Chapter 2

Random Variables

Random variables are a fundamental tool in probabilistic modeling. They allow us to model numerical quantities that are *uncertain*: the temperature in New York tomorrow, the time of arrival of a flight, the position of a satellite... Reasoning about such quantities probabilistically allows us to structure the information we have about them in a principled way.

2.1 Definition

Formally, we define a random variables as a function mapping each outcome in a probability space to a real number.

Definition 2.1.1 (Random variable)**.** *Given a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*, a random variable* X *is a function from the sample space* Ω *to the real numbers* $\mathbb{R}$*. Once the outcome* $\omega \in \Omega$ *of the experiment is revealed, the corresponding* $X(\omega)$ *is known as a* ***realization*** *of the random variable.*

Remark 2.1.2 (Rigorous definition)**.** *If we want to be completely rigorous, Definition 2.1.1 is missing some details. Consider two sample spaces* Ω_1 *and* Ω_2*, and a* σ*-algebra* $\mathcal{F}_2$ *of sets in* Ω_2*. Then, for* X *to be a random variable, there must exist a* σ*-algebra* $\mathcal{F}_1$ *in* Ω_1 *such that for any set* S *in* $\mathcal{F}_2$ *the inverse image of* S*, defined by*

$$X^{-1}(S) := \{\omega \mid X(\omega) \in S\}, \tag{2.1}$$

belongs to $\mathcal{F}_1$*. Usually, we take* Ω_2 *to be the reals* $\mathbb{R}$ *and* $\mathcal{F}_2$ *to be the Borel* σ*-algebra, which is defined as the smallest* σ*-algebra defined on the reals that contains all open intervals (amazingly, it is possible to construct sets of real numbers that do not belong to this* σ*-algebra). In any case, for the purpose of these notes, Definition 2.1.1 is sufficient (more information about the formal foundations of probability can be found in any book on measure theory and advanced probability theory).*

Remark 2.1.3 (Notation)**.** *We often denote events of the form*

$$\{X(\omega) \in \mathcal{S} \ : \ \omega \in \Omega\} \tag{2.2}$$

for some random variable X *and some set* $\mathcal{S}$ *as*

$$\{X \in \mathcal{S}\} \tag{2.3}$$

to alleviate notation, since the underlying probability space is often of no significance once we have specified the random variables of interest.

A random variable quantifies our uncertainty about the quantity it represents, *not* the value that it happens to finally take once the outcome is revealed. You should *never* think of a random variable as having a fixed numerical value. If the outcome is known, then that determines a *realization* of the random variable. In order to stress the difference between random variables and their realizations, we denote the former with uppercase letters (X, Y, ...) and the latter with lowercase letters (x, y, ...).

If we have access to the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ in which the random variable is defined, then it is straightforward to compute the probability of a random variable X belonging to a certain set S:[1] it is the probability of the event that comprises all outcomes in Ω which X maps to S,

$$\mathrm{P}\left(X \in S\right) = \mathrm{P}\left(\{\omega \mid X\left(\omega\right) \in S\}\right). \tag{2.4}$$

However, we almost never model the probability space directly, since this requires estimating the probability of every possible event in the corresponding σ-algebra. As we explain in Sections 2.2 and 2.3, there are more practical methods to specify random variables, which automatically imply that a valid underlying probability space exists. The existence of this probability space ensures that the whole framework is mathematically sound, but you don't really have to worry about it.

2.2 Discrete random variables

Discrete random variables take values on a *finite or countably infinite* subset of $\mathbb{R}$ such as the integers. They are used to model discrete numerical quantities: the outcome of the roll of a die, the score in a basketball game, etc.

2.2.1 Probability mass function

To specify a discrete random variable it is enough to determine the probability of each value that it can take. In contrast to the case of continuous random variables, this is tractable because these values are countable by definition.

Definition 2.2.1 (Probability mass function)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : \Omega \to \mathbb{Z}$ a random variable. The probability mass function (pmf) of X is defined as*

$$p_X\left(x\right) := \mathrm{P}\left(\{\omega \mid X\left(\omega\right) = x\}\right). \tag{2.5}$$

In words, $p_X\left(x\right)$ is the probability that X equals x.

We usually say that a random variable is **distributed** according to a certain pmf.

If the discrete range of X is denoted by D, then the triplet $\left(D, 2^D, p_X\right)$ is a valid probability space (recall that 2^D is the power set of D). In particular, p_x is a valid probability measure

[1] Strictly speaking, S needs to belong to the Borel σ-algebra. Again, this comprises essentially any subset of the reals that you will ever encounter in probabilistic modeling

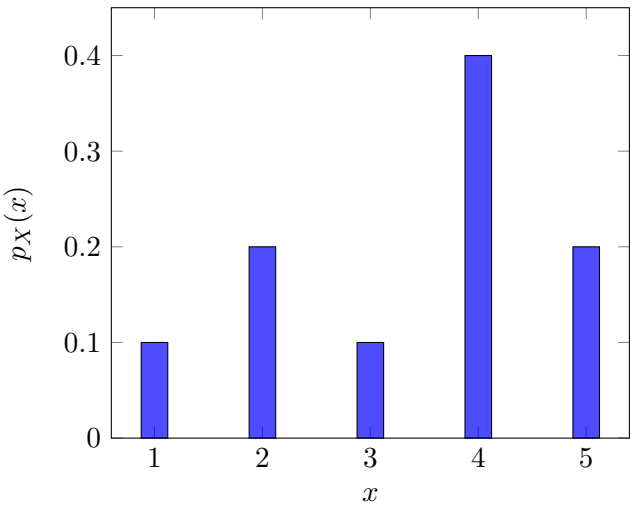

Figure 2.1: Probability mass function of the random variable X in Example 2.2.2.

which satisfies

$$p_X(x) \geq 0 \quad \text{for any } x \in D, \tag{2.6}$$

$$\sum_{x \in D} p_X(x) = 1. \tag{2.7}$$

The converse is also true, if a function defined on a countable subset D of the reals is nonnegative and adds up to one, then it may be interpreted as the pmf of a random variable. In fact, in practice we usually define discrete random variables by just specifying their pmf.

To compute the probability that a random variable X is in a certain set S we take the sum of the pmf over all the values contained in S:

$$\mathrm{P}(X \in S) = \sum_{x \in S} p_X(x). \tag{2.8}$$

Example 2.2.2 (Discrete random variable)**.** Figure 2.1 shows the probability mass function of a discrete random variable X (check that it adds up to one). To compute the probability of X belonging to different sets we apply (2.8):

$$\mathrm{P}(X \in \{1, 4\}) = p_X(1) + p_X(4) = 0.5, \tag{2.9}$$

$$\mathrm{P}(X > 3) = p_X(4) + p_X(5) = 0.6. \tag{2.10}$$

△

2.2.2 Important discrete random variables

In this section we describe several discrete random variables that are useful for probabilistic modeling.

Bernoulli

Bernoulli random variables are used to model experiments that have two possible outcomes. By convention we usually represent an outcome by 0 and the other outcome by 1. A canonical example is flipping a biased coin, such that the probability of obtaining heads is p. If we encode heads as 1 and tails as 0, then the result of the coin flip corresponds to a Bernoulli random variable with parameter p.

Definition 2.2.3 (Bernoulli)**.** *The pmf of a Bernoulli random variable with parameter $p \in [0, 1]$ is given by*

$$p_X(0) = 1 - p, \tag{2.11}$$
$$p_X(1) = p. \tag{2.12}$$

A special kind of Bernoulli random variable is the indicator random variable of an event. This random variable is particularly useful in proofs.

Definition 2.2.4 (Indicator)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. The indicator random variable of an event $S \in \mathcal{F}$ is defined as*

$$1_S(\omega) = \begin{cases} 1, & \text{if } \omega \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

By definition the distribution of an indicator random variable is Bernoulli with parameter $\mathrm{P}(S)$.

Geometric

Imagine that we take a biased coin and flip it until we obtain heads. If the probability of obtaining heads is p and the flips are independent then the probability of having to flip k times is

$$\mathrm{P}(k \text{ flips}) = \mathrm{P}(\text{1st flip} = \text{tails}, \ldots, k-1\text{th flip} = \text{tails}, k\text{th flip} = \text{heads}) \tag{2.14}$$
$$= \mathrm{P}(\text{1st flip} = \text{tails}) \cdots \mathrm{P}(k-1\text{th flip} = \text{tails})\, \mathrm{P}(k\text{th flip} = \text{heads}) \tag{2.15}$$
$$= (1-p)^{k-1} p. \tag{2.16}$$

This reasoning can be applied to any situation in which a random experiment with a fixed probability p is repeated until a particular outcome occurs, as long as the independence assumption is met. In such cases the number of repetitions is modeled as a geometric random variable.

Definition 2.2.5 (Geometric)**.** *The pmf of a geometric random variable with parameter p is given by*

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \ldots \tag{2.17}$$

Figure 2.2 shows the probability mass function of geometric random variables with different parameters. The larger p is, the more the distribution concentrates around smaller values of k.

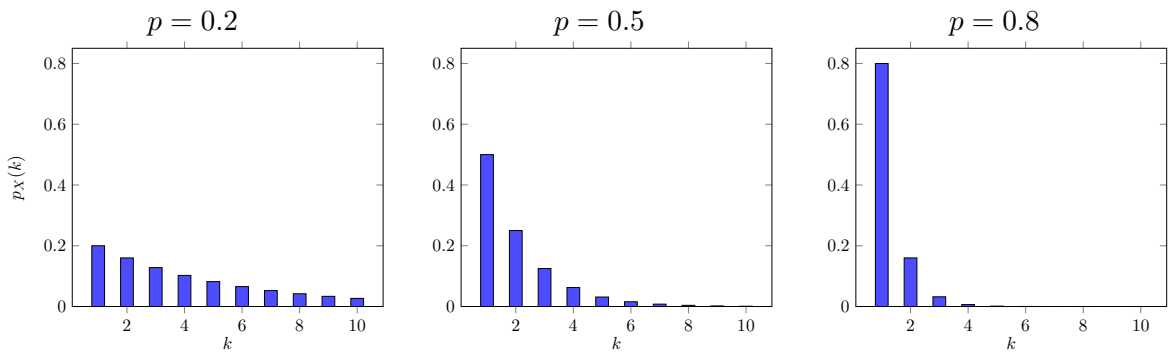

Figure 2.2: Probability mass function of three geometric random variables with different parameters.

Binomial

Binomial random variables are extremely useful in probabilistic modeling. They are used to model the number of positive outcomes of n trials modeled as independent Bernoulli random variables with the same parameter. The following example illustrates this with coin flips.

Example 2.2.6 (Coin flips)**.** If we flip a biased coin n times, what is the probability that we obtain exactly k heads if the flips are independent and the probability of heads is p?

Let us first consider a simpler problem: what is the probability of first obtaining k heads and then $n - k$ tails? By independence, the answer is

$$\begin{aligned} &\mathrm{P}\left(k \text{ heads, then } n-k \text{ tails}\right) && (2.18)\\ &= \mathrm{P}\left(1\text{st flip} = \text{heads}, \ldots, k\text{th flip} = \text{heads}, k+1\text{th flip} = \text{tails}, \ldots, n\text{th flip} = \text{tails}\right)\\ &= \mathrm{P}\left(1\text{st flip} = \text{heads}\right) \cdots \mathrm{P}\left(k\text{th flip} = \text{heads}\right) \mathrm{P}\left(k+1\text{th flip} = \text{tails}\right) \cdots \mathrm{P}\left(n\text{th flip} = \text{tails}\right)\\ &= p^k \left(1-p\right)^{n-k}. && (2.19) \end{aligned}$$

Note that the same reasoning implies that this is also the probability of obtaining exactly k heads *in any fixed order*. The probability of obtaining exactly k heads is the union of all of these events. Because these events are disjoint (we cannot obtain exactly k heads in two different orders simultaneously) we can add their individual to compute the probability of our event of interest. We just need to know the number of possible orderings. By basic combinatorics, this is given by the binomial coefficient $\binom{n}{k}$, defined as

$$\binom{n}{k} := \frac{n!}{k!\left(n-k\right)!}. \tag{2.20}$$

We conclude that

$$\mathrm{P}\left(k \text{ heads out of } n \text{ flips}\right) = \binom{n}{k} p^k \left(1-p\right)^{(n-k)}. \tag{2.21}$$

△

The random variable representing the number of heads in the example is called a binomial random variable.

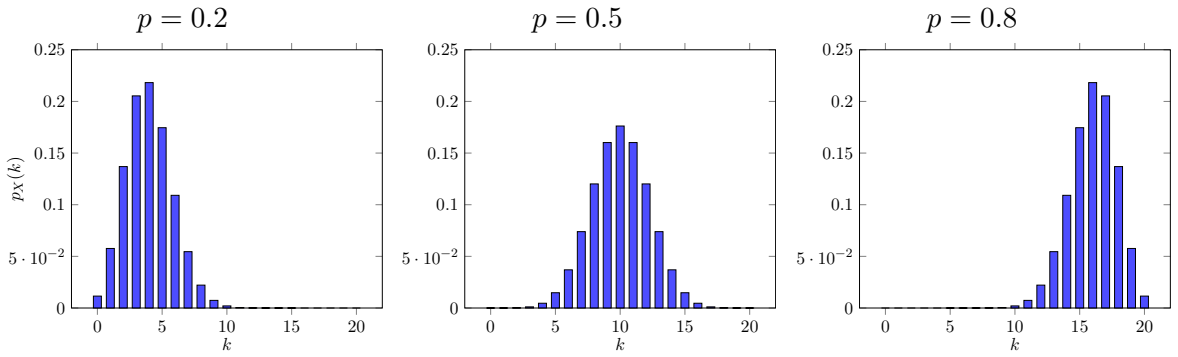

Figure 2.3: Probability mass function of three binomial random variables with different values of p and $n = 20$.

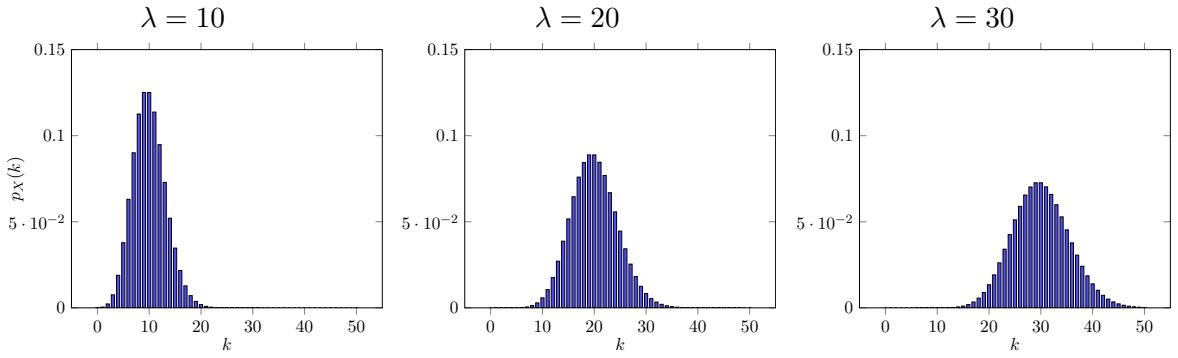

Figure 2.4: Probability mass function of three Poisson random variables with different parameters.

Definition 2.2.7 (Binomial)**.** *The pmf of a binomial random variable with parameters n and p is given by*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \quad k = 0, 1, 2, \ldots, n. \tag{2.22}$$

Figure 2.3 shows the probability mass function of binomial random variables with different values of p.

Poisson

We motivate the definition of the Poisson random variable using an example.

Example 2.2.8 (Call center)**.** A call center wants to model the number of calls they receive over a day in order to decide how many people to hire. They make the following assumptions:

1. Each call occurs independently from every other call.
2. A given call has the same probability of occurring at any given time of the day.
3. Calls occur at a rate of λ calls per day.

In Chapter 5, we will see that these assumptions define a Poisson process.

Our aim is to compute the probability of receiving exactly k calls during a given day. To do this we discretize the day into n intervals, compute the desired probability assuming each interval is very small and then let $n \to \infty$.

The probability that a call occurs in an interval of length $1/n$ is λ/n by Assumptions 2 and 3. The probability that $m > 1$ calls occur is $(\lambda/n)^m$. If n is very large this probability is negligible compared to the probability that either one or zero calls are received in the interval, in fact it tends to zero when we take the limit $n \to \infty$. The total number of calls occurring over the whole hour can consequently be approximated by the number of intervals in which a call occurs, as long as n is large enough. Since a call occurs in each interval with the same probability and calls happen independently, the number of calls over a whole day can be modeled as a binomial random variable with parameters n and $p := \lambda/n$.

We now compute the distribution of calls when the intervals are arbitrarily small, i.e. when $n \to \infty$:

$$\begin{aligned}
\mathrm{P}\left(k \text{ calls during the day }\right) &= \lim_{n\to\infty} \mathrm{P}\left(k \text{ calls in } n \text{ small intervals}\right) && (2.23)\\
&= \lim_{n\to\infty} \binom{n}{k} p^k \left(1-p\right)^{(n-k)} && (2.24)\\
&= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{(n-k)} && (2.25)\\
&= \lim_{n\to\infty} \frac{n!\,\lambda^k}{k!\,(n-k)!\,\left(n-\lambda\right)^k} \left(1-\frac{\lambda}{n}\right)^n && (2.26)\\
&= \frac{\lambda^k\, e^{-\lambda}}{k!}. && (2.27)
\end{aligned}$$

The last step follows from the following lemma proved in Section 2.7.1.

Lemma 2.2.9.

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\,\left(n-\lambda\right)^k} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{2.28}$$

△

Random variables with the pmf that we have derived in the example are called Poisson random variables. They are used to model situations where something happens from time to time at a constant rate: packets arriving at an Internet router, earthquakes, traffic accidents, etc. The number of such events that occur over a fixed interval follows a Poisson distribution, as long as the assumptions we listed in the example hold.

Definition 2.2.10 (Poisson)**.** *The pmf of a Poisson random variable with parameter λ is given by*

$$p_X\left(k\right) = \frac{\lambda^k\, e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \tag{2.29}$$

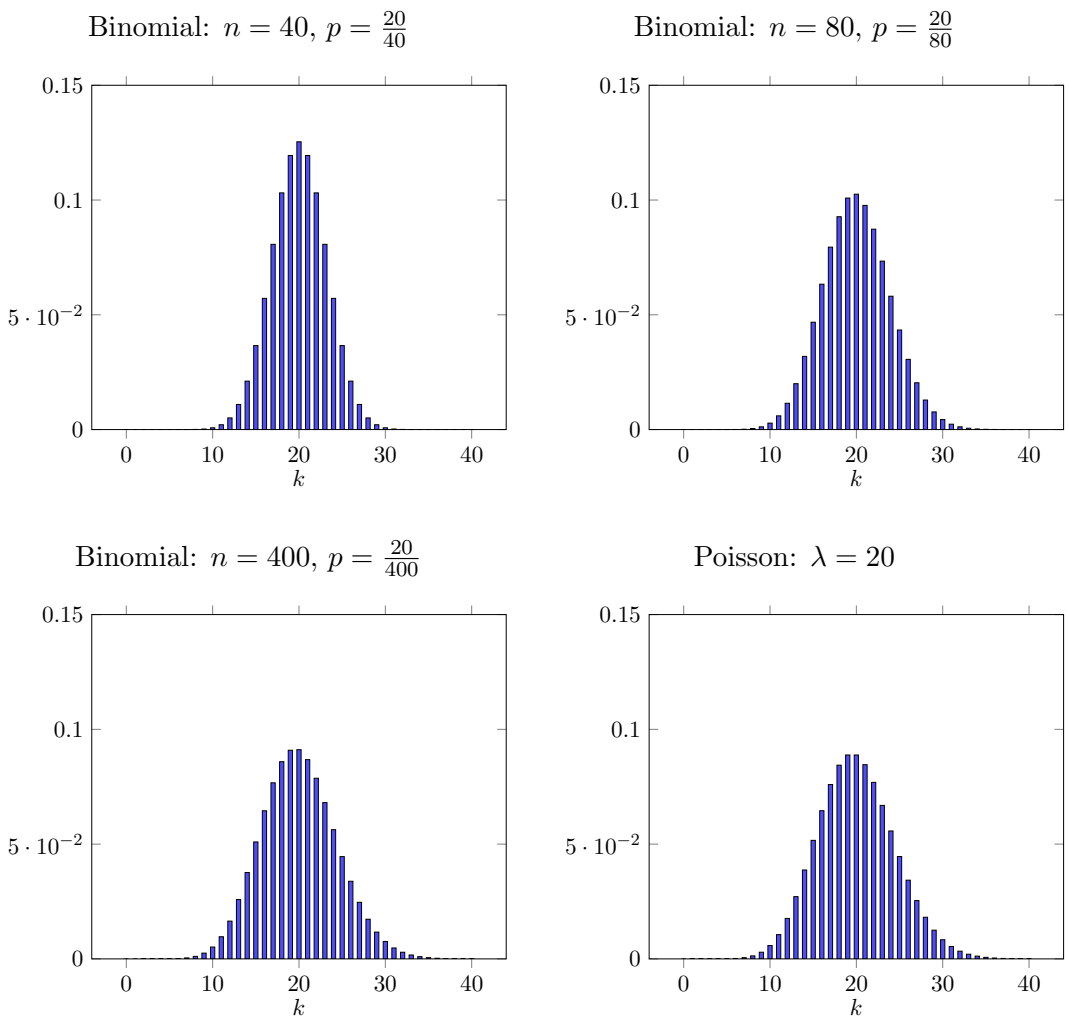

Figure 2.5: Convergence of the binomial pmf with $p = \lambda/n$ to a Poisson pmf of parameter λ as n grows.

Figure 2.4 shows the probability mass function of Poisson random variables with different values of λ. In Example 2.2.8 we prove that as $n \to \infty$ the pmf of a binomial random variable with parameters n and λ/n tends to the pmf of a Poisson with parameter λ (as we will see later in the course, this is an example of *convergence in distribution*). Figure 2.5 shows an example of this phenomenon numerically; the convergence is quite fast.

You might feel a bit skeptical about Example 2.2.8: the probability of receiving a call surely changes over the day and it must be different on weekends! That is true, but the model is actually very useful if we restrict our attention to shorter periods of time. In Figure 2.6 we show the result of modeling the number of calls received by a call center in Israel[2] over an interval of four hours (8 pm to midnight) using a Poisson random variable. We plot the histogram of the number of calls received during that interval for two months (September and October of 1999) together with a Poisson pmf fitted to the data (we will learn how to fit distributions to data later on in the course). Despite the fact that our assumptions do not hold exactly, the model

[2]The data is available here.

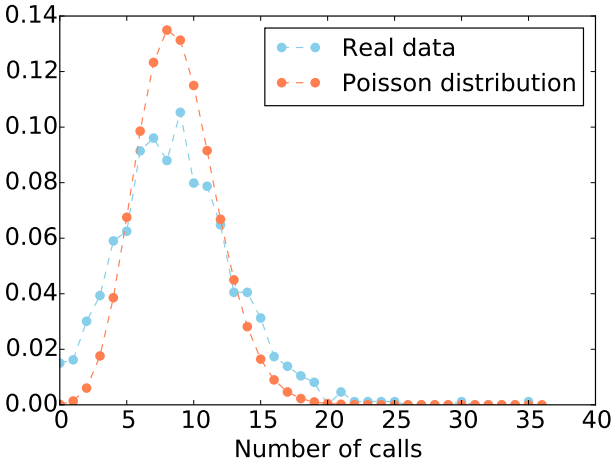

Figure 2.6: In blue, we see the histogram of the number of calls received during an interval of four hours over two months at a call center in Israel. A Poisson pmf approximating the distribution of the data is plotted in orange.

produces a reasonably good fit.

2.3 Continuous random variables

Physical quantities are often best described as continuous: temperature, duration, speed, weight, etc. In order to model such quantities probabilistically we could discretize their domain and represent them as discrete random variables. However, we may not want our conclusions to depend on how we choose the discretization grid. Constructing a continuous model allows us to obtain insights that are valid for *sufficiently fine* grids without worrying about discretization.

Precisely because continuous domains model the limit when discrete outcomes have an arbitrarily fine granularity, we *cannot* characterize the probabilistic behavior of a continuous random variable by just setting values for the probability of X being equal to individual outcomes, as we do for discrete random variables. In fact, we *cannot* assign nonzero probabilities to specific outcomes of an uncertain continuous quantity. This would result in uncountable disjoint outcomes with nonzero probability. The sum of an uncountable number of positive values is infinite, so the probability of their union would be greater than one, which does not make sense.

More rigorously, it turns out that we cannot define a valid probability measure on the power set of $\mathbb{R}$ (justifying this requires measure theory and is beyond the scope of these notes). Instead, we consider events that are composed of *unions of intervals of* $\mathbb{R}$. Such events form a σ-algebra called the Borel σ-algebra. This σ-algebra is granular enough to represent any set that you might be interested in (try thinking of a set that cannot be expressed as a countable union of intervals), while allowing for valid probability measures to be defined on it.

2.3.1 Cumulative distribution function

To specify a random variable on the Borel σ-algebra it suffices to determine the probability of the random variable belonging to all intervals of the form $(-\infty, x)$ for any $x \in \mathbb{R}$.

Definition 2.3.1 (Cumulative distribution function)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. The cumulative distribution function (cdf) of X is defined as*

$$F_X(x) := \mathrm{P}(X \leq x). \tag{2.30}$$

In words, $F_X(x)$ is the probability of X being smaller than x.

Note that the cumulative distribution function can be defined for *both continuous and discrete* random variables.

The following lemma describes some basic properties of the cdf. You can find the proof in Section 2.7.2.

Lemma 2.3.2 (Properties of the cdf)**.** *For any continuous random variable X*

$$\lim_{x \to -\infty} F_X(x) = 0, \tag{2.31}$$

$$\lim_{x \to \infty} F_X(x) = 1, \tag{2.32}$$

$$F_X(b) \geq F_X(a) \quad \text{if } b > a, \quad \text{i.e. } F_X \text{ is nondecreasing.} \tag{2.33}$$

To see why the cdf completely determines a random variable recall that we are only considering sets that can be expressed as unions of intervals. The probability of a random variable X belonging to an interval $(a, b]$ is given by

$$\begin{aligned} \mathrm{P}(a < X \leq b) &= \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) && (2.34)\\ &= F_X(b) - F_X(a). && (2.35) \end{aligned}$$

Remark 2.3.3. *Since individual points have zero probability, for any continuous random variable X*

$$\mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X < b) = \mathrm{P}(a \leq X < b). \tag{2.36}$$

Now, to find the probability of X belonging to any particular set, we only need to decompose it into disjoint intervals and apply (2.35), as illustrated by the following example.

Example 2.3.4 (Continuous random variable)**.** Consider a continuous random variable X with a cdf given by

$$F_X(x) := \begin{cases} 0 & \text{for } x < 0, \\ 0.5\,x & \text{for } 0 \leq x \leq 1, \\ 0.5 & \text{for } 1 \leq x \leq 2, \\ 0.5\left(1 + (x-2)^2\right) & \text{for } 2 \leq x \leq 3, \\ 1 & \text{for } x > 3. \end{cases} \tag{2.37}$$

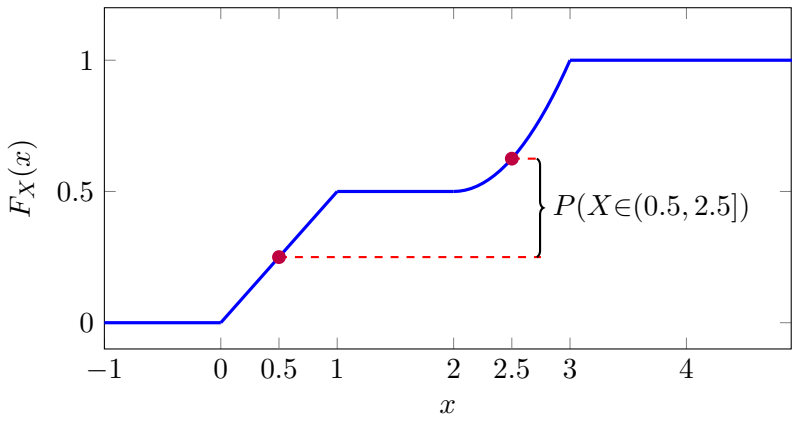

Figure 2.7: Cumulative distribution function of the random variable in Examples 2.3.4 and 2.3.7.

Figure 2.7 shows the cdf on the left image. You can check that it satisfies the properties in Lemma 2.3.2. To determine the probability that X is between 0.5 and 2.5, we apply (2.35),

$$\mathrm{P}\left(0.5 < X \leq 2.5\right) = F_X\left(2.5\right) - F_X\left(0.5\right) = 0.375, \tag{2.38}$$

as illustrated in Figure 2.7. △

2.3.2 Probability density function

If the cdf of a continuous random variable is differentiable, its derivative can be interpreted as a density function. This density can then be integrated to obtain the probability of the random variable belonging to an interval or a union of intervals (and hence to any Borel set).

Definition 2.3.5 (Probability density function)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable with cdf F_X. If F_X is differentiable then the probability density function or pdf of X is defined as*

$$f_X\left(x\right) := \frac{dF_X\left(x\right)}{d\ x}. \tag{2.39}$$

Intuitively, $f_X\left(x\right)\Delta$ is the probability of X belonging to an interval of width Δ around x as $\Delta \to 0$. By the fundamental theorem of calculus, the probability of a random variable X belonging to an interval is given by

$$\mathrm{P}\left(a < X \leq b\right) = F_X\left(b\right) - F_X\left(a\right) \tag{2.40}$$

$$= \int_a^b f_X\left(x\right) \mathrm{d}x. \tag{2.41}$$

Our sets of interest belong the Borel σ-algebra, and hence can be decomposed into unions of intervals, so we can obtain the probability of X belonging to any such set S by integrating its pdf over S

$$\mathrm{P}\left(X \in S\right) = \int_S f_X\left(x\right) \mathrm{d}x. \tag{2.42}$$

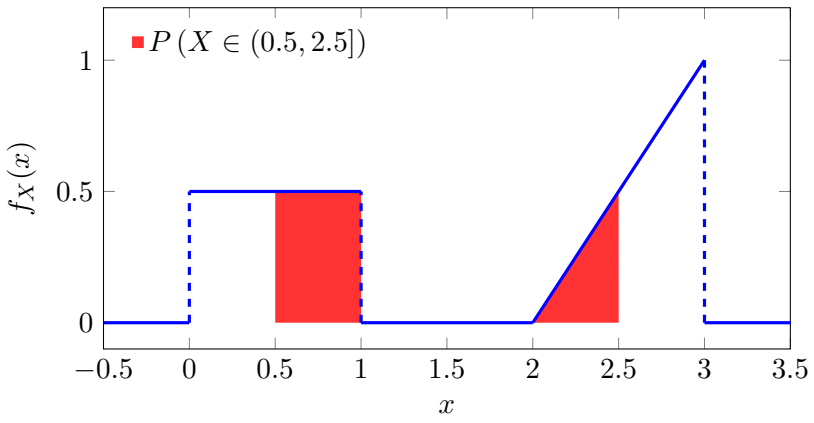

Figure 2.8: Probability density function of the random variable in Examples 2.3.4 and 2.3.7.

In particular, since X belongs to $\mathbb{R}$ by definition

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \mathrm{P}\left(X \in \mathbb{R}\right) = 1. \tag{2.43}$$

It follows from the monotonicity of the cdf (2.33) that the pdf is nonnegative

$$f_X(x) \geq 0, \tag{2.44}$$

since otherwise we would be able to find two points $x_1 < x_2$ for which $F_X(x_2) < F_X(x_1)$.

Remark 2.3.6 (The pdf is not a probability measure)**.** *The pdf is a* density *which must be integrated to yield a probability. In particular, it is not necessarily smaller than one (for example, take $a = 0$ and $b = 1/2$ in Definition 2.3.8 below).*

Finally, just as in the case of discrete random variables, we often say that a random variable is **distributed** according to a certain pdf or cdf, or that we know its distribution. The reason is that the pmf, pdf or cdf suffice to characterize the underlying probability space.

Example 2.3.7 (Continuous random variable (continued))**.** To compute the pdf of the random variable in Example 2.3.4 we differentiate its cdf, to obtain

$$f_X(x) = \begin{cases} 0 & \text{for } x < 0, \\ 0.5 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{for } 1 \leq x \leq 2 \\ x - 2 & \text{for } 2 \leq x \leq 3 \\ 0 & \text{for } x > 3. \end{cases} \tag{2.45}$$

Figure 2.8 shows the pdf. You can check that it integrates to one. To determine the probability that X is between 0.5 and 2.5, we can just integrate over that interval to obtain the same answer as in Example 2.3.4,

$$\mathrm{P}\left(0.5 < X \leq 2.5\right) = \int_{0.5}^{2.5} f_X(x)\,\mathrm{d}x \tag{2.46}$$

$$= \int_{0.5}^{1} 0.5\,\mathrm{d}x + \int_{2}^{2.5} x - 2\,\mathrm{d}x = 0.375. \tag{2.47}$$

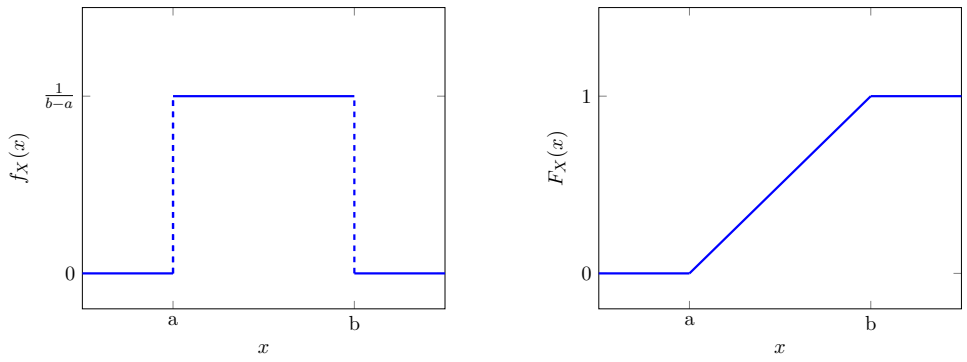

Figure 2.9: Probability density function (left) and cumulative distribution function (right) of a uniform random variable X.

Figure 2.8 illustrates that the probability of an event is equal to the area under the pdf once we restrict it to the corresponding subset of the real line.

△

2.3.3 Important continuous random variables

In this section we describe several continuous random variables that are useful in probabilistic modeling and statistics.

Uniform

A uniform random variable models an experiment in which every outcome within a continuous interval is equally likely. As a result the pdf is constant over the interval. Figure 2.9 shows the pdf and cdf of a uniform random variable.

Definition 2.3.8 (Uniform)**.** *The pdf of a uniform random variable with domain* $[a, b]$*, where* $b > a$ *are real numbers, is given by*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases} \tag{2.48}$$

Exponential

Exponential random variables are often used to model the time that passes until a certain event occurs. Examples include decaying radioactive particles, telephone calls, earthquakes and many others.

Definition 2.3.9 (Exponential)**.** *The pdf of an exponential random variable with parameter* λ *is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.49}$$

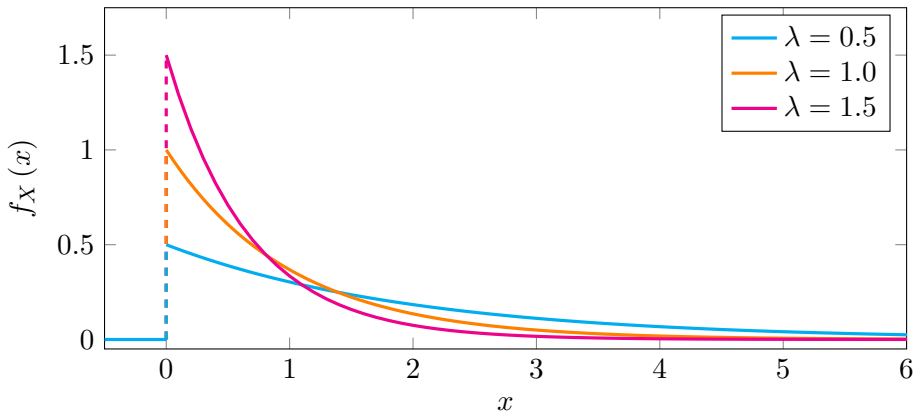

Figure 2.10: Probability density functions of exponential random variables with different parameters.

Figure 2.10 shows the pdf of three exponential random variables with different parameters. In order to illustrate that the potential of exponential distributions for modeling real data, in Figure 2.11 we plot the histogram of inter-arrival times of calls at the same call center in Israel we mentioned earlier. In more detail, these inter-arrival times are the times between consecutive calls occurring between 8 pm and midnight over two days in September 1999. An exponential model fits the data quite well.

An important property of an exponential random variable is that it is *memoryless*. We elaborate on this property, which is shared by the geometric distribution, in Section 2.4.

Gaussian or Normal

The Gaussian or normal random variable is arguably the most popular random variable in all of probability and statistics. It is often used to model variables with unknown distributions in the natural sciences. This is motivated by the fact that sums of independent random variables often converge to Gaussian distributions. This phenomenon is captured by the Central Limit Theorem, which we discuss in Chapter 6.

Definition 2.3.10 (Gaussian)**.** *The pdf of a Gaussian or normal random variable with mean μ and standard deviation σ is given by*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.50}$$

A Gaussian distribution with mean μ and standard deviation σ is usually denoted by $\mathcal{N}\left(\mu, \sigma^2\right)$.

We provide formal definitions of the mean and the standard deviation of a random variable in Chapter 4. For now, you can just think of them as quantities that parametrize the Gaussian pdf.

It is not immediately obvious that the pdf of the Gaussian integrates to one. We establish this in the following lemma.

Lemma 2.3.11 (Proof in Section 2.7.3)**.** *The pdf of a Gaussian random variable integrates to one.*

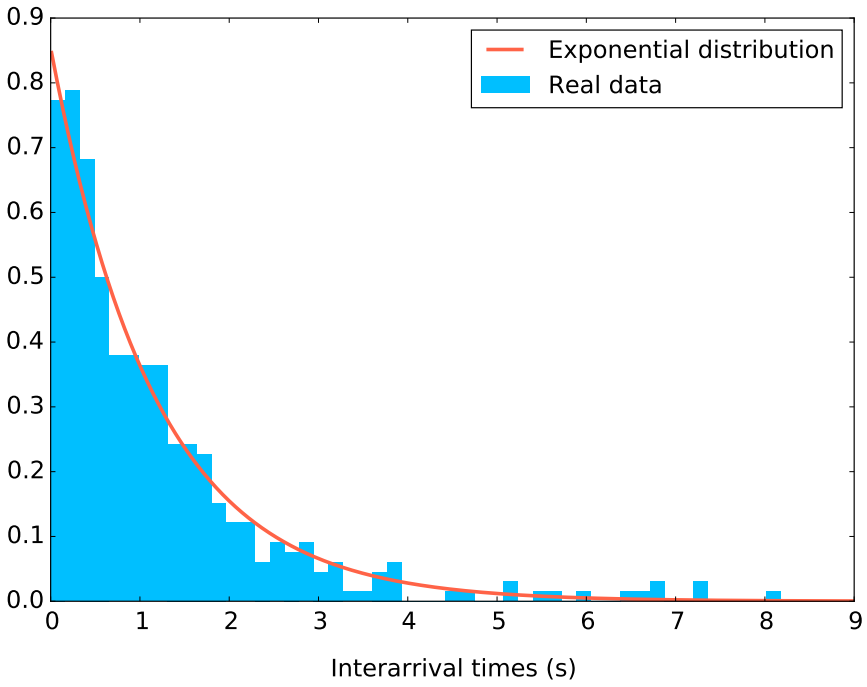

Figure 2.11: Histogram of inter-arrival times of calls at a call center in Israel (red) compared to its approximation by an exponential pdf.

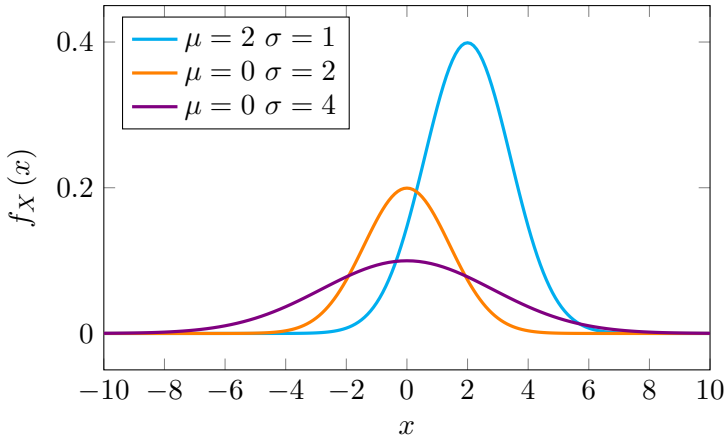

Figure 2.12: Gaussian random variable with different means and standard deviations.

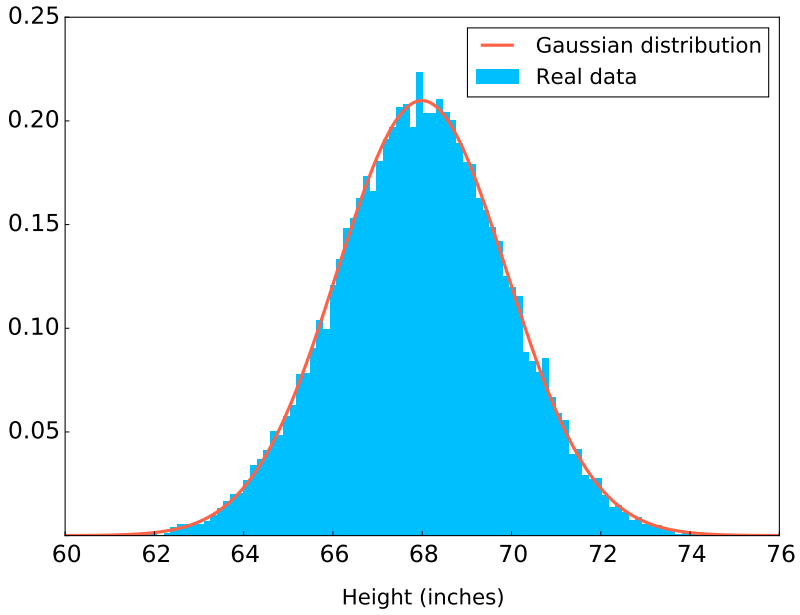

Figure 2.13: Histogram of heights in a population of 25,000 people (blue) and its approximation using a Gaussian distribution (orange).

Figure 2.12 shows the pdfs of two Gaussian random variables with different values of μ and σ. Figure 2.13 shows the histogram of the heights in a population of 25,000 people and how it is very well approximated by a Gaussian random variable[3].

An annoying feature of the Gaussian random variable is that its cdf does not have a closed form solution, in contrast to the uniform and exponential random variables. This complicates the task of determining the probability that a Gaussian random variable is in a certain interval. To tackle this problem we use the fact that if X is a Gaussian random variable with mean μ and standard deviation σ, then

$$U := \frac{X - \mu}{\sigma} \tag{2.51}$$

is a **standard** Gaussian random variable, which means that its mean is zero and its standard deviation equals one. See Lemma 2.5.1 for the proof. This allows us to express the probability of X being in an interval $[a, b]$ in terms of the cdf of a standard Gaussian, which we denote by Φ,

$$\mathrm{P}\left(X \in [a, b]\right) = \mathrm{P}\left(\frac{X - \mu}{\sigma} \in \left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right]\right) \tag{2.52}$$

$$= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \tag{2.53}$$

As long as we can evaluate Φ, this formula allows us to deal with arbitrary Gaussian random variables. To evaluate Φ people used to resort to lists of tabulated values, compiled by computing the corresponding integrals numerically. Nowadays you can just use Matlab, WolframAlpha, SciPy, etc.

[3]The data is available here.

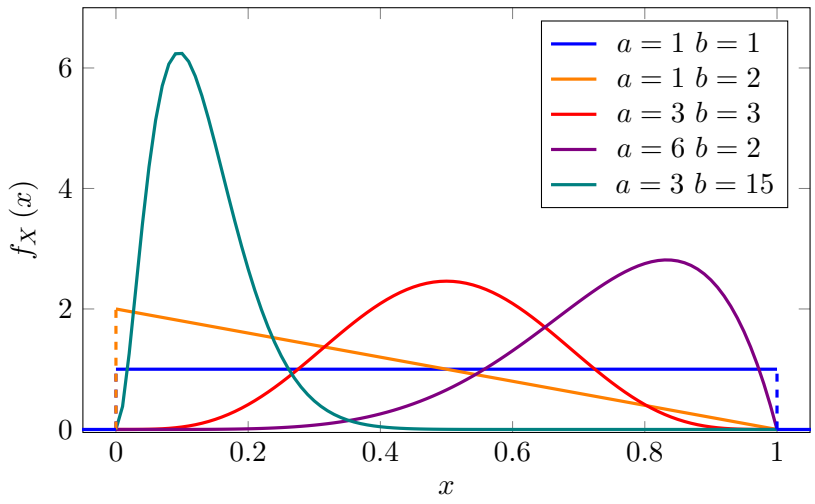

Figure 2.14: Pdfs of beta random variables with different values of the a and b parameters.

Beta

Beta distributions allow us to parametrize unimodal continuous distributions supported on the unit interval. This is useful in Bayesian statistics, as we discuss in Chapter 10.

Definition 2.3.12 (Beta distribution)**.** *The pdf of a beta distribution with parameters a and b is defined as*

$$f_\beta(\theta; a, b) := \begin{cases} \frac{\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)}, & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.54}$$

where

$$\beta(a, b) := \int_u u^{a-1}(1-u)^{b-1}\, du. \tag{2.55}$$

$\beta(a, b)$ is a special function called the beta function or Euler integral of the first kind, which must be computed numerically. The uniform distribution is an example of a beta distribution (where $a = 1$ and $b = 1$). Figure 2.14 shows the pdf of several different beta distributions.

2.4 Conditioning on an event

In Section 1.2 we explain how to modify the probability measure of a probability space to incorporate the assumption that a certain event has occurred. In this section, we review this situation when random variables are involved. In particular, we consider a random variable X with a certain distribution represented by a pmf, cdf or pdf and explain how its distribution changes if we assume that $X \in \mathcal{S}$, for any set $\mathcal{S}$ belonging to the Borel σ-algebra (remember that this includes essentially any useful set you can think of).

If X is discrete with pmf p_X, the conditional pmf of X given $X \in \mathcal{S}$ is

$$p_{X|X\in\mathcal{S}}(x) := \mathrm{P}(X = x|X \in \mathcal{S}) \tag{2.56}$$

$$= \begin{cases} \frac{p_X(x)}{\sum_{s\in\mathcal{S}} p_X(s)} & \text{if } x \in \mathcal{S} \\ 0 \quad \text{otherwise.} \end{cases} \tag{2.57}$$

This is a valid pmf in the new probability space restricted to the event $\{X \in \mathcal{S}\}$.

Similarly if X is continuous with pdf f_X, the conditional cdf of X given the event $X \in \mathcal{S}$ is

$$F_{X|X\in\mathcal{S}}(x) := \mathrm{P}(X \leq x|X \in \mathcal{S}) \tag{2.58}$$

$$= \frac{\mathrm{P}(X \leq x, X \in \mathcal{S})}{\mathrm{P}(X \in \mathcal{S})} \tag{2.59}$$

$$= \frac{\int_{u\leq x, u\in\mathcal{S}} f_X(u)\,\mathrm{d}u}{\int_{u\in\mathcal{S}} f_X(u)\,\mathrm{d}u}, \tag{2.60}$$

again by the definition of conditional probability. One can check that this is a valid cdf in the new probability space. To obtain the conditional pdf we just differentiate this cdf,

$$f_{X|X\in\mathcal{S}}(x) := \frac{\mathrm{d}F_{X|X\in\mathcal{S}}(x)}{\mathrm{d}x}. \tag{2.61}$$

We now apply this ideas to show that the geometric and exponential random variables are memoryless.

Example 2.4.1 (Geometric random variables are memoryless)**.** We flip a coin repeatedly until we obtain heads, but pause after a couple of flips (which were tails). Let us assume that the flips are independent and have the same bias p (i.e. the probability of obtaining heads in every flip is p). What is the probability of obtaining heads in k more flips? Perhaps surprisingly, it is exactly the same as the probability of obtaining a heads after k flips from the beginning.

To establish this rigorously we compute the conditional pmf of a geometric random variable X conditioned on the event $\{X > k_0\}$ (i.e. the first k_0 were tails in our example). Applying (2.56) we have

$$p_{X|X>k_0}(k) = \frac{p_X(k)}{\sum_{m=k_0+1}^{\infty} p_X(m)} \tag{2.62}$$

$$= \frac{(1-p)^{k-1}\,p}{\sum_{m=k_0+1}^{\infty}(1-p)^{m-1}\,p} \tag{2.63}$$

$$= (1-p)^{k-k_0-1}\,p \tag{2.64}$$

if $k > k_0$ and zero otherwise. We have used the fact that the geometric series

$$\sum_{m=k_0+1}^{\infty} \alpha^m = \frac{\alpha^{k_0+1}}{1-\alpha} \tag{2.65}$$

for any $\alpha < 1$.

In the new probability space where the count starts at $k_0 + 1$ the conditional pmf is that of a geometric random variable with the same parameter as the original one. The first k_0 flips don't affect the future, once it is revealed that they were tails.

△

Example 2.4.2 (Exponential random variables are memoryless)**.** Let us assume that the inter-arrival times of your emails follow an exponential distribution (over intervals of several hours this is probably a good approximation, let us know if you check). You receive an email. The time until you receive your next email is exponentially distributed with a certain parameter λ. No email arrives in the next t_0 minutes. Surprisingly, the time from then until you receive your next email is again exponentially distributed with the same parameter, no matter the value of t_0. Just like geometric random variables, exponential random variables are memoryless.

Let us prove this rigorously. We compute the conditional cdf of an exponential random variable T with parameter λ conditioned on the event $\{T > t_0\}$– for an arbitrary $t_0 > 0$– by applying (2.60)

$$F_{T|T>t_0}(t) = \frac{\int_{t_0}^{t} f_T(u)\,\mathrm{d}u}{\int_{t_0}^{\infty} f_T(u)\,\mathrm{d}u} \tag{2.66}$$

$$= \frac{e^{-\lambda t} - e^{-\lambda t_0}}{-e^{-\lambda t_0}} \tag{2.67}$$

$$= 1 - e^{-\lambda(t-t_0)}. \tag{2.68}$$

Differentiating with respect to t yields an exponential pdf $f_{T|T>t_0}(t) = \lambda e^{-\lambda(t-t_0)}$ starting at t_0. △

2.5 Functions of random variables

Computing the distribution of a function of a random variable is often very useful in probabilistic modeling. For example, if we model the current in a circuit using a random variable X, we might be interested in the power $Y := rX^2$ dissipated across a resistor with deterministic resistance r. If we apply a deterministic function $g : \mathbb{R} \to \mathbb{R}$ to a random variable X, then the result $Y := g(X)$ is *not* a deterministic quantity. Recall that random variables are functions from a sample space Ω to $\mathbb{R}$. If X maps elements of Ω to $\mathbb{R}$, then so does Y since $Y(\omega) = g(X(\omega))$. This means that Y is also a random variable. In this section we explain how to characterize the distribution of Y when the distribution of X is known.

If X is discrete, then it is straightforward to compute the pmf of $g(X)$ from the pmf of X,

$$p_Y(y) = \mathrm{P}(Y = y) \tag{2.69}$$

$$= \mathrm{P}(g(X) = y) \tag{2.70}$$

$$= \sum_{\{x \mid g(x)=y\}} p_X(x). \tag{2.71}$$

If X is continuous, the procedure is more subtle. We first compute the cdf of Y by applying the definition,

$$F_Y(y) = \mathrm{P}(Y \leq y) \tag{2.72}$$

$$= \mathrm{P}(g(X) \leq y) \tag{2.73}$$

$$= \int_{\{x \mid g(x)\leq y\}} f_X(x)\,\mathrm{d}x, \tag{2.74}$$

where the last equality obviously only holds if X has a pdf. We can then obtain the pdf of Y from its cdf if it is differentiable. This idea can be used to prove a useful result about Gaussian random variables.

Lemma 2.5.1 (Gaussian random variable)**.** *If X is a Gaussian random variable with mean μ and standard deviation σ, then*

$$U := \frac{X - \mu}{\sigma} \tag{2.75}$$

is a standard Gaussian random variable.

Proof. We apply (2.74) to obtain

$$F_U(u) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \leq u\right) \tag{2.76}$$

$$= \int_{(x-\mu)/\sigma \leq u} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, \mathrm{d}x \tag{2.77}$$

$$= \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \, \mathrm{d}w \qquad \text{by the change of variables } w = \frac{x-\mu}{\sigma}. \tag{2.78}$$

Differentiating with respect to u yields

$$f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \tag{2.79}$$

so U is indeed a standard Gaussian random variable. □

2.6 Generating random variables

Simulation is a fundamental tool in probabilistic modeling. Simulating the outcome of a model requires sampling from the random variables included in it. The most widespread strategy for generating samples from a random variable decouples the process into two steps:

1. Generating samples uniformly from the unit interval $[0, 1]$.
2. Transforming the uniform samples so that they have the desired distribution.

Here we focus on the second step, assuming that we have access to a random-number generator that produces independent samples following a uniform distribution in $[0, 1]$. The construction of good uniform random generators is an important problem, which is beyond the scope of these notes.

2.6.1 Sampling from a discrete distribution

Let X be a discrete random variable with pmf p_X and U a uniform random variable in $[0, 1]$. Our aim is to transform a sample from U so that it is distributed according to p_X. We denote the values that have nonzero probability under p_X by x_1, x_2, ...

For a fixed i, assume that we assign all samples of U within an interval of length $p_X(x_i)$ to x_i. Then the probability that a given sample from U is assigned to x_i is exactly $p_X(x_i)$!

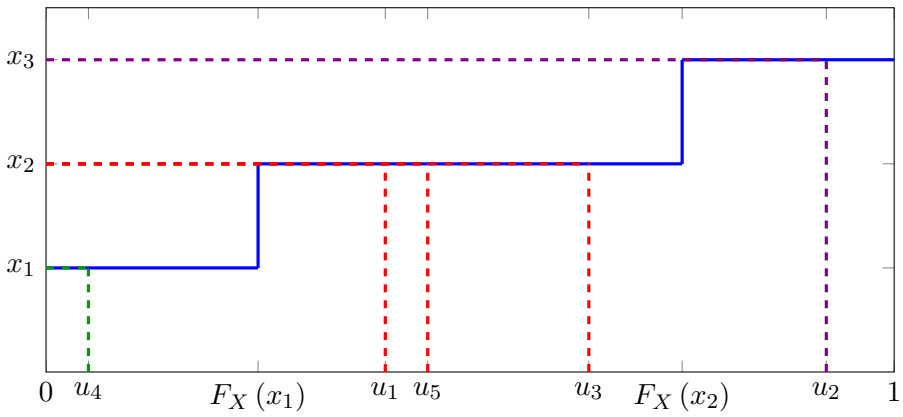

Figure 2.15: Illustration of the method to generate samples from an arbitrary discrete distribution described in Section 2.6.1. The cdf of a discrete random variable is shown in blue. The samples u_4 and u_2 from a uniform distribution are mapped to x_1 and x_3 respectively, whereas u_1, u_3 and u_5 are mapped to x_3.

Very conveniently, the unit interval can be partitioned into intervals of length $p_X(x_i)$. We can consequently generate X by sampling from U and setting

$$X = \begin{cases} x_1 & \text{if } 0 \leq U \leq p_X(x_1), \\ x_2 & \text{if } p_X(x_1) \leq U \leq p_X(x_1) + p_X(x_2), \\ \dots & \\ x_i & \text{if } \sum_{j=1}^{i-1} p_X(x_j) \leq U \leq \sum_{j=1}^{i} p_X(x_j), \\ \dots & \end{cases} \tag{2.80}$$

Recall that the cdf of a discrete random variable equals

$$F_X(x) = \mathrm{P}(X \leq x) \tag{2.81}$$

$$= \sum_{x_i \leq x} p_X(x_i), \tag{2.82}$$

so our algorithm boils down to obtaining a sample u from U and then outputting the x_i such that $F_X(x_{i-1}) \leq u \leq F_X(x_i)$. This is illustrated in Figure 2.15.

2.6.2 Inverse-transform sampling

Inverse-transform sampling makes it possible to sample from an arbitrary distribution with a known cdf by applying a deterministic transformation to uniform samples. Intuitively, we can interpret it as a generalization of the method in Section 2.6.1 to continuous distributions.

Algorithm 2.6.1 (Inverse-transform sampling)**.** *Let X be a continuous random variable with cdf F_X and U a random variable that is uniformly distributed in $[0, 1]$ and independent of X.*

1. *Obtain a sample u of U.*
2. *Set $x := F_X^{-1}(u)$.*

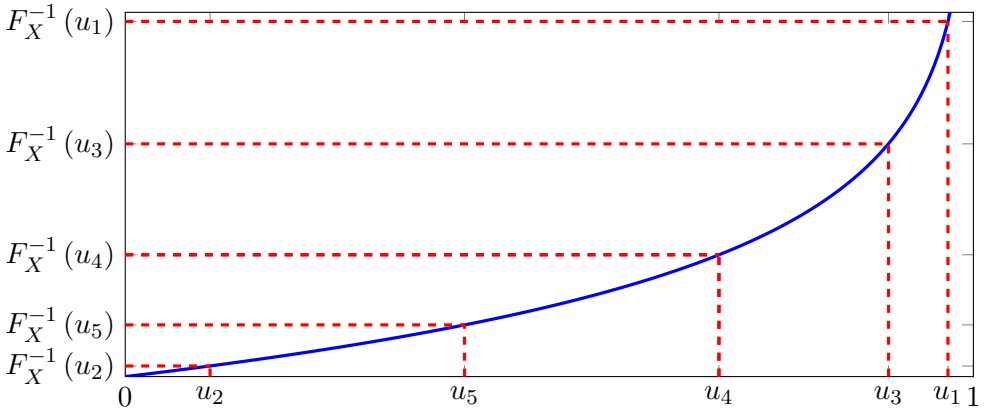

Figure 2.16: Samples from an exponential distribution with parameter $\lambda = 1$ obtained by inverse-transform sampling as described in Example 2.6.4. The samples $u_1, \ldots, u_5$ are generated from a uniform distribution.

The careful reader will point out that F_X may not be invertible at every point. To avoid this problem we define the generalized inverse of the cdf as

$$F_X^{-1}(u) := \min_x \{F_X(x) = u\}. \tag{2.83}$$

The function is well defined because all cdfs are non-decreasing, so F_X is equal to a constant c in any interval $[x_1, x_2]$ where it is not invertible.

We now prove that Algorithm 2.6.1 works.

Theorem 2.6.2 (Inverse-transform sampling works)**.** *The distribution of $Y = F_X^{-1}(U)$ is the same as the distribution of X.*

Proof. We just need to show that the cdf of Y is equal to F_X. We have

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) && (2.84)\\ &= \mathrm{P}\left(F_X^{-1}(U) \leq y\right) && (2.85)\\ &= \mathrm{P}(U \leq F_X(y)) && (2.86)\\ &= \int_{u=0}^{F_X(y)} \mathrm{d}u && (2.87)\\ &= F_X(y), && (2.88) \end{aligned}$$

where in step (2.86) we have to take into account that we are using the generalized inverse of the cdf. This is resolved by the following lemma proved in Section 2.7.4.

Lemma 2.6.3. *The events $\left\{F_X^{-1}(U) \leq y\right\}$ and $\{U \leq F_X(y)\}$ are equivalent.*

□

Example 2.6.4 (Sampling from an exponential distribution)**.** Let X be an exponential random variable with parameter λ. Its cdf $F_X(x) := 1 - e^{-\lambda x}$ is invertible in $[0, \infty]$. Its inverse equals

$$F_X^{-1}(u) = \frac{1}{\lambda} \log\left(\frac{1}{1-u}\right). \tag{2.89}$$

$F_X^{-1}(U)$ is an exponential random variable with parameter λ by Theorem 2.6.2. Figure 2.16 shows how the samples of U are transformed into samples of X.

△

2.7 Proofs

2.7.1 Proof of Lemma 2.2.9

For any fixed constants c_1 and c_2

$$\lim_{n\to\infty} \frac{n - c_1}{n - c_2} = 1, \tag{2.90}$$

so that

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\,(n-\lambda)^k} = \frac{n}{n-\lambda} \cdot \frac{n-1}{n-\lambda} \cdots \frac{n-k+1}{n-\lambda} = 1. \tag{2.91}$$

The result follows from the following basic calculus identity:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{2.92}$$

2.7.2 Proof of Lemma 2.3.2

To establish (2.31)

$$\lim_{x\to-\infty} F_X(x) = 1 - \lim_{x\to-\infty} \mathrm{P}(X > x) \tag{2.93}$$

$$= 1 - \mathrm{P}(X > 0) - \lim_{n\to\infty} \sum_{i=0}^{n} \mathrm{P}(-i \geq X > -(i+1)) \tag{2.94}$$

$$= 1 - \mathrm{P}\left(\lim_{n\to\infty} \{X > 0\} \cup \cup_{i=0}^{n} \{-i \geq X > -(i+1)\}\right) \tag{2.95}$$

$$= 1 - \mathrm{P}(\Omega) = 0. \tag{2.96}$$

The proof of (2.32) follows from this result. Let $Y = -X$, then

$$\lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} \mathrm{P}(X \leq x) \tag{2.97}$$

$$= 1 - \lim_{x\to\infty} \mathrm{P}(X > x) \tag{2.98}$$

$$= 1 - \lim_{x\to-\infty} \mathrm{P}(-X < x) \tag{2.99}$$

$$= 1 - \lim_{x\to-\infty} F_Y(x) = 1 \quad \text{by (2.32)}. \tag{2.100}$$

Finally, (2.33) holds because $\{X \leq a\} \subseteq \{X \leq b\}$.

2.7.3 Proof of Lemma 2.3.11

The result is a consequence of the following lemma.

Lemma 2.7.1.

$$\int_{-\infty}^{\infty} e^{-t^2}\, dt = \sqrt{\pi}. \tag{2.101}$$

Proof. Let us define

$$I = \int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x. \tag{2.102}$$

Now taking the square and changing to polar coordinates,

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x \int_{-\infty}^{\infty} e^{-y^2} \mathrm{d}y && (2.103)\\ &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} e^{-\left(x^2+y^2\right)} \mathrm{d}x \mathrm{d}y && (2.104)\\ &= \int_{\theta=0}^{2\pi} \int_{r=-\infty}^{\infty} r e^{-\left(r^2\right)} \mathrm{d}\theta \mathrm{d}r && (2.105)\\ &= \pi e^{-\left(r^2\right)}]_0^{\infty} = \pi. && (2.106) \end{aligned}$$

□

To complete the proof we use the change of variables $t = (x-\mu)/\sqrt{2}\sigma$.

2.7.4 Proof of Lemma 2.6.3

$\underline{\left\{F_X^{-1}(U) \leq y\right\} \text{ implies } \{U \leq F_X(y)\}}$

Assume that $U > F_X(y)$, then for all x, such that $F_X(x) = U$, $x > y$ because the cdf is nondecreasing. In particular $\min_x \{F_X(x) = U\} > y$.

$\underline{\{U \leq F_X(y)\} \text{ implies } \left\{F_X^{-1}(U) \leq y\right\}}$

Assume that $\min_x \{F_X(x) = U\} > y$, then $U > F_X(y)$ because the cdf is nondecreasing. The inequality is strict because $U = F_X(y)$ would imply that y belongs to $\{F_X(x) = U\}$, which cannot be the case as we are assuming that it is smaller than the minimum of that set.