Part III

Continuous Random Variables

In this part of the book, we repeat the material in Part II, but this time we focus on continuous random variables, which can take on an uncountable number of values. Continuous random variables are very relevant to computer systems – how else can we model response time, for example? Working in continuous time also allows us to leverage everything we know about calculus.

Because continuous-time analysis is often harder for students (no one seems to remember how to integrate!), we split up our discussion of continuous random variables into two parts. In Chapter 7, we consider the case of random variables drawn from a single distribution. Here we introduce the two most common continuous distributions: the Uniform and the Exponential. In Chapter 8, we move on to multiple distributions and introduce jointly distributed continuous random variables. All the topics, such as conditioning, Bayes' Law, independence, that were covered in Part II are reintroduced in these two chapters, from the continuous perspective.

Chapter 9 is devoted to one very important continuous distribution, the Normal, a.k.a., Gaussian distribution, which occurs throughout nature. We also introduce the Central Limit Theorem, which we will use multiple times in the book as a tail approximation.

In Chapter 10 we discuss another very important continuous distribution, the Pareto distribution. This distribution also occurs throughout nature and is particularly relevant to computer science. We discuss properties of the Pareto distribution, in particular the heavy-tailed property and decreasing failure rate, and their implications for the design of computer systems.

Finally, Chapter 11 is the counterpart to Chapter 6. While z-transforms are the moment-generating function of choice for discrete random variables, the Laplace transform is the moment-generating function of choice for continuous random variables. We illustrate how the Laplace transform can be used to generate all moments of continuous random variables, and we also show how one can combine Laplace transforms and z-transforms.

7 Continuous Random Variables: Single Distribution

Until now we have only studied discrete random variables. These are defined by a probability mass function (p.m.f.). This chapter introduces continuous random variables, which are defined by a probability density function.

7.1 Probability Density Functions

Definition 7.1 *A* **continuous random variable (r.v.)** *has a continuous range of values that it can take on. This might be an interval or a set of intervals. Thus a continuous r.v. can take on an uncountable set of possible values.*

Continuous random variables are extremely common. They might be used to represent the time of an event, the speed of a device, the location of a satellite, or the distance between people's eyeballs. All these quantities can be discretized, of course, but it's more accurate to think of them as continuous random variables, and the math also gets much easier as well, since one can invoke calculus.

The probability that a continuous r.v., X, is equal to any particular value is defined to be zero. We define probability for a continuous r.v. in terms of a density function.

Definition 7.2 *The* **probability density function (p.d.f.)** *of a continuous r.v. X is a non-negative function $f_X(\cdot)$, where*

$$\mathbf{P}\{a \leq X \leq b\} = \int_a^b f_X(x)dx \qquad \text{and where} \qquad \int_{-\infty}^{\infty} f_X(x)dx = 1.$$

Definition 7.2 is illustrated in Figure 7.1. To interpret the p.d.f., $f_X(x)$, think about a very skinny rectangle of height $f_X(x)$ and width dx with area $f_X(x)dx$. This area represents a tiny probability:

$$f_X(x)dx \approx \mathbf{P}\{x \leq X \leq x + dx\}.$$

Now the integral from a to b of $f_X(x)dx$ is the sum of all these tiny probabilities.

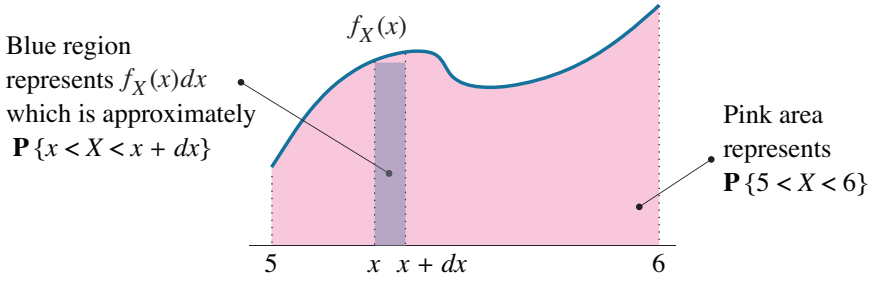

Figure 7.1 *The area under the curve represents the probability that X is between 5 and 6, namely $\int_5^6 f_X(x)dx$.*

Question: How does $\mathbf{P}\{a \leq X \leq b\}$ compare with $\mathbf{P}\{a < X < b\}$?

Answer: These are the same. For continuous distributions we don't have to be careful about differentiating between $<$ and $\leq$, because there is no mass at any particular value.

Question: Does $f_X(x)$ have to be below 1 for all x?

Answer: No, $f_X(x)$ is not a probability.

Density functions are used everywhere, and are not necessarily related to probability. We start with a typical example from a calculus class.

Example 7.3 (Density as a rate)

Imagine that we're filling a bathtub, as in Figure 7.2, where the rate of water out of the faucet starts out slow but increases over time. Specifically, let

$$f(t) = t^2, \qquad t \geq 0$$

denote the rate (in gallons/s) at which water comes out of the faucet.

Question: If we start filling at time 0, what is the total amount of water in the bathtub by time 4 seconds?

Answer: In this example, $f(t) = t^2$ is a density function, where $f(t)$ is the instantaneous rate at time t. If we want to talk about a *total amount of water*, we need to integrate the rate (density) over some period of time:

$$\int_0^4 t^2 dt = \frac{64}{3} = 21\frac{1}{3} \text{ gallons.}$$

Question: Is $f(t) = t^2$, where $t > 0$, a p.d.f.?

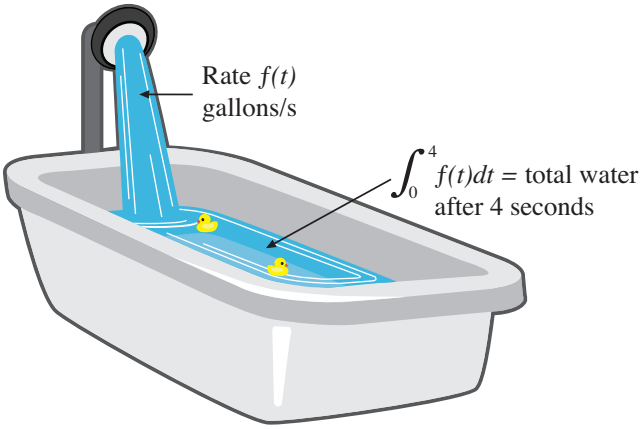

Figure 7.2 *Here, $f(t) = t^2$ represents the gallons/s coming out at time t.*

Answer: No. For $f(t)$ to be a p.d.f., it must be the case that $\int_{-\infty}^{\infty} f(t)dt = 1$, which is not true. Also, in our example $f(t)$ has no relation to probability.

Now for an example involving a p.d.f.

Example 7.4 (Weight of two-year-olds)

Let's say that the weight of two-year-olds can range anywhere from 15 pounds to 35 pounds. Let $f_W(x)$ denote the p.d.f. of weight for two-year-olds, where

$$f_W(x) = \begin{cases} \frac{3}{40} - \frac{3}{4000}(x-25)^2 & \text{if } 15 \le x \le 35 \\ 0 & \text{otherwise} \end{cases}.$$

Question: What is the fraction of two-year-olds who weigh > 30 pounds?

Answer: As illustrated in Figure 7.3,

$$\mathbf{P}\{\text{Two-year-old weighs } > 30 \text{ pounds}\} = \int_{30}^{\infty} f_W(x)dx = \int_{30}^{35} f_W(x)dx \approx 16\%.$$

Definition 7.5 *The* **cumulative distribution function (c.d.f.)** *$F(\cdot)$ of a continuous r.v. X is defined by*

$$F_X(a) = \mathbf{P}\{-\infty < X \le a\} = \int_{-\infty}^{a} f_X(x)dx.$$

We can express the **tail** *of X by*

$$\overline{F_X}(a) = 1 - F_X(a) = \mathbf{P}\{X > a\}.$$

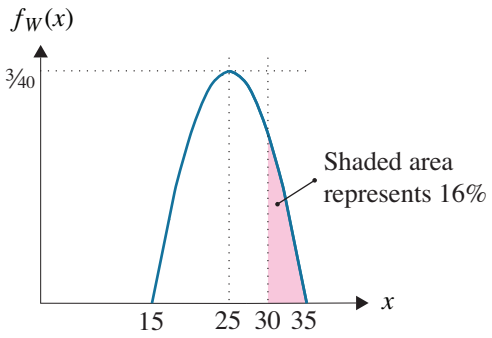

Figure 7.3 *Probability density function for the weight of two-year-olds.*

Question: We know how to get $F_X(x)$ from $f_X(x)$. How do we get $f_X(x)$ from $F_X(x)$?

Answer: By the Fundamental Theorem of Calculus (explained in Section 1.3),

$$f_X(x) = \frac{d}{dx}\int_{-\infty}^{x} f_X(t)dt = \frac{d}{dx}F_X(x).$$

7.2 Common Continuous Distributions

There are many common continuous distributions. Below we briefly define just a couple: the Uniform and Exponential distributions.

Uniform(a, b), often written $U(a, b)$, models the fact that any interval of length δ between a and b is equally likely. Specifically, if $X \sim U(a, b)$, then

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}.$$

Question: For $X \sim U(a, b)$, what is $F_X(x)$?

Answer:

$$F_X(x) = \int_a^x \frac{1}{b-a}dt = \frac{x-a}{b-a}, \qquad a \le x \le b.$$

Figure 7.4 depicts $f_X(x)$ and $F_X(x)$ graphically.

Exp(λ) denotes the Exponential distribution, whose p.d.f. drops off exponen-

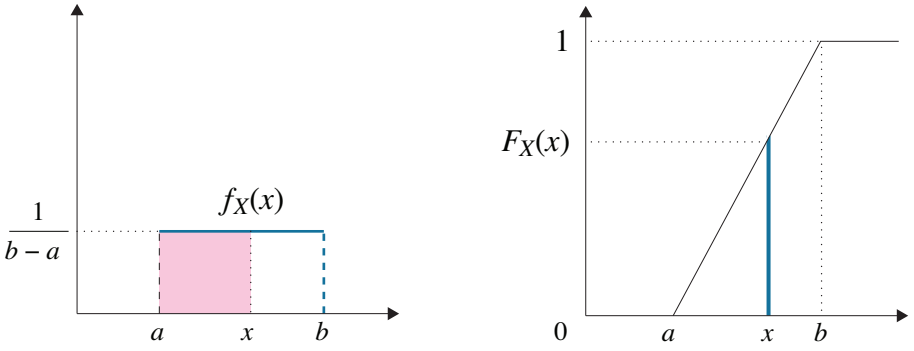

Figure 7.4 *The p.d.f., $f_X(x)$, and c.d.f., $F_X(x)$, functions for $X \sim Uniform(a, b)$. The shaded (pink) region under the p.d.f. has an area equal to the height of the blue segment in the c.d.f.*

tially. We say that a r.v. X is distributed Exponentially with *rate* $\lambda > 0$, written $X \sim \text{Exp}(\lambda)$, if

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

The graph of the p.d.f. is shown in Figure 7.5.

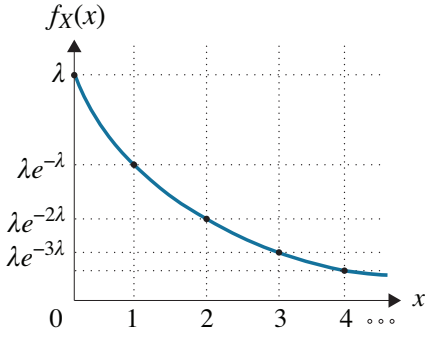

Figure 7.5 *Exponential probability density function, where $\lambda = 0.5$.*

The c.d.f., $F_X(x) = \mathbf{P}\{X \leq x\}$, is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}.$$

$$\overline{F}_X(x) = 1 - F_X(x) = e^{-\lambda x}, \quad \text{if } x \geq 0.$$

Both $f_X(x)$ and $\overline{F}_X(x)$ drop off by a *constant* factor, $e^{-\lambda}$, with each unit increase of x.

The Exponential distribution has a property called **memorylessness**.

Definition 7.6 *We say that r.v. X has the* **memoryless property** *if*

$$\mathbf{P}\{X > t+s \mid X > s\} = \mathbf{P}\{X > t\} \qquad \forall s, t \geq 0.$$

To understand memorylessness, think of X as representing the time until I win the lottery. Suppose we know that I haven't yet won the lottery by time s. Then the probability that I will need $> t$ more time to win the lottery is independent of s (that is, it's independent of how long I've been trying so far).

Equivalently, we say X is memoryless if

$$[X \mid X > s] \stackrel{d}{=} s + X, \qquad \forall s \geq 0.$$

That is, the r.v. $[X \mid X > s]$ and the r.v. $s + X$ have the same distribution.

Question: Prove that if $X \sim \text{Exp}(\lambda)$, then X has the memoryless property.

Answer:

$$\mathbf{P}\{X > t+s \mid X > s\} = \frac{\mathbf{P}\{X > t+s\}}{\mathbf{P}\{X > s\}} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbf{P}\{X > t\}.$$

Question: What other distribution has the memoryless property?

Answer: The Geometric distribution.

Question: Does the Uniform distribution also have the memoryless property?

Answer: No. If $X \sim \text{Uniform}(a, b)$ and we are given that $X > b - \epsilon$, then we know that X will end very soon.

The memoryless property is a little counter-intuitive, because it says that history doesn't affect the future.

Example 7.7 (The naked mole-rat)

Most living beings have the property that their mortality rate increases as they age. The naked mole-rat is an exception in that its remaining lifetime is independent of its age [65].

Question: Let X denote the lifetime of the naked mole-rat in years, where

Figure 7.6 *The naked mole-rat's mortality rate does not increase with age.*

$X \sim \text{Exp}(1)$. If a naked mole-rat is four years old, what is its probability of surviving at least one more year?

Answer:

$$\mathbf{P}\{X > 4+1 \mid X > 4\} = \frac{\mathbf{P}\{X > 5\}}{\mathbf{P}\{X > 4\}} = \frac{e^{-5}}{e^{-4}} = e^{-1}.$$

Question: If a naked mole-rat is 24 years old, what is its probability of surviving at least one more year?

Answer: Same thing! $\mathbf{P}\{X > 24+1 \mid X > 24\} = e^{-1}$.

Example 7.8 (Post office)

Suppose that a post office has two clerks. When customer A walks in, customer B is being served by one clerk, and customer C is being served by the other clerk. All service times are Exponentially distributed with rate λ.

Question: What is $\mathbf{P}\{A \text{ is the last to leave}\}$?

Answer: $\frac{1}{2}$. Note that one of B or C will leave first. Without loss of generality, let us say B leaves first. Then C and A will have the same distribution on their remaining service time. It does not matter that C has been served for a while.

We will return to the memoryless property of the Exponential distribution in Chapter 12. There are additional important continuous distributions, including the **Normal** and **Pareto** distributions, which we defer to Chapters 9 and 10, respectively.

7.3 Expectation, Variance, and Higher Moments

The moments of a continuous distribution are derived from its p.d.f., just as we used the p.m.f. in the case of discrete distributions. Likewise, we can also define arbitrary functions of a continuous random variable.

Definition 7.9 *For a continuous r.v. X, with p.d.f. $f_X(\cdot)$, we have:*

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$$

$$\mathbf{E}\left[X^i\right] = \int_{-\infty}^{\infty} x^i \cdot f_X(x)dx.$$

For any function $g(\cdot)$, we have:

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx.$$

In particular,

$$\mathbf{Var}(X) = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \int_{-\infty}^{\infty} (x - \mathbf{E}[X])^2 \cdot f_X(x)dx$$

Example 7.10 (The Uniform distribution)

Question: Derive the mean and variance of $X \sim \text{Uniform}(a, b)$.

Answer: Recall that

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}.$$

Thus,

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} f_X(t)tdt = \int_a^b \frac{1}{b-a}tdt = \frac{1}{b-a} \cdot \frac{b^2-a^2}{2} = \frac{a+b}{2}.$$

This answer should make sense! Likewise,

$$\mathbf{E}\left[X^2\right] = \int_{-\infty}^{\infty} f_X(t)t^2dt = \int_a^b \frac{1}{b-a}t^2dt = \frac{1}{b-a} \cdot \frac{b^3-a^3}{3} = \frac{b^2+ab+a^2}{3}.$$

After some algebra, this yields:

$$\mathbf{Var}(X) = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \frac{(b-a)^2}{12}.$$

Example 7.11 (The Exponential distribution)

Question: Derive the mean and variance of $X \sim \text{Exp}(\lambda)$.

Answer: Recall that

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

Thus,

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} f_X(t) t dt = \int_0^{\infty} \lambda e^{-\lambda t} t dt = \frac{1}{\lambda} \qquad \text{(integration by parts).}$$

Likewise,

$$\mathbf{E}\left[X^2\right] = \int_{-\infty}^{\infty} f_X(t) t^2 dt = \int_0^{\infty} \lambda e^{-\lambda t} t^2 dt = \frac{2}{\lambda^2} \qquad \text{(double integration by parts).}$$

Thus,

$$\mathbf{Var}(X) = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \frac{1}{\lambda^2}.$$

Observe that whereas the λ parameter for the Poisson distribution is also its mean, for the Exponential distribution, the λ parameter is the reciprocal of the mean. We thus refer to λ as the **rate** of the Exponential. For example, if the time until the next arrival is Exponentially distributed with rate three arrivals per second, then the expected time until the next arrival is $\frac{1}{3}$ seconds.

Example 7.12 (Time to get from NYC to Boston)

Figure 7.7 *What is the expected time to get from NYC to Boston?*

Suppose that the distance from NYC to Boston is 180 miles. You decide to buy a motorized bicycle for the trip. Suppose that motorized bikes have speeds that are Uniformly distributed between 30 and 60 m.p.h., and you buy a random motorized bike. Let T be the time to get from NYC to Boston. What is $\mathbf{E}[T]$?

Consider two ideas for figuring this out:

Idea 1: Average speed is 45 m.p.h. Thus, $\mathbf{E}[T] = \frac{180}{45} = 4$ hours.

Idea 2: $\mathbf{E}[T]$ is the average of $\frac{180}{30}$ and $\frac{180}{60}$. Thus $\mathbf{E}[T]$ is the average of 6 and 3, which is 4.5 hours.

Question: Which of ideas 1 and 2 is correct?

Answer: Neither is correct! We are interested in

$$T = \frac{180}{S},$$

where $S \sim \text{Uniform}(30, 60)$ represents the speed of the bike. Then,

$$\begin{aligned}
\mathbf{E}[T] = \mathbf{E}\left[\frac{180}{S}\right] &= \int_{30}^{60} \frac{180}{s} \cdot f_S(s) ds \\
&= \int_{30}^{60} \frac{180}{s} \cdot \frac{1}{60-30} ds \\
&= 6 \int_{30}^{60} \frac{1}{s} ds \\
&= 6 \cdot (\ln(60) - \ln(30)) \\
&\approx 4.15 \text{ hours.}
\end{aligned}$$

7.4 Computing Probabilities by Conditioning on a R.V.

Recall the Law of Total Probability for discrete random variables (Theorem 3.7) which said the following: For any event A and any *discrete* r.v. X,

$$\mathbf{P}\{A\} = \sum_x \mathbf{P}\{A \cap (X = x)\} = \sum_x \mathbf{P}\{A \mid X = x\} \cdot p_X(x) \tag{7.1}$$

The same result holds when conditioning on a continuous r.v., expect that: (1) We are working with densities, rather than probabilities, (2) we need to integrate the densities, rather than summing probabilities, and (3) when we condition on a continuous r.v., we're conditioning on a zero-probability event, which can feel a little odd but is still well defined.

Theorem 7.13 (Law of Total Probability: Continuous) *Given any event A and continuous r.v. X, we can compute $\mathbf{P}\{A\}$ by conditioning on the value of X, as follows:*

$$\mathbf{P}\{A\} = \int_{-\infty}^{\infty} f_X(x \cap A) dx = \int_{-\infty}^{\infty} \mathbf{P}\{A \mid X = x\} f_X(x) dx.$$

Here, $f_X(x \cap A)$ is notation that we're adopting to denote the density of the intersection of the event A with $X = x$.

Theorem 7.13 is analogous to (7.1), except that now the state space that we're conditioning on has been partitioned into an *uncountable* number of events of zero mass.

As an example, suppose A is the event $X > 50$. Then,

$$f_X(x \cap A) = \begin{cases} f_X(x) & \text{if } x > 50 \\ 0 & \text{if } x \leq 50 \end{cases}.$$

That is, $\forall x \leq 50$, the quantity $f_X(x \cap A)$ is simply 0, because the intersection of $X = x$ and $X > 50$ is zero. Similarly, $\forall x > 50$, the quantity $f_X(x \cap A)$ is just $f_X(x)$ because the event $X > 50$ doesn't add any new information.

Using Theorem 7.13,

$$\mathbf{P}\{X > 50\} = \mathbf{P}\{A\} = \int_{-\infty}^{\infty} f_X(x \cap A)dx = \int_{50}^{\infty} f_X(x)dx.$$

Likewise, we get this same answer by writing:

$$\mathbf{P}\{X > 50\} = \int_{-\infty}^{\infty} \mathbf{P}\{X > 50 \mid X = x\} \cdot f_X(x)dx = \int_{50}^{\infty} 1 \cdot f_X(x)dx.$$

Question: It may seem confusing to think about $\mathbf{P}\{A \mid X = x\}$. How can this possibly be well defined? If we write:

$$\mathbf{P}\{A \mid X = x\} = \frac{\mathbf{P}\{A \cap X = x\}}{\mathbf{P}\{X = x\}},$$

don't we have zero in the denominator?

Answer: Yes, we do have zero in the denominator, but we also have zero in the numerator, so this is not necessarily a problem. Both the numerator and denominator are actually densities. The correct notation is:

$$\mathbf{P}\{A \mid X = x\} = \frac{f_X(x \cap A)}{f_X(x)}.$$

Conditioning on a zero-probability event is best explained via an example.

Example 7.14 (Coin whose probability of heads is a r.v.)

Suppose we have a coin with probability P of heads, where P is drawn from a Uniform$(0, 1)$ distribution.

Question: What is the probability that the next 10 flips are all heads?

Answer:

$$\begin{aligned}\mathbf{P}\{10 \text{ Heads}\} &= \int_0^1 \mathbf{P}\{10 \text{ Heads} \mid P = p\} \cdot f_P(p)dp \\ &= \int_0^1 \mathbf{P}\{10 \text{ Heads} \mid P = p\} \cdot 1 dp \\ &= \int_0^1 p^{10} dp \\ &= \frac{1}{11}.\end{aligned}$$

As we saw above, conditioning on a zero-probability event, as in $\mathbf{P}\{10 \text{ Heads} \mid P = p\}$, makes perfect sense.

Definition 7.15 defines the conditional p.d.f., $f_{X|A}(x)$.

Definition 7.15 (Conditional p.d.f. and Bayes' Law) *For a continuous r.v. X and an event A, we define the* **conditional p.d.f.** *of r.v. X given event A as:*

$$f_{X|A}(x) = \frac{f_X(x \cap A)}{\mathbf{P}\{A\}} = \frac{\mathbf{P}\{A \mid X = x\} \cdot f_X(x)}{\mathbf{P}\{A\}}.$$

Once again, $f_X(x \cap A)$ denotes the density of the intersection of the event A with $X = x$.

Observe that $f_{X|A}(x)$ has a value of 0 when x is outside of A. The conditional p.d.f. is still a proper p.d.f. in the sense that:

$$\int_x f_{X|A}(x)dx = 1.$$

Example 7.16 (Pictorial view of conditional density)

A conditional density function can be viewed as a density function whose domain has been restricted in some way, and then scaled up to compensate. To see this, imagine we have a density function $f_X(x)$, where

$$f_X(x) > 0 \quad \text{for} \quad 0 < x < 100.$$

Now let A be the event that $X > 50$. Figure 7.8 shows $f_X(x)$ in blue/dashed and $f_{X|A}(x)$ in red/solid. The $f_X(x)$ curve is positive over the interval $[0, 100]$. The $f_{X|A}(x)$ curve is positive over the interval $[50, 100]$. The $f_{X|A}(x)$ curve is a scaled-up version of $f_X(x)$, where the scaling factor is $\frac{1}{\mathbf{P}\{X>50\}}$. This allows the

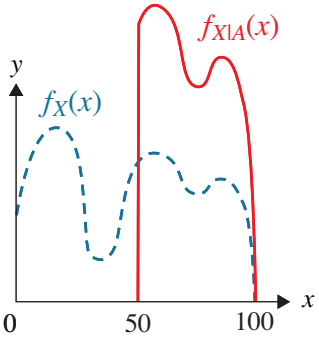

Figure 7.8 *In blue/dashed we see the p.d.f. $f_X(x)$. In red/solid we see the conditional p.d.f. $f_{X|X>50}(x)$.*

area under each curve to be 1, so both are proper probability density functions. Specifically,

$$f_{X|A}(x) = f_{X|X>50}(x) = \frac{f_X(x \cap X > 50)}{\mathbf{P}\{X > 50\}} = \begin{cases} \frac{f_X(x)}{\mathbf{P}\{X>50\}} & \text{if } x > 50 \\ 0 & \text{if } x \leq 50 \end{cases}.$$

Here we've used the fact that

$$f_X(x \cap X > 50) = \begin{cases} f_X(x) & \text{if } x > 50 \\ 0 & \text{if } x \leq 50 \end{cases}.$$

We furthermore see that the conditional p.d.f. integrates to 1:

$$\int_{x=0}^{100} f_{X|A}(x)dx = \int_{x=50}^{100} \frac{f_X(x)}{\mathbf{P}\{X > 50\}}dx = \frac{\mathbf{P}\{X > 50\}}{\mathbf{P}\{X > 50\}} = 1.$$

7.5 Conditional Expectation and the Conditional Density

One is often interested in the expected value of a random variable, conditioned on some event, A. In the continuous world this could, for example, be the expected height of people if we're restricted to people of height greater than 6 feet.

It is useful to start by recalling the definition of conditional expectation for the **discrete** space, given in Definitions 4.18 and 4.14: For a *discrete* r.v. X, and an event A, where $\mathbf{P}\{A\} > 0$, the conditional expectation of X given event A is:

$$\mathbf{E}[X \mid A] = \sum_x x \cdot p_{X|A}(x), \tag{7.2}$$

where

$$p_{X|A}(x) = \mathbf{P}\{X = x \mid A\} = \frac{\mathbf{P}\{(X = x) \cap A\}}{\mathbf{P}\{A\}}. \tag{7.3}$$

Definition 7.17 provides the corresponding definitions for a *continuous* r.v. X and an event A. Note the use of a **conditional p.d.f.** for the continuous case, where we used a conditional p.m.f. for the discrete case.

Definition 7.17 *For the case of a continuous r.v. X, corresponding to (7.2), we similarly define the* **conditional expectation** *of r.v. X given event A, where $\mathbf{P}\{A\} > 0$, as:*

$$\mathbf{E}[X \mid A] = \int_x x \cdot f_{X|A}(x)dx,$$

where $f_{X|A}(x)$ is the conditional p.d.f. defined in Definition 7.15.

Example 7.18 (Pittsburgh Supercomputing Center)

The Pittsburgh Supercomputing Center (PSC) runs large parallel jobs for scientists from all over the country. Jobs are grouped into different bins based on their *size*, where "size" denotes the required number of CPU-hours. Suppose that job sizes are Exponentially distributed with *mean* 1000 CPU-hours. Further suppose that all jobs of size less than 500 CPU-hours are sent to bin 1, and all remaining jobs are sent to bin 2.

Question: Consider the following questions:

(a) What is $\mathbf{P}\{\text{Job is sent to bin 1}\}$?
(b) What is $\mathbf{P}\{\text{Job size} < 200 \mid \text{job is sent to bin 1}\}$?
(c) What is $f_{X|A}(x)$, where X is the job size and A is the event that the job is sent to bin 1?
(d) What is $\mathbf{E}[\text{Job size} \mid \text{job is in bin 1}]$?

Answer: Start by recalling that for $X \sim \text{Exp}\left(\frac{1}{1000}\right)$ we have

$$f_X(x) = \begin{cases} \frac{1}{1000} e^{-\frac{x}{1000}} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \mathbf{P}\{X \leq x\} = 1 - e^{-\frac{1}{1000}x}.$$

(a)

$$\mathbf{P}\{\text{Job is sent to bin 1}\} = F_X(500) = 1 - e^{-\frac{500}{1000}} = 1 - e^{-\frac{1}{2}} \approx 0.39.$$

(b)

$$\mathbf{P}\{\text{Job size} < 200 \mid \text{job is sent to bin 1}\} = \frac{\mathbf{P}\{X < 200 \cap \text{bin 1}\}}{\mathbf{P}\{\text{bin 1}\}} = \frac{F_X(200)}{F_X(500)} \approx 0.46.$$

(c)

$$f_{X|A}(x) = \frac{f_X(x \cap A)}{\mathbf{P}\{A\}} = \frac{f_X(x \cap A)}{F_X(500)} = \begin{cases} \frac{f_X(x)}{F_X(500)} = \frac{\frac{1}{1000}e^{-\frac{x}{1000}}}{1-e^{-\frac{1}{2}}} & \text{if } x < 500 \\ 0 & \text{otherwise} \end{cases}.$$

We have used the fact that $f_X(x \cap A) = f_X(x)$ if and only if $x < 500$.

(d)

$$\mathbf{E}[\text{Job size} \mid \text{job in bin 1}] = \int_{-\infty}^{\infty} x f_{X|A}(x)dx = \int_0^{500} x \frac{\frac{1}{1000}e^{-\frac{x}{1000}}}{1-e^{-\frac{1}{2}}}dx \approx 229.$$

Question: Why is the expected size of jobs in bin 1 less than 250?

Answer: Consider the shape of the Exponential p.d.f. Now truncate it at 500, and scale everything by a constant needed to make it integrate to 1. There is still more weight on the smaller values, so the expected value is less than the midpoint.

Question: How would the answer to question (d) change if the job sizes were distributed Uniform$(0, 2000)$, still with mean 1000?

Answer: Logically, given that the job is in bin 1 and the distribution is Uniform, we should find that the expected job size is 250 CPU-hours. Here is an algebraic argument:

$$f_{X|A}(x) = \frac{f_X(x \cap A)}{\mathbf{P}\{A\}} = \frac{f_X(x \cap A)}{F_X(500)} = \begin{cases} \frac{f_X(x)}{F_X(500)} = \frac{\frac{1}{2000}}{\frac{500}{2000}} = \frac{1}{500} & \text{if } x < 500 \\ 0 & \text{otherwise} \end{cases}.$$

$$\mathbf{E}[\text{Job size} \mid \text{job in bin 1}] = \int_{-\infty}^{\infty} x f_{X|A}(x)dx = \int_0^{500} x \frac{1}{500}dx = 250.$$

This next example talks about a coin. However, it represents the type of math used all the time when learning the bias of humans, such as a human's likelihood for clicking on a particular type of ad, or their likelihood for buying a particular brand of shoes, etc.

Example 7.19 (Learning the bias of a coin, or a human)

Suppose that we have a biased coin, with probability P of heads. P is a r.v. in that we don't know what it is. Since we know nothing, our initial assumption is that $P \sim \text{Uniform}(0,1)$. We are interested in the expected value of P, given that the coin has resulted in 10 heads out of the first 10 flips.

At first, one might think that the best estimator of P is the fraction of heads obtained. For example, if the coin has resulted in 7 heads and 3 tails out of 10 flips, then one might be tempted to say that $\mathbf{E}[P] = 0.7$. Likewise, if the coin has resulted in 10 heads out of 10 flips, one might be tempted to say that $\mathbf{E}[P] = 1$. However, this reasoning seems shakier if you've only seen 1 flip so far, and in fact the reasoning is incorrect.

We define A as the event that 10 heads have occurred in 10 flips. By Definition 7.17,

$$\mathbf{E}[P \mid A] = \int_0^1 f_{P|A}(p) \cdot p dp,$$

where, by Definition 7.15,

$$f_{P|A}(p) = \frac{\mathbf{P}\{A \mid P = p\} \cdot f_P(p)}{\mathbf{P}\{A\}} = \frac{p^{10} \cdot 1}{\mathbf{P}\{A\}}$$

and where

$$\mathbf{P}\{A\} = \int_0^1 \mathbf{P}\{A \mid P = p\} \cdot f_P(p) dp = \int_0^1 p^{10} \cdot 1 dp = \frac{1}{11}.$$

Putting these together, we have:

$$\begin{aligned}
\mathbf{E}[P \mid A] &= \int_0^1 f_{P|A}(p) \cdot p dp \\
&= \int_0^1 \frac{p^{10} \cdot 1}{\mathbf{P}\{A\}} \cdot p dp \\
&= \int_0^1 11 p^{10} \cdot p dp \\
&= \frac{11}{12}.
\end{aligned}$$

Thus, the expected bias of the coin is not 1 but is close to 1, as one would intuit. Observe that the answer depends on our initial assumption that $P \sim \text{Uniform}(0,1)$. That initial assumption is referred to as **"the prior"** and will be the focus of Chapter 17.

7.6 Exercises

7.1 **Valid p.d.f.s**
Which of the following are plausible probability density functions?

$$f_X(x) = \begin{cases} 0.5x^{-.5} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_X(x) = \begin{cases} 2x^{-2} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_X(x) = \begin{cases} x^{-2} & \text{if } 1 < x < \infty \\ 0 & \text{otherwise} \end{cases}.$$

7.2 **Translation**
Let $X \sim \text{Exp}(\mu)$. Let $Y = 3X$. What is $f_Y(t)$?

7.3 **Weight of two-year-olds**
For Example 7.4, where W denotes the weight of two-year-olds:
(a) Derive $\mathbf{E}[W]$.
(b) Derive $\mathbf{Var}(W)$.

7.4 **Exponential distribution warm-up**
Suppose that the time a customer spends in a bank is Exponentially distributed with mean 10 minutes.
(a) What is $\mathbf{P}\{\text{Customer spends } > 5 \text{ min in bank}\}$?
(b) What is $\mathbf{P}\{\text{Customer spends } > 15 \text{ min total } \mid \text{ he is there after 10 min}\}$?

7.5 **Memorylessness**
Let $X \sim \text{Exp}(\lambda)$. What is $\mathbf{E}[X \mid X > 10]$? Solve this in two ways:
(a) By integrating the conditional p.d.f.
(b) By a two-line argument via the memoryless property of Exponential distribution.

7.6 **Memorylessness continued**
Given $X \sim \text{Exp}(1)$, what is $\mathbf{E}\left[X^2 \mid X > 10\right]$?

7.7 **Practice with conditional expectations**
Let X be a continuous r.v. with the following p.d.f.:

$$f_X(t) = \begin{cases} \frac{3}{2t^2} & \text{if } 1 < t < 3 \\ 0 & \text{otherwise} \end{cases}.$$

Derive $\mathbf{E}[X \mid 1 < X < 2]$.

7.8 **When will I hear back?**
More than 20 days ago, I interviewed at U-co for a software engineer position, but I still haven't heard back. Turns out that this is a common phenomenon. There are two types of recruiters at U-co:
- Type A: Get back to you in time Exponentially distributed with mean 20 days.
- Type B: Never get back to you.

There are an equal number of Type A and Type B recruiters at U-co. What is $\mathbf{P}\{\text{My recruiter is type B} \mid \text{I've been waiting more than 20 days}\}$?

7.9 **Alternative definition of expectation: summing the tail**
Let X be a non-negative, continuous r.v.
(a) Prove

$$\mathbf{E}[X] = \int_{x=0}^{\infty} \mathbf{P}\{X > x\}\, dx.$$

(b) What is a nicer name for this quantity?

$$\int_{x=0}^{\infty} x\mathbf{P}\{X > x\}\, dx.$$

7.10 **Transformations**
Transforming probability density functions must be handled carefully, through the cumulative distribution functions.
(a) Let $f_X(\cdot)$ denote the p.d.f. of r.v. X and $f_Y(\cdot)$ denote the p.d.f. of r.v. Y. Suppose that

$$Y = aX + b,$$

where $a > 0$ and $b > 0$ are constants. Express $f_Y(\cdot)$ in terms of $f_X(\cdot)$. You will need to work with $F_Y(y)$, the c.d.f. of Y, or you will get the wrong answer.
(b) Let $X \sim \text{Uniform}(-1, 1)$. Let $Y = e^X$. Derive the p.d.f. of Y from that of X.

7.11 **When the first alarm goes off**
Before I go to bed, I set three alarms.
- Alarm A goes off after X_A time, where $X_A \sim \text{Exp}(\lambda_A)$.
- Alarm B goes off after X_B time, where $X_B \sim \text{Exp}(\lambda_B)$.
- Alarm C goes off after X_C time, where $X_C \sim \text{Exp}(\lambda_C)$.

Assume that $X_A \perp X_B \perp X_C$. Let T denote the time until the first alarm goes off. What is $\mathbf{E}[T]$? What is $\mathbf{Var}(T)$? [Hint: It helps to start by analyzing the tail distribution of T.]

7.12 **Reliability: when the last server dies**
Nivedita has bought two very old servers to host her new online game. At

the time of purchase, she was told that each of the servers will fail at some Uniformly distributed random time during the next year, where the servers fail independently of each other. Half a year later, her game is still up, which means that at least one server did not yet fail. What is the expected time until the last server fails?

(a) Start by solving the following easier problem: Let $X_1 \sim \text{Uniform}(0, 1)$ and $X_2 \sim \text{Uniform}(0, 1)$, where $X_1 \perp X_2$. Let $X = \max(X_1, X_2)$. Derive $\mathbf{E}[X]$.

(b) The original problem is asking, what is: $\mathbf{E}\left[X \mid X > \frac{1}{2}\right]$. Derive this quantity.