{length of } [a, b]}{\text{length of } [0, \ell]} = \frac{b - a}{\ell}. \quad (2)$$

In other words, the probability that $X$ lies in an interval is proportional to its length.

So we rescue the situation by assigning probabilities to events that $X$ lies in an interval instead of to events that $X$ takes on a specific value. But what about probabilities of other events? Actually, by specifying the
probabilities that $X$ lies in intervals we have also specified the probability of that $X$ lies in the disjoint union of (a finite or countably infinite number of) intervals, $I = \cup_i I_i$. For then we can write $P(X \in I) = \sum_i P(X \in I_i)$, in analogous fashion to the discrete case. Thus for example the probability that the pointer ends up in the first or third quadrants of the wheel is $\ell/4 + \ell/4 = 1/2$.

General continuous random variables

The continuous random variable we dealt with above was uniformly distributed over their ranges of values. What about random variables which are not uniformly distributed? How should we define their distributions?

The strategy is essentially the same as for uniform random variables: instead of specifying $P(X = a)$, we instead specify $P(a < X \leq b)$ for all intervals $[a, b]$.$^1$ To do this formally, we need to introduce the concept of a probability density function (sometimes referred to just as a “density”, or a “pdf”).

**Definition 10.1 (Density):** A probability density function for a random variable $X$ is a function $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$P(a < X \leq b) = \int_a^b f(x) \, dx \quad \text{for all } a \leq b. \quad (3)$$

Let’s examine this definition. Note that the definite integral is just the area under the curve $f$ between the values $a$ and $b$ (Figure 1(a)). Thus $f$ plays a similar role to the “histogram” we sometimes draw to picture the distribution of a discrete random variable.

In order for the definition to make sense, $f$ must obey certain properties. Some of these are technical in nature, which basically just ensure that the integral is always well defined; we shall not dwell on this issue here since all the densities that we will meet will be well behaved. What about some more basic properties of $f$? First, it must be the case that $f$ is a non-negative function; for if $f$ took on negative values we could find an interval in which the integral is negative, so we would have a negative probability for some event!

$^1$Note that it does not matter whether or not we include the endpoints $a, b$; since $P(X = a) = P(X = b) = 0$, we have $P(a < X < b) = P(a < X \leq b) = P(a < X < b)$. 

---

EE 178, Autumn 2017, Lecture 10
Second, since the r.v. \( X \) must take on some value everywhere in the space, we must have

\[
\int_{-\infty}^{\infty} f(x)dx = P(-\infty < X < \infty) = 1. \tag{4}
\]

In other words, the total area under the curve \( f \) must be 1.

A caveat is in order here. Following the “histogram” analogy above, it is tempting to think of \( f(x) \) as a “probability.” However, \( f(x) \) doesn’t itself correspond to the probability of anything! For one thing, there is no requirement that \( f(x) \) be bounded by 1 (and indeed, we shall see examples of densities in which \( f(x) \) is greater than 1 for some \( x \)). To connect \( f(x) \) with probabilities, we need to look at a very small interval \([x,x+\delta]\) close to \( x \). Assuming that the interval \([x,x+\delta]\) is so small that the function \( f \) doesn’t change much over that interval. we have

\[
P(x < X \leq x + \delta) = \int_{x}^{x+\delta} f(z)dz \approx \delta f(x). \tag{5}
\]

This approximation is illustrated in Figure 1(b). Equivalently,

\[
f(x) \approx \frac{P(x < X \leq x + \delta)}{\delta}. \tag{6}
\]

The approximation in (6) becomes more accurate as \( \delta \) becomes small. Hence, more formally, we can relate density and probability by taking limits:

\[
f(x) = \lim_{\delta \to 0} \frac{P(x < X \leq x + \delta)}{\delta}. \tag{7}
\]

Thus we can interpret \( f(x) \) as the “probability per unit length” in the vicinity of \( x \). Note that while the equation (3) allows us to compute probabilities given the probability density function, the equation (7) allows us to compute the probability density function given probabilities. Both relationships are useful in problems.

Now let’s go back and put our wheel-of-fortune r.v. \( X \) into this framework. What should be the density of \( X \)? Well, we want \( X \) to have non-zero probability only on the interval \([0,\ell]\), so we should certainly have \( f(x) = 0 \) for \( x < 0 \) and for \( x > \ell \). Within the interval \([0,\ell]\] we want the distribution of \( X \) to be uniform, which means we should take \( f(x) = c \) for \( 0 \leq x \leq \ell \). What should be the value of \( c \)? This is determined by the requirement (4) that the total area under \( f \) is 1. The area under the above curve is \( \int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\ell} cdx = c\ell \), so we must take \( c = \frac{1}{\ell} \). Summarizing, then, the density of the uniform distribution on \([0,\ell]\) is given by

\[
f(x) = \begin{cases} 
0 & \text{for } x < 0; \\
1/\ell & \text{for } 0 \leq x \leq \ell; \\
0 & \text{for } x > \ell.
\end{cases}
\]

This is plotted in Figure 2. Note that \( f(x) \) can certainly be greater than 1, depending on the value of \( \ell \).

**Exponential distribution**

The exponential distribution is a continuous version of the geometric distribution, which we have already met. Recall that the geometric distribution describes the number of tosses of a coin until the first Heads appears; the distribution has a single parameter \( p \), which is the bias (Heads probability) of the coin. Of
In real life applications, we are usually not waiting for a coin to come up Heads but rather waiting for a system to fail, a clock to ring, an experiment to succeed etc.

In such applications, we are frequently not dealing with discrete events or discrete time, but rather with continuous time: for example, if we are waiting for an apple to fall off a tree, it can do so at any time at all, not necessarily on the tick of a discrete clock. This situation is naturally modeled by the exponential distribution, defined as follows:

**Definition 10.2 (Exponential distribution):** For any $\lambda > 0$, a continuous random variable $X$ with pdf $f$ given by

$$f(x) = \begin{cases} Ae^{-\lambda x} & \text{if } x \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

is called an exponential random variable with parameter $\lambda$.

Since we know that the pdf has to integrate to one, we can evaluate $A$:

$$\int_0^\infty Ae^{-\lambda x} = 1 \implies A = \lambda.$$ 

The exponential distribution is frequently used to model interarrival times, e.g., time between two customers at a bank and lifetimes, e.g., lifetime of a bulb, lifetime of a radioactive particle etc.

To see one reason for the applicability of the exponential distribution in many cases, consider $P(X > x | X > x_1)$. For example, if $X$ is the duration of a call, and given that the call has been going on for $x_1$ minutes, what is the probability that it will go up to $x$ minutes.

$$P(X > x | X > x_1) = \frac{P(X > x, X > x_1)}{P(X > x_1)}$$

$$P(X > x | X > x_1) = \frac{P(X > x)}{P(X > x_1)}$$

$$P(X > x | X > x_1) = \frac{\int_x^\infty \lambda e^{-\lambda y}dy}{\int_{x_1}^\infty \lambda e^{-\lambda y}dy}$$

$$P(X > x | X > x_1) = \frac{e^{-\lambda x}}{e^{-\lambda x_1}}$$
\[ P(X > x | X > x_1) = e^{-\lambda(x-x_1)} \]
\[ P(X > x | X > x_1) = P(X > (x-x_1)) \]

Thus, the probability is the same as the probability of the call lasting at least \(x - x_1\) minutes. In other words, the fact that you were already on the phone for a certain amount of time has absolutely no relevance for the question of how much longer you are likely to be on the phone.

This property of the exponential distribution is known as “memorylessness”. That is, if the event has not occurred till some time, the time basically starts afresh. The discrete analogue, geometric distribution, also has this property.

**Gaussian Distribution**

The last continuous distribution we will look at, and by far the most prevalent in applications, is called the Normal or Gaussian distribution. It has two parameters, \(\mu\) and \(\sigma\).

**Definition 10.3 (Gaussian distribution):** For any \(\mu\) and \(\sigma > 0\), a continuous random variable \(X\) with pdf \(f\) given by

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

is called a Gaussian random variable with parameters \(\mu\) and \(\sigma\). In the special case \(\mu = 0\) and \(\sigma = 1\), \(X\) is said to have the standard Gaussian distribution and the density becomes

\[ f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}} \]

A plot of the pdf \(f\) reveals a classical "bell-shaped" curve, centered at (and symmetric around) \(x = \mu\), and with “width” determined by \(\sigma\). (The precise meaning of this latter statement will become clear when we discuss the variance below.) See Figure 3.

![Figure 3: Pdf of a Gaussian random variable with mean \(\mu\) and variance \(\sigma^2\).](image-url)
Let’s run through the usual calculations for this distribution. We first check that the density integrates to 1:

\[
\int_{-\infty}^{\infty} f(x) \, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} \, dx = 1.
\]  

(8)

The fact that this integral evaluates to 1 is a routine exercise in integral calculus, and is left as an exercise (or feel free to look it up in any standard book on probability or on the internet).

The Gaussian distribution is ubiquitous throughout the sciences and the social sciences, because it is the standard model for any aggregate data that results from a large number of independent observations of the same random variable (such as the heights of females in the US population, or the observational error in a physical experiment). Such data, as is well known, tends to cluster around its mean in a “bell-shaped” curve, with the correspondence becoming more accurate as the number of observations increases. A theoretical explanation of this phenomenon is the Central Limit Theorem, which we will discuss later.

**Expectation and variance of a continuous random variable**

By analogy with the discrete case, we define the expectation of a continuous r.v. as follows:

**Definition 10.4 (Expectation):** The expectation of a continuous random variable \( X \) with probability density function \( f \) is

\[
\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) \, dx.
\]

Note that the integral plays the role of the summation in the discrete formula \( \mathbb{E}[X] = \sum a \mathbb{P}(X = a) \). Since variance is really just another expectation, we can immediately port its definition to the continuous setting as well:

**Definition 10.5 (Variance):** The variance of a continuous random variable \( X \) with probability density function \( f \) is

\[
\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{\infty} x^2 f(x) \, dx - \left( \int_{-\infty}^{\infty} x f(x) \, dx \right)^2.
\]

As an example, we’ll now compute the expectation and variance of an exponentially distributed random variable \( X \) with pdf \( f_X(x) = \lambda e^{-\lambda x} \) for \( x \geq 0 \) and \( f_X(x) = 0 \) for \( x < 0 \).

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

\[
= \int_{0}^{\infty} x \lambda e^{-\lambda x} \, dx
\]

Using integration by parts,

\[
= -xe^{-\lambda x}\bigg|_{0}^{\infty} - \int_{0}^{\infty} (-e^{-\lambda x}) \, dx
\]

\[
= 0 - \frac{1}{\lambda} e^{-\lambda x}\bigg|_{0}^{\infty}
\]

\[
= \frac{1}{\lambda}
\]

\[
\mathbb{E}[X] = \frac{1}{\lambda}
\]
With similar computations, one can show that
\[ E[X^2] = \frac{2}{\lambda^2} \]
and hence,
\[ \text{Var}(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \]
Thus, for large \( \lambda \), the distribution is more concentrated around 0, has expectation close to zero and has low variance.
We’ll compute the mean and variance of the Gaussian distribution later.

**Cumulative Distribution Function**

Till now we’ve looked at discrete and continuous random variables. In some situations, we need to work with a mixture of these. For example, consider the waiting time at a bank. There is some positive probability of no waiting, in which case the random variable is 0. If you do have to wait, the waiting time is continuously distributed. To handle such situations, we now introduce the notion of cumulative distribution function (CDF).

**Definition 10.6 (Cumulative Distribution Function):**
The cumulative distribution function for a random variable \( X \) is a function \( F : \mathbb{R} \to \mathbb{R} \) defined to be:
\[ F_X(x) = P(X \leq x). \]  
(9)
For a continuous random variable, its relationship with the probability density function \( f \) of \( X \) is given by
\[ f_X(x) = \frac{d}{dx} F_X(x), \quad F_X(x) = \int_{-\infty}^{x} f_X(a) da. \]

The cumulative distribution function satisfies the following properties:
1. \( 0 \leq F_X(x) \leq 1 \)
2. \( \lim_{x \to -\infty} F_X(x) = 0 \)
3. \( \lim_{x \to \infty} F_X(x) = 1 \)
4. \( F_X(x) \) is monotonically non-decreasing, i.e., if \( y \geq x \), then \( F_X(y) \geq F_X(x) \).
5. \( P(X \in (a, b]) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) \)
6. \( P(X < a) = F_X(a^-) \), where \( F_X(a^-) = \lim_{\epsilon \to 0} F_X(a - \epsilon) \)
   Note that here we are working with the probability of being strictly less than \( a \). To get this probability, we need to find the left hand limit of the CDF at \( a \).