2

Discrete Variables

Overview

In this chapter, we explain how to model uncertain numerical quantities that are *discrete*, which means that they take a finite or countably infinite number of values. Examples include the number of students attending a class, the number of goals scored in a soccer game, or the number of earthquakes that will occur in the San Francisco Bay Area in the next twenty years. Section 2.1 introduces the machinery of discrete random variables, which allows us to represent and manipulate such quantities within the mathematical framework of probability spaces. Section 2.2 describes how to build discrete nonparametric models. Section 2.3 defines several popular discrete parametric distributions. Section 2.4 explains how to fit parametric models based on these distributions to data via maximum-likelihood estimation. Finally, in Section 2.5 we compare nonparametric and parametric models, and discuss how to evaluate them in practice.

2.1 Discrete Random Variables

Random variables are mathematical objects designed to represent uncertain quantities. They do not have a fixed value. Instead, they can take multiple values with different probabilities. We do not make statements such as *the random variable* $\tilde{a}$ *equals 3*, but rather *the probability that the random variable* $\tilde{a}$ *equals 3 is 0.5*. We use a tilde to indicate that variables are random (e.g. $\tilde{a}$, $\tilde{b}$, $\tilde{x}$, $\tilde{y}$), in order to distinguish them from non-random or *deterministic* variables (a, b, x, y) representing quantities that are not uncertain.

In Section 2.1.1 we define random variables as functions that map outcomes in a probability space to real numbers. However, in data science we don't often think of random variables as functions, unless we need to prove precise mathematical statements about them. Instead, we treat them as uncertain numerical quantities with certain associated probabilities. In fact, we usually define and manipulate the random variables exclusively through these probabilities, as explained in Sections 2.1.2 and 2.1.3.

2.1.1 Random Variables As Functions Of Outcomes

In order to define random variables mathematically, we build upon the framework of probability spaces described in the previous chapter. The random variable is defined as a function of the outcomes in the probability space. Intuitively, this allows us to *capture* all the uncertainty in the probability space. This is best illustrated with an example.

Example 2.1 (Rolling a die twice)**.** Consider a probability space representing two rolls of a six-sided die. Each possible outcome can be encoded as a vector with two entries,

$$\omega := \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix}, \quad \omega_1, \omega_2 \in \{1, 2, 3, 4, 5, 6\}. \tag{2.1}$$

The first entry ω_1 is the result of the first roll, the second entry ω_2 is the result of the second roll. The sample space Ω has 36 possible outcomes.

We are interested in modeling the result of the first roll. This is an uncertain numerical quantity that can take a discrete number of values (six). The quantity can be represented as a function of the outcome of the probability space:

$$\tilde{a}(\omega) := \omega_1. \tag{2.2}$$

We call such functions random variables. Since a random variable is a function, we refer to the set of values it can take as its *range*. The range of $\tilde{a}$ is $\{1, 2, 3, 4, 5, 6\}$. The random variable allows us to describe events related to the roll of the first die very succinctly. The event

$$\textit{First roll equals one} := \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 \\ 6 \end{bmatrix} \right\} \tag{2.3}$$

can be expressed as $\{\omega : \tilde{a}(\omega) = 1\}$ or just $\tilde{a} = 1$.

Similarly, we can define random variables to represent the second roll,

$$\tilde{b}(\omega) := \omega_2, \tag{2.4}$$

or the sum of the two rolls

$$\tilde{c}(\omega) := \omega_1 + \omega_2. \tag{2.5}$$

.....

The definition of a random variable as a function confines all the uncertainty to the outcome of the probability space. Once the outcome is revealed, it simultaneously determines the value of all random variables associated to it. In Example 2.1, if $\omega = \left[\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right]$, then automatically $\tilde{a}(\omega) = 3$, $\tilde{b}(\omega) = 1$, and $\tilde{c}(\omega) = 4$. Crucially, this allows to model the dependence between different uncertain quantities, as we discuss in more detail in Chapter 4.

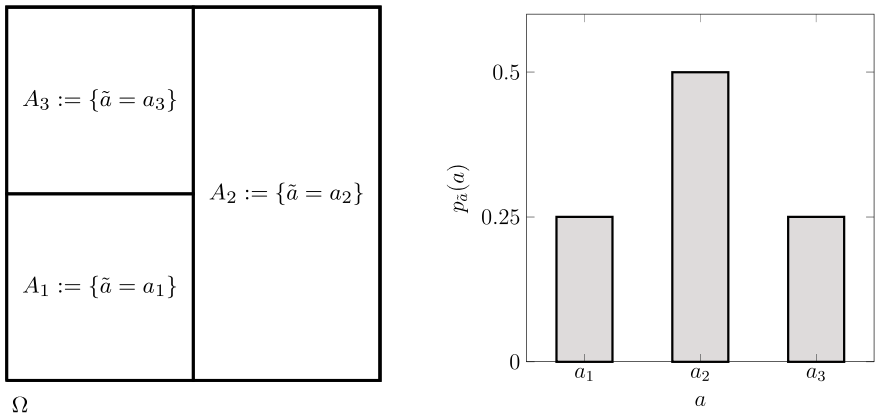

Figure 2.1 Probability mass function of a discrete random variable. The discrete random variable $\tilde{a}$ maps the outcomes in the sample space Ω to three possible values: a_1, a_2 and a_3. The three events A_1, A_2 and A_3 contain the outcomes that map to a_1, a_2 and a_3, respectively. A_1, A_2 and A_3 form a partition, as shown in the Venn diagram on the left. The graph on the right depicts the pmf of $\tilde{a}$. For $i \in \{1, 2, 3\}$, the pmf maps a_i to the probability of A_i, $p_{\tilde{a}}(a_i) = P(A_i)$, represented by the area of A_i in the Venn diagram.

2.1.2 The Probability Mass Function

Describing a random variable by explicitly defining its associated probability space is very cumbersome. For this reason, we usually manipulate random variables through their *probability mass function* (pmf), which encodes the information we are really interested in: the probability that the random variable is equal to any element of its range. Probability mass is just another name for probability, motivated by the fact that it is a measure just like mass (see Section 1.2). The *probability distribution* of a random variable $\tilde{a}$, meaning its probabilistic behavior, is completely determined by its pmf $p_{\tilde{a}}$ (we make this more precise below). Because of this, we often say that a random variable $\tilde{a}$ is *distributed* according to a certain pmf $p_{\tilde{a}}$.

Definition 2.2 (Probability mass function)**.** *Let $\tilde{a}$ be a discrete random variable defined on a probability space $(\Omega, \mathcal{C}, \mathrm{P})$ with range $a_1, a_2, \ldots$ The probability mass function (pmf) $p_{\tilde{a}} : \mathbb{R} \to [0, 1]$ of $\tilde{a}$ is the probability that $\tilde{a}$ equals each element of its range:*

$$p_{\tilde{a}}(a_i) := \mathrm{P}(A_i), \qquad i = 1, 2, \ldots, \tag{2.6}$$

where

$$A_i := \{\omega : \tilde{a}(\omega) = a_i\}. \tag{2.7}$$

Figure 2.1 illustrates this definition with a simple example. The more mathematically oriented readers might be wondering whether Definition 2.2 is sound.

How do we know that these probabilities exist? For the probabilities to be well defined, the events

$$A_i := \{\omega : \tilde{a}(\omega) = a_i\}, \qquad i = 1, 2, \ldots \tag{2.8}$$

need to belong to the collection of events of the probability space, so that the probability measure of the probability space assigns them a probability. In mathematical jargon, these events have to be *measurable*. This condition is essential for the random variable to be well defined, so we impose it explicitly in our formal mathematical definition of discrete random variables.

Definition 2.3 (Formal definition of discrete random variable)**.** *Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space. Let $\tilde{a}$ be a function mapping elements in the sample space Ω to a finite or countably infinite set of values $\{a_1, a_2, \ldots\}$. The function $\tilde{a}$ is a discrete random variable if the preimage of each element of the range*

$$A_i := \{\omega : \tilde{a}(\omega) = a_i\}, \qquad i = 1, 2, \ldots \tag{2.9}$$

is measurable, meaning that this event belongs to the collection $\mathcal{C}$ and is assigned a probability by the probability measure P, so

$$\mathrm{P}(\tilde{a} = a_i) := \mathrm{P}(A_i), \qquad i = 1, 2, \ldots \tag{2.10}$$

is well defined.

When I first came across the formal definition of random variables, I found it quite daunting. The definition seems to suggest that in order to model an uncertain quantity, we need to: (1) build a probability space with the corresponding sample space, collection of events, and probability measure, (2) define the random variable representing the quantity as a function of the probability space, (3) ensure that all the preimages are in the collection. This sounds pretty complicated. Thankfully, in practice *we never actually do this!*

The probability space associated to a random variable is a mathematical abstraction, which ensures that the random variable is properly defined. However, we only describe this probability space explicitly in simple pedagogical examples, such as Example 2.1. Otherwise, we manipulate the random variable using only its pmf, which completely characterizes its behavior, as established in Theorem 2.5. You can think of the pmf as the *user interface* of a random variable. In contrast, the formal definition of the random variable as a function is its mathematical *implementation*, which we don't need to worry about most of the time. Indeed, when using discrete random variables to model data, we estimate their pmf directly, without ever defining the underlying probability space. This is mathematically legitimate because there always exists a valid underlying probability space, as we show in Theorem 2.10.

2.1.3 Properties Of The Probability Mass Function

In Figure 2.1, the preimages A_1, A_2 and A_3 of the possible values a_1, a_2 and a_3 of the random variable form a partition of the sample space. This is a direct

consequence of the definition of a random variable as a function of the outcomes in a sample space.

Lemma 2.4. *Let $\tilde{a} : \Omega \to \mathbb{R}$ be a discrete random variable with range $\{a_1, a_2, \ldots\}$ associated to a probability space $(\Omega, \mathcal{C}, \mathrm{P})$. The events*

$$A_i := \{\omega : \tilde{a}(\omega) = a_i\}, \qquad i = 1, 2, \ldots \tag{2.11}$$

form a partition of the sample space Ω, i.e. they are disjoint and cover all of Ω.

Proof Recall that a function assigns a unique value to every element of its domain. If there is an $\omega \in \Omega$ such that $\omega \in A_i$ and $\omega \in A_j$ for $a_i \neq a_j$, then $\tilde{a}(\omega)$ would have to equal both a_i and a_j, which is impossible. The events cover Ω if $\tilde{a}$ is a valid function: every $\omega \in \Omega$ is mapped to an element of $\{a_1, a_2, \ldots\}$ and therefore belongs to A_i for some i. ■

In Example 1.14 we build a partition of events associated to a die roll explicitly, in order to define the probability measure of the probability space. By Lemma 2.4, when we define a discrete random variable and its pmf we are implicitly applying the same strategy! For instance, in Example 2.1 each random variable is implicitly associated to a different partition of the sample space.

Lemma 2.4 implies that the pmf can be used to determine the probability of a random variable belonging to any subset of its range. This is great; it frees us from having to worry about the underlying probability space! In fact, we often refer to events of the form $\{\omega : \tilde{a}(\omega) \in S\}$ using the simplified notation $\tilde{a} \in S$, without even mentioning the probability space.

Theorem 2.5. *Let $\tilde{a} : \Omega \to \mathbb{R}$ be a discrete random variable with range R associated to a probability space $(\Omega, \mathcal{C}, \mathrm{P})$. The probability that $\tilde{a}$ belongs to any subset $S \subseteq R$ of its range equals,*

$$\mathrm{P}(\tilde{a} \in S) := \mathrm{P}(\{\omega : \tilde{a}(\omega) \in S\}) = \sum_{a \in S} p_{\tilde{a}}(a). \tag{2.12}$$

Proof The proof follows from Lemma 2.4. We represent S as a union of the sets defined in (2.11). Since the sets are disjoint, by the properties of probability measures and the definition of pmf,

$$\begin{aligned} \mathrm{P}(\tilde{a} \in S) &= \mathrm{P}(\{\omega : \tilde{a}(\omega) \in S\}) && (2.13)\\ &= \mathrm{P}(\cup_{a_i \in S} A_i) && (2.14)\\ &= \sum_{a_i \in S} \mathrm{P}(A_i) && (2.15)\\ &= \sum_{a_i \in S} p_{\tilde{a}}(a_i). && (2.16) \end{aligned}$$

■

A direct consequence of Theorem 2.5 is that the sum of all the values assigned by a pmf to elements in its range must equal one.

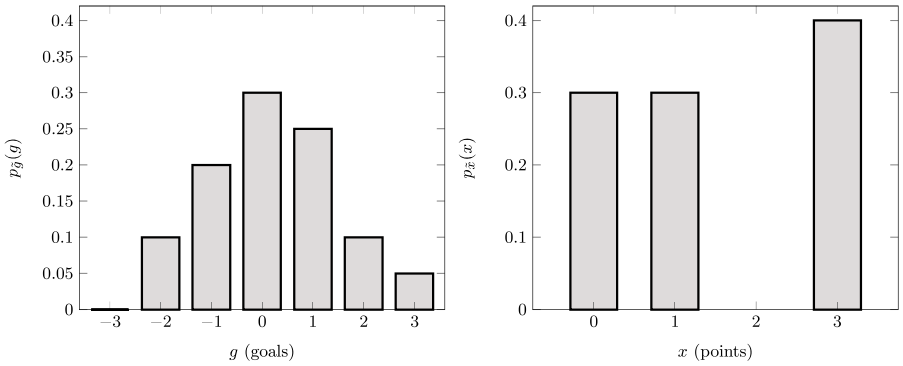

Figure 2.2 Soccer game. The figure shows the pmfs of the random variables $\tilde{g}$ and $\tilde{x}$ from Examples 2.7 and 2.9. The random variable $\tilde{g}$ represents goal difference in a soccer game. The random variable $\tilde{x}$ represents the corresponding number of points earned in the same soccer game.

Lemma 2.6 (Pmfs sum to one)**.** *For any discrete random variable* $\tilde{a} : \Omega \to \mathbb{R}$ *with range* $R := \{a_1, a_2, \ldots\}$,

$$\sum_{a \in R} p_{\tilde{a}}(a) = 1. \tag{2.17}$$

Proof The events A_1, A_2, ... in Lemma 2.4 form a partition of the sample space Ω, so by Theorem 2.5

$$\sum_{i=1,2,\ldots} p_{\tilde{a}}(a_i) = \mathrm{P}\left(\cup_i A_i\right) \tag{2.18}$$

$$= \mathrm{P}(\Omega) = 1. \tag{2.19}$$

■

Example 2.7 (Goal difference in soccer game)**.** We use a discrete random variable $\tilde{g}$ to model the goal difference in a soccer game between Barcelona and Atlético de Madrid. Figure 2.2 shows the probability mass function of $\tilde{g}$. To compute the probability of $\tilde{g}$ belonging to different sets we apply Theorem 2.5. The probability of the difference being exactly two in favor of either team is

$$\mathrm{P}\left(\tilde{g} \in \{-2, 2\}\right) = p_{\tilde{g}}\left(-2\right) + p_{\tilde{g}}\left(2\right) = 0.2. \tag{2.20}$$

The probability of Barcelona winning by more than one goal is

$$\mathrm{P}\left(\tilde{g} > 1\right) = p_{\tilde{g}}\left(2\right) + p_{\tilde{g}}\left(3\right) = 0.15. \tag{2.21}$$

2.1.4 Functions Of Random Variables

In probabilistic modeling, we often encounter deterministic functions of uncertain quantities. For instance, in Example 2.7 we might be interested in whether a team wins or not, which is a deterministic function of the goal difference (if a team scores more goals, they win). Functions of discrete random variables are discrete random variables themselves. Intuitively, if the input to a function is uncertain, then the output is also uncertain, so it makes sense for it to be represented as a random variable. The following lemma makes this mathematically rigorous. It also provides a simple formula for the pmf of a function of a random variable. To compute the probability of an output we just add up the probabilities of all the inputs mapping to that output.

Theorem 2.8 (Function of a discrete random variable)**.** *Let $\tilde{a} : \Omega \to R_{\tilde{a}}$ be a discrete random variable with range $R_{\tilde{a}} := \{a_1, a_2, \ldots\}$ associated to the probability space $(\Omega, \mathcal{C}, \mathrm{P})$, and let $h : R_{\tilde{a}} \to \mathbb{R}$ be an arbitrary function. Then $\tilde{b} := h \circ \tilde{a}$, also denoted by $h(\tilde{a})$, is a discrete random variable, and its pmf is given by*

$$p_{\tilde{b}}(b) = \sum_{\{a:h(a)=b\}} p_{\tilde{a}}(a). \tag{2.22}$$

Proof Let us denote the range of $\tilde{b}$, which contains $h(a)$ for any $a \in R_{\tilde{a}}$, as $R_{\tilde{b}} := \{b_1, b_2, \ldots\}$. Since $\tilde{a}$ is a function from Ω, $\tilde{b} := h \circ \tilde{a}$ is a function from Ω to $R_{\tilde{b}}$. If $R_{\tilde{a}}$ is discrete, then so is $R_{\tilde{b}}$, because it has at most the same cardinality (for each $b \in R_{\tilde{b}}$ there is some $a \in R_{\tilde{a}}$ such that $h(a) = b$ by definition). We just need to verify that the function is measurable, meaning that the preimages

$$B_i := \{\omega : \tilde{b}(\omega) = b_i\}, \qquad i = 1, 2, \ldots \tag{2.23}$$

belong to the collection $\mathcal{C}$. For all i, we can express B_i as

$$B_i := \cup_{h(a_j)=b_i} \{\omega : \tilde{a}(\omega) = a_j\}. \tag{2.24}$$

The set $\{\omega : \tilde{a}(\omega) = a_j\}$ for any $a_j \in R_{\tilde{a}}$ is in $\mathcal{C}$; otherwise the pmf of $\tilde{a}$ would not be well defined and $\tilde{a}$ would not be a valid random variable. Consequently, their union is also in $\mathcal{C}$, so $\tilde{b}$ is indeed measurable. Finally, by Theorem 2.5

$$\begin{aligned} p_{\tilde{b}}(b) &= \mathrm{P}(\tilde{b} = b) && (2.25)\\ &= \mathrm{P}(h(\tilde{a}) = b) && (2.26)\\ &= \sum_{\{a:h(a)=b\}} p_{\tilde{a}}(a). && (2.27) \end{aligned}$$

■

Example 2.9 (Converting goal difference to points)**.** In most soccer leagues, a win is worth three points, a draw one point, and a loss zero. Imagine that we want to model the number of points obtained by Barcelona in the soccer game of Example 2.7, but only have access to the pmf of the random variable $\tilde{g}$ modeling

the goal difference. We need to derive the pmf of a new random variable $\tilde{x} := h(\tilde{g})$ representing the points, where h is the deterministic function

$$h(g) := \begin{cases} 0 & \text{if } g < 0, \\ 1 & \text{if } g = 0, \\ 3 & \text{if } g > 0. \end{cases} \tag{2.28}$$

In order to underscore that (2.22) just follows from common sense, we compute $p_{\tilde{x}}$ from first principles:

$$\begin{aligned} p_{\tilde{x}}(0) &= \mathrm{P}\left(\tilde{x} = 0\right) && (2.29)\\ &= \mathrm{P}\left(\tilde{g} \in \{-2, -1\}\right) && (2.30)\\ &= p_{\tilde{g}}(-2) + p_{\tilde{g}}(-1) && (2.31)\\ &= 0.3. && (2.32) \end{aligned}$$

By the same argument, $p_{\tilde{x}}(1) = 0.3$, $p_{\tilde{x}}(3) = 0.4$, and $p_{\tilde{x}}(x) = 0$ for any other value of x. Figure 2.2 shows a plot of the pmf, which reassuringly sums up to one.

2.2 The Empirical Probability Mass Function

In Section 2.1 we define discrete random variables formally, and show that their pmf completely characterizes their behavior. We now explain how to apply these concepts to model discrete data in practice. The idea is to interpret the data as *realizations* or *samples* obtained by observing a random variable. To estimate the probability distribution of this random variable from the data, all we need to do is estimate its pmf. We do *not* need to define the underlying probability space. As established in the following theorem, as long as the pmf is nonnegative and sums up to one, we can rest assured that a valid probability space exists.

Theorem 2.10. *Any function $p : A \to [0, 1]$ that maps a discrete set A to nonnegative real numbers in the interval $[0, 1]$ can be interpreted as a valid pmf of a random variable, as long as*

$$\sum_{a \in A} p(a) = 1. \tag{2.33}$$

Proof We need to define a valid underlying probability space $(\Omega, \mathcal{C}, \mathrm{P})$. We set the sample space to be $\Omega := A$, the collection of events to be the power set (all possible subsets of A), and the probability measure to be p (we leave it as an exercise to check that this is a valid probability space). Note that for this probability space, the random variable associated to the pmf is the identity function. ■

In Section 1.4 we explain how to compute the empirical probability of an event from data. The pmf of a discrete random variable encodes the probability that the random variable equals each element of its range. It is therefore natural to

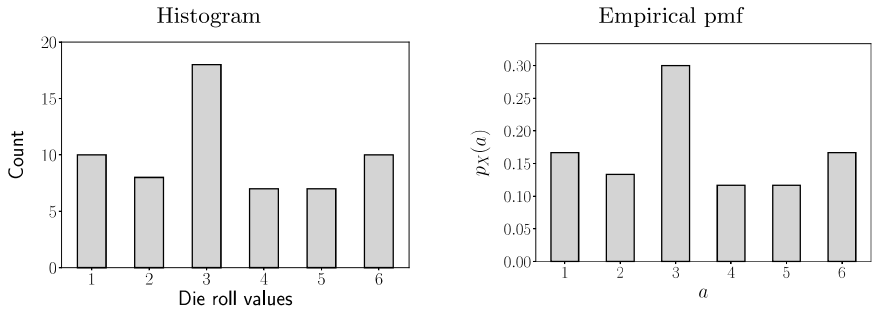

Figure 2.3 Histogram and empirical pmf for a die roll. The left plot shows the histogram of the data in Example 1.23. The empirical pmf of a random variable $\tilde{a}$ representing the die roll, shown on the right, is obtained by dividing the counts in each bin by the total number of data.

estimate pmfs using empirical probabilities. The resulting statistical estimator is known as the *empirical pmf*.

Definition 2.11 (Empirical probability mass function)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a dataset with values in a discrete set A. The empirical probability mass function $p_X : A \to [0, 1]$ maps each value $a \in A$ to the fraction of data that equal a,*

$$p_X(a) := \frac{\sum_{i=1}^{n} 1(x_i = a)}{n}, \tag{2.34}$$

where $1(x_i = a)$ is an indicator function that is equal to one if $x_i = a$ and to zero otherwise.

By Theorem 2.10 and the following simple lemma, the empirical pmf is guaranteed to be a valid pmf.

Lemma 2.12 (The empirical pmf sums up to one)**.** *Let $p_X : A \to [0, 1]$ be the empirical pmf of a dataset $X := \{x_1, x_2, \ldots, x_n\}$ taking values in a discrete set A. Then p_X sums to one,*

$$\sum_{a \in A} p_X(a) = 1. \tag{2.35}$$

Proof All x_i are in A, so $\sum_{a \in A} 1(x_i = a) = 1$, which implies

$$\sum_{a \in A} p_X(a) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in A} 1(x_i = a) = 1. \tag{2.36}$$

■

The empirical pmf has a direct connection to the *histogram*, a widely used technique to visualize datasets.

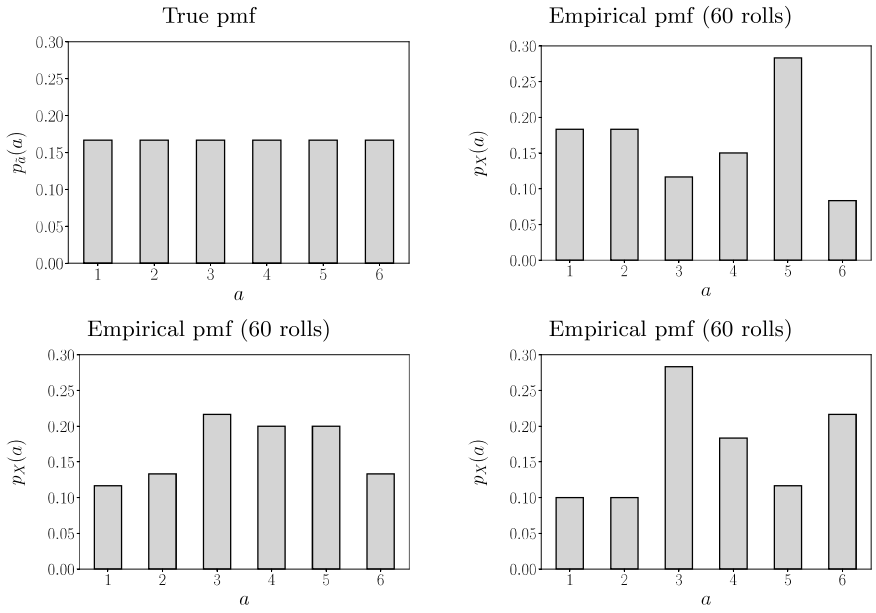

Figure 2.4 True and empirical pmf for a fair die roll. The upper left plot shows the true pmf of a random variable $\tilde{a}$ representing a fair die roll. The remaining three plots show the empirical pmfs estimated from 60 independent and identically distributed rolls (see Definition 2.23). Due to the limited data, there are substantial fluctuations in the empirical pmfs.

Definition 2.13 (Histogram of discrete data)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ be a dataset with values in a discrete set A. To build a histogram of the data we assign a* bin *or* bucket *to every element in A. We then count how many elements of X are in each bin.*

The empirical pmf is just a normalized histogram. Figure 2.3 shows the histogram of the data in Example 1.23, corresponding to rolls from a six-sided die. The count in each bin is the number of observed die rolls with the corresponding value. The empirical pmf is obtained by normalizing the counts by the total number of data.

As we discuss in Section 1.4, empirical probabilities computed from a finite number of data are only approximations, even if the data truly reflect idealized repetitions of a random phenomenon. As a result, empirical pmfs cannot be expected to be completely accurate. This is illustrated in Figure 2.4, where we show three different empirical pmf estimates obtained from 60 realizations of a fair die roll. Due to the limited number of data, there are substantial fluctuations in the estimated pmfs.

Example 2.14 (Free-throw streaks)**.** Kevin Durant is a fantastic free-throw shooter. In this example, we model the length of his streaks of consecutive made

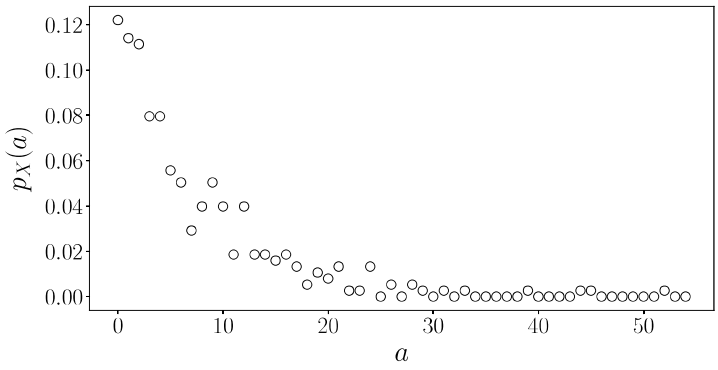

Figure 2.5 Empirical probability mass function of free-throw streaks. Probability of Kevin Durant making a certain number of consecutive free throws before missing, estimated by computing the empirical pmf of the first 3,015 free throws in his career.

free throws, which we represent by a random variable $\tilde{a}$. To this end, we compute the empirical pmf of $\tilde{a}$ using the first 3,015 free throws shot by Durant in his career, extracted from Dataset 2. To compute the empirical pmf, the 3,015 free throws are divided into 377 different streaks of varying length. For example, Durant made 2 free throws in a row (followed by a miss) 42 times. The value of the empirical pmf at 2 is therefore estimated to equal $42/377 = 0.11$. Figure 2.5 shows the empirical pmf.

The general shape of the pmf estimated in Example 2.14 looks quite reasonable, but if we look closer, we observe imprecisions due to the small number of data. The probability of a streak of length 7 is lower than a streak of length 12, even though the general trend is for longer streaks to be less likely. It seems plausible that either streaks of length 7 are underrepresented or streaks of length 12 are overrepresented in the data by mere chance.

Similarly, the probability of a streak of length 25 is zero, because no such streaks were observed. However, there are five streaks of length 24 and two streaks of length 26, which suggests that this is also due to chance. Indeed, imagine that the true probability of a streak of length 25 is around 0.01. Since the total number of streaks that we observe is less than 400, we would expect to see around four examples of such streaks in the data. However, if the same holds for streaks with similar lengths (23, 24, 26, 27, etc.), we are bound to get unlucky for one of them and not observe any at all.

2.3 Discrete Parametric Distributions

As illustrated by Example 2.14, when the available data are limited, the empirical pmf can be an inaccurate estimator of the pmf of a random variable. Parametric

modeling is a strategy to address this issue by incorporating additional assumptions about the random variable. The idea is to design a pmf that only depends on a small number of parameters, which can then estimated robustly even from limited data. We explain how to estimate these parameters in Section 2.4. In the remainder of this section, we derive several popular discrete distributions, which are widely used to build parametric models: Bernoulli (Section 2.3.1), geometric (Section 2.3.3), binomial (Section 2.3.2), and Poisson (Section 2.3.4).

2.3.1 The Bernoulli Distribution

The Bernoulli distribution is used to model uncertain phenomena that have two possible results, such as a basketball game (win or loss) or a driving test (pass or fail). By convention we often represent one of the results by 0 and the other by 1. A canonical example is flipping a biased coin, where the probability of obtaining heads is θ. 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 θ.

Definition 2.15 (Bernoulli distribution)**.** *The pmf of a Bernoulli random variable $\tilde{a}$ with parameter $\theta \in [0, 1]$ is*

$$p_{\tilde{a}}(0) = 1 - \theta, \tag{2.37}$$

$$p_{\tilde{a}}(1) = \theta. \tag{2.38}$$

2.3.2 The Binomial Distribution

We motivate the definition of the binomial distribution with a simple example involving coin flips.

Example 2.16 (Coin flips)**.** We flip a biased coin n times. If the flips are independent and the probability of heads is θ, what is the probability of obtaining a heads? Let us first consider the case where $n = 3$ and $a = 2$ to gain some intuition. We are interested in the probability the event *obtaining two heads*, which occurs if the sequence of flips equals *heads-heads-tails*, *heads-tails-heads*, or *tails-heads-heads*. Since these individual events are disjoint (the sequence cannot be *heads-tails-heads* and *tails-heads-heads* at the same time), we can just add their probabilities to compute the probability of their union. This suggests a strategy to answer the question: compute the probabilities of obtaining a heads and $n - a$ tails *in every possible order* and then sum them up.

The probability of first obtaining a heads (denoted by h) and then $n - a$ tails (denoted by t) equals

$$\begin{aligned}
&\mathrm{P}\left(1\text{st flip} = \mathrm{h}, \ldots, a\text{th flip} = \mathrm{h}, a+1\text{th flip} = \mathrm{t}, \ldots, n\text{th flip} = \mathrm{t}\right) \\
&= \mathrm{P}\left(1\text{st flip} = \mathrm{h}\right) \cdots \mathrm{P}\left(a\text{th flip} = \mathrm{h}\right) \mathrm{P}\left(a+1\text{th flip} = \mathrm{t}\right) \cdots \mathrm{P}\left(n\text{th flip} = \mathrm{t}\right) \\
&= \theta^a \left(1-\theta\right)^{n-a}.
\end{aligned} \tag{2.39}$$

By the same reasoning, this is the probability of obtaining exactly a heads in any

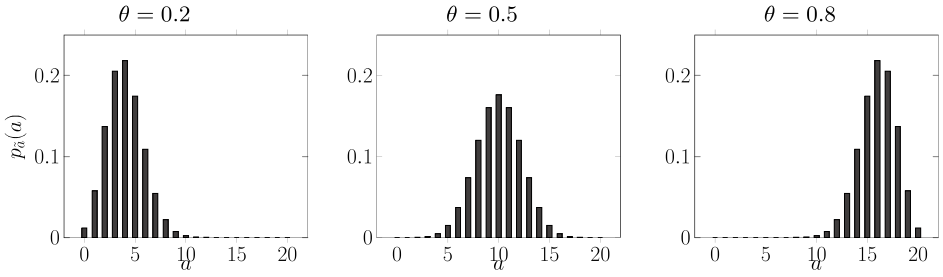

Figure 2.6 Binomial pmf. The figure shows the pmf of a binomial random variable $\tilde{a}$ for $n = 20$ and $\theta = 0.2, 0.5, 0.8$.

fixed order. By basic combinatorics, the number of possible orders is given by the binomial coefficient

$$\binom{n}{a} := \frac{n!}{a!\,(n-a)!}. \tag{2.40}$$

We conclude that

$$\mathrm{P}\left(a \text{ heads out of } n \text{ flips}\right) = \binom{n}{a} \theta^a \left(1-\theta\right)^{n-a}. \tag{2.41}$$

. . . . . .

The distribution derived in the example is known as the binomial distribution. It is widely used to model situations where there are n independent trials with a binary result that has constant probability. Figure 2.6 shows binomial pmfs with different values of θ.

Definition 2.17 (Binomial distribution)**.** *The pmf of a binomial random variable $\tilde{a}$ with parameters n and θ is given by*

$$p_{\tilde{a}}\left(a\right) = \binom{n}{a} \theta^a \left(1-\theta\right)^{n-a}, \quad a = 0, 1, \ldots, n. \tag{2.42}$$

The binomial distribution enables us to analyze the behavior of the empirical-probability estimator. As we discuss in Section 1.4, whenever we compute an empirical probability using a finite number of data, the estimate is subject to random fluctuations. In the following example, we analyze these random fluctuations under the assumption that the data are sampled according to a true underlying probability, and that these samples are independent. This illustrates an important application of probability-theory concepts in statistics: the theoretical analysis of statistical estimators with the goal of understanding their behavior in idealized settings.

Example 2.18 (Analysis of the empirical-probability estimator)**.** Recall that to

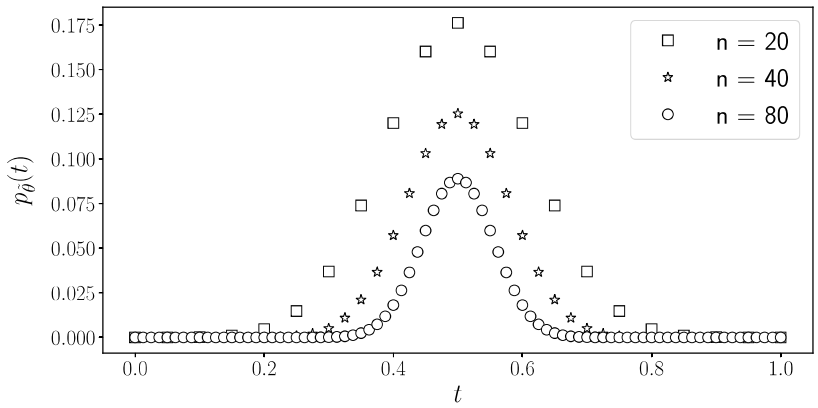

Figure 2.7 Analysis of the empirical-probability estimator. Pmfs of the empirical-probability estimator under the assumptions of Example 2.18 when $\theta = 0.5$ and the number of data n is 20, 40 and 80. As n increases, the estimator concentrates around the true probability θ.

compute the empirical probability of an event A from n data $x_1, x_2, \ldots, x_n$, we count how many data points are in A and divide by n (see Definition 1.22).

In order to analyze the estimator, we model the data probabilistically. Let B_i denote the event that the ith data point belongs to A. We assume

$$\mathrm{P}\left(B_i\right) = \theta \qquad \text{for } 1 \leq i \leq n \tag{2.43}$$

and also that the events and their complements are all independent. To be more precise for any sequence $S_1, S_2, \ldots, S_n$, where S_i is either B_i or B_i^c, the events in the sequence are mutually independent.

We represent the number of data points in A as a random variable $\tilde{c}$. Under our assumptions, $\tilde{c}$ has exactly the same distribution as the number of coin flips in Example 2.16, and is therefore binomial with parameters n and θ,

$$p_{\tilde{c}}(c) = \binom{n}{c} \theta_{\text{true}}^c \left(1 - \theta_{\text{true}}\right)^{n-c}, \quad c = 0, 1, 2, \ldots, n. \tag{2.44}$$

Our goal is to model the empirical probability, which is equal to the normalized count represented by the random variable $\tilde{\theta} := \frac{\tilde{c}}{n}$. By (2.44) and Theorem 2.8

$$p_{\tilde{\theta}}(t) = \mathrm{P}(\tilde{c} = nt) \tag{2.45}$$

$$= \binom{n}{nt} \theta_{\text{true}}^{nt} \left(1 - \theta_{\text{true}}\right)^{n-nt}, \quad t = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1. \tag{2.46}$$

Figure 2.7 shows the pmf of the empirical-pmf estimator under these assumptions when $\theta_{\text{true}} = 0.5$ and the number of data is 20, 40 and 80. As we gather more data, the pmf concentrates around the true value of the probability. To make this

statement more precise, let us compute the probability of making an error of less than 0.1 when approximating θ_{true}. By Theorem 2.5,

$$\mathrm{P}(|\tilde{\theta} - \theta_{\text{true}}| \leq 0.1) = \sum_{t \in [\theta_{\text{true}} - 0.1, \theta_{\text{true}} + 0.1]} p_{\tilde{\theta}}(t) \tag{2.47}$$

$$= \sum_{k \in [n\theta_{\text{true}} - 0.1n, n\theta_{\text{true}} + 0.1n]} \binom{n}{k} \theta_{\text{true}}^k \left(1 - \theta_{\text{true}}\right)^{n-k}. \tag{2.48}$$

For $\theta_{\text{true}} = 0.5$, the probabilities are 0.737 ($n = 20$), 0.846 ($n = 40$), and 0.943 ($n = 80$). Reassuringly, the probability increases with the number of data points, and is already very high for $n := 80$; the error is more than 0.1 only about 5% of the time. Theorem 9.24 establishes that this is not a coincidence: the empirical-probability estimator is guaranteed to converge in probability to the true probability of the event of interest, as long as it is computed using independent data.

2.3.3 The Geometric Distribution

In the following example we derive a parametric model for the free-throw data in Example 2.14.

Example 2.19 (Parametric model for free-throw streaks)**.** Our parametric model is based on two simplifying assumptions:

1 The free throws are all independent.
2 The probability of making each individual free throw is always the same.

Let θ denote the probability of making each free throw and let $\tilde{s}$ be a discrete random variable representing the length of a streak of made free throws. Under our assumptions, the value of the pmf of $\tilde{s}$ at s, $s \in \{0, 1, 2, \ldots\}$ equals

$$p_{\tilde{s}}(s) = \mathrm{P}(s \text{ free throws are made, followed by a miss}) \tag{2.49}$$

$$= \mathrm{P}(\text{1st made})\mathrm{P}(\text{2nd made}) \cdots \mathrm{P}(s\text{th made})\mathrm{P}(s+1\text{th missed}) \tag{2.50}$$

$$= \theta^s(1 - \theta). \tag{2.51}$$

You may be questioning the assumptions used to derive the model in Example 2.19, and you would be right to do so. The probability of making a free throw probably changes depending on the type of game (e.g. regular season vs. playoffs), the time at which it is taken (e.g. beginning of the game vs. end of the game), and the state of mind of the player (e.g. if the game is close, contract negotiations are coming up, etc.). However, in practice, we often need to make assumptions that are not entirely correct in order to design tractable models. The key, paraphrasing George Box, is to ensure that the model is *wrong, but useful*. The simple model derived in Example 2.19 is definitely useful, in Section 2.5 we show that it yields good predictive performance.

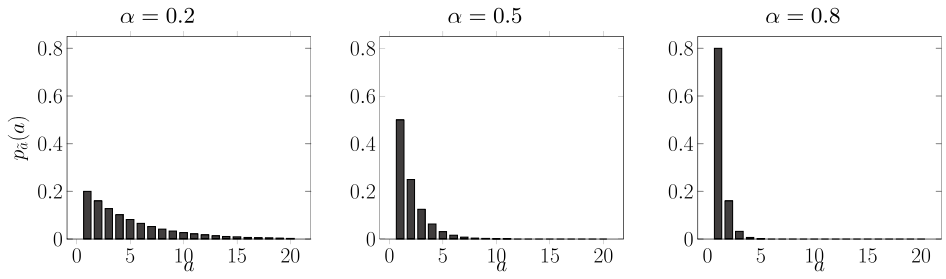

Figure 2.8 Geometric pmf. The figure shows the first twenty entries of the pmf of a geometric random variable $\tilde{a}$ for $\alpha = 0.2, 0.5, 0.8$

The model derived in Example 2.19 is a popular parametric distribution called the *geometric* distribution. It can be used to model the number of coin flips we need until we obtain heads, or the number of times we use a device such as a computer or a washing machine before it stops working. The following standard definition of the geometric distribution is slightly different, but equivalent to our derivation (set $\alpha := 1 - \theta$, and $a := s + 1$). Figure 2.8 shows some examples of geometric pmfs.

Definition 2.20 (Geometric distribution)**.** *The pmf of a geometric random variable $\tilde{a}$ with parameter $\alpha \in (0, 1)$ is*

$$p_{\tilde{a}}(a) = (1 - \alpha)^{a-1} \alpha, \quad a = 1, 2, \ldots \tag{2.52}$$

2.3.4 The Poisson Distribution

The Poisson distribution is used to model phenomena that occur randomly but at a constant rate: particle emissions by a radioactive substance, packets arriving at an Internet router, neuronal spikes, action potentials in neurons, etc. We describe the assumptions underlying Poisson models more precisely in the following example, and use them to derive the corresponding pmf.

Example 2.21 (Number of earthquakes)**.** We would like to model the number of earthquakes occurring in the San Francisco Bay Area over one year. We decide that the following assumptions are reasonable:

1 For any period of time of length t, if t is small enough, the probability of an earthquake occurring in that period is equal to λt and the probability of more than one earthquake is negligible. λ is a fixed parameter representing the total earthquakes per year that we expect to occur.
2 Each earthquake occurs independently from other earthquakes.

Our aim is to compute the probability that exactly a earthquakes occur during a year, based on these assumptions. We begin by discretizing the year into n intervals of length $1/n$. By Assumption 1, if n is large enough, the probability

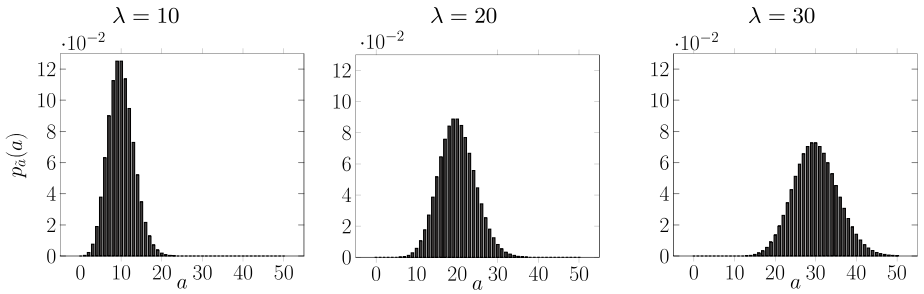

Figure 2.9 Poisson pmf. The graphs show the first fifty entries of the pmf of a Poisson random variable $\tilde{a}$ for $\lambda = 10, 20, 30$.

of an earthquake occurring in each interval is λ/n, and the probability of more earthquakes occurring is zero. By Assumption 2, earthquakes occur independently in the different intervals. Consequently, the total number of earthquakes occurring in the n slots is a binomial random variable with parameters n and λ/n. To ensure that the intervals are small enough that Assumption 1 holds, we take the limit $n \to \infty$. This yields

$$\begin{aligned}
&\mathrm{P}\left(a \text{ earthquakes during the whole year }\right) && (2.53)\\
&= \lim_{n\to\infty} \mathrm{P}\left(a \text{ earthquakes in } n \text{ small intervals}\right) && (2.54)\\
&= \lim_{n\to\infty} \binom{n}{a} \left(\frac{\lambda}{n}\right)^a \left(1-\frac{\lambda}{n}\right)^{(n-a)} && (2.55)\\
&= \lim_{n\to\infty} \frac{n!\,\lambda^a}{a!\,(n-a)!\,(n-\lambda)^a} \left(1-\frac{\lambda}{n}\right)^n && (2.56)\\
&= \frac{\lambda^a\, e^{-\lambda}}{a!}. && (2.57)
\end{aligned}$$

The last step requires some calculus. By definition of Euler's constant e,

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{2.58}$$

Also, for any fixed constants c_1 and c_2

$$\lim_{n\to\infty} \frac{n-c_1}{n-c_2} = 1, \tag{2.59}$$

which implies

$$\lim_{n\to\infty} \frac{n!}{(n-a)!\,(n-\lambda)^a} = \frac{n}{n-\lambda} \cdot \frac{n-1}{n-\lambda} \cdots \frac{n-a+1}{n-\lambda} = 1. \tag{2.60}$$

The pmf derived in Example 2.21 is the pmf of the Poisson distribution. Figure 2.9 shows examples of this pmf for different values of the rate parameter λ.

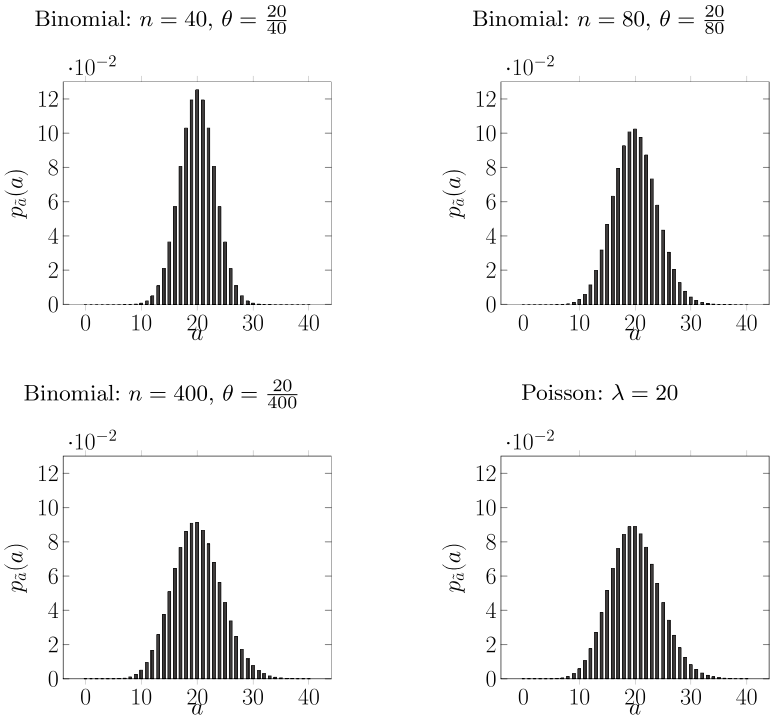

Figure 2.10 Convergence of the binomial distribution to a Poisson distribution. As n increases the binomial pmf with parameters n and $p = \lambda/n$ converges to a Poisson pmf of parameter λ.

Definition 2.22 (Poisson distribution)**.** *The pmf of a Poisson random variable $\tilde{a}$ with parameter $\lambda > 0$ is*

$$p_{\tilde{a}}(a) = \frac{\lambda^a e^{-\lambda}}{a!}, \quad a = 0, 1, 2, \ldots \tag{2.61}$$

In our derivation of the Poisson pmf, 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 random variable with parameter λ. This is an example of *convergence in distribution*, where the pmf of a random variable dependent on n converges to a certain fixed shape as $n \to \infty$. Figure 2.10 illustrates this phenomenon numerically; the convergence is quite fast. The most famous example of convergence in distribution is the central limit theorem, which we discuss in Section 9.7.

2.4 Maximum-Likelihood Estimation

The previous section describes several popular discrete parametric distributions. Each distribution is characterized by a pmf that depends on one or more parameters. In order to model data using these distributions, we need a procedure to

compute the parameters based on the data. This is known as *fitting* the parametric model to the data. Let $p_\theta : A \to [0, 1]$ be the pmf of a discrete random variable with range A that depends on a vector of parameters θ. The set of possible parameter values is denoted by S. For each value of θ in S, p_θ is nonnegative and sums to one, and is therefore a valid pmf. Our goal is to select the value of θ that is most consistent with the available data.

For any $a \in A$, $p_\theta(a)$ encodes the probability of observing a according to the model. Assume that we only have access to one data point x. Then a reasonable choice for θ is the one that maximizes the probability of observing that particular data point, $p_\theta(x)$. Note the change of perspective: we are suddenly thinking of $p_\theta(x)$ *as a function of* θ. This function is known as the *likelihood* of the parametric model. Maximizing the likelihood to fit the parameters of a parametric model is known as *maximum-likelihood estimation*.

Of course, we won't ever fit a model with a single data point. In order to extend the concept of likelihood to multiple data, we need to make some assumptions. A reasonable choice is to interpret the data are realizations of independent, identically-distributed (i.i.d.) random variables. This means that the value of each data point is generated according to the same distribution, and is independent of the rest of the data. This is the same assumption that we used to study the empirical-probability estimator in Example 2.18.

Definition 2.23 (Independent identically distributed random variables)**.** *Let* $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ *be discrete random variables defined on the same probability space. The random variables are identically distributed if they have the same pmf, so that for each element* a *of their range* A*:*

$$\mathrm{P}(\tilde{a}_1 = a) = \mathrm{P}(\tilde{a}_2 = a) = \cdots = \mathrm{P}(\tilde{a}_n = a). \tag{2.62}$$

The random variables are independent, if for any choice of a_1, a_2, ..., a_n *belonging to* A*, the events* $\tilde{a}_1 = a_1$, $\tilde{a}_2 = a_2$, ..., $\tilde{a}_n = a_n$ *are mutually independent.*

Let x_1, x_2, ..., x_n be the available data. If we model them as realizations from i.i.d. discrete random variables with a parametric pmf p_θ, the probability of observing precisely those values equals

$$\mathrm{P}(\tilde{a}_1 = x_1, \tilde{a}_2 = x_2, \ldots, \tilde{a}_n = x_n) = \mathrm{P}(\tilde{a}_1 = x_1)\mathrm{P}(\tilde{a}_2 = x_2) \cdots \mathrm{P}(\tilde{a}_n = x_n) \tag{2.63}$$

$$= \prod_{i=1}^{n} p_\theta(x_i). \tag{2.64}$$

If we fix the data, we can interpret this probability as a function of θ to obtain the likelihood of the parametric model. Notice that the likelihood is a product of probabilities, which are numbers between zero and one. As a result, it can become extremely small when n is large. To avoid numerical instabilities, we often consider the logarithm of the likelihood, or *log-likelihood*, instead.

Definition 2.24 (Likelihood function under i.i.d assumptions)**.** *Let* $p_\theta : A \to [0, 1]$ *be a parametric pmf model dependent on a parameter vector* θ*, and* $X :=$

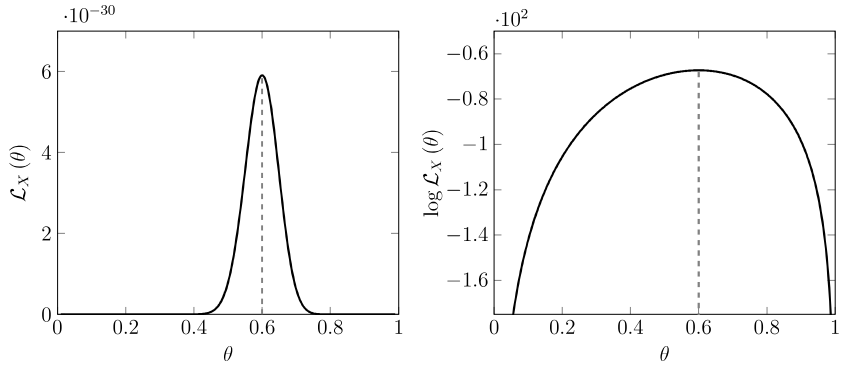

Figure 2.11 Likelihood and log-likelihood functions of a Bernoulli parametric model. Likelihood (left) and log-likelihood (right) functions of a Bernoulli parametric model for a dataset where there are 60 data points equal to 1 and 40 data points equal to 0. The values of the likelihood are extremely small, which can lead to numerical instabilities. The maximum of both functions occurs at 0.6, which is the maximum-likelihood estimate.

$\{x_1, x_2, \ldots, x_n\}$ a dataset with values in A. The likelihood of the model given these data under i.i.d. assumptions is equal to

$$\mathcal{L}_X(\theta) := \prod_{i=1}^{n} p_\theta\left(x_i\right). \tag{2.65}$$

The log-likelihood function is equal to the logarithm of the likelihood function,

$$\log \mathcal{L}_X(\theta) = \sum_{i=1}^{n} \log p_\theta\left(x_i\right). \tag{2.66}$$

Maximum-likelihood estimation selects the parameters that maximize the likelihood or the log-likelihood, and therefore the probability of the observed data according to the parametric model under i.i.d. assumptions. Maximizing the likelihood or the log-likelihood is equivalent because the logarithm is a monotone function.

Definition 2.25 (Maximum-likelihood estimator)**.** *Let $p_\theta : A \to [0, 1]$ be a parametric pmf model dependent on a parameter vector θ, $X := \{x_1, x_2, \ldots, x_n\}$ a dataset with values in A, and S the set of parameters for which p_θ is a valid pmf. The maximum-likelihood estimate of θ is*

$$\theta_{\text{ML}} := \arg\max_{\theta \in S} \mathcal{L}_X(\theta) \tag{2.67}$$

$$= \arg\max_{\theta \in S} \log \mathcal{L}_X(\theta). \tag{2.68}$$

Example 2.26 (Maximum-likelihood estimator for the Bernoulli distribution)**.**

Let $X := \{x_1, \ldots, x_n\}$ be n data points equal to zero or one, representing the occurrence of some event of interest. Assuming that the data are i.i.d., we decide to fit a Bernoulli model with parameter θ (in this case there is only one parameter). The likelihood function is equal to

$$\mathcal{L}_X(\theta) = \prod_{i=1}^{n} p_\theta(x_i) \tag{2.69}$$

$$= \theta^{n_1}(1-\theta)^{n_0}, \tag{2.70}$$

where n_0 and n_1 are the number of observations equal to zero and one, respectively. The log-likelihood function equals

$$\log \mathcal{L}_X(\theta) = n_1 \log \theta + n_0 \log(1-\theta). \tag{2.71}$$

Figure 2.11 shows the likelihood and log-likelihood functions for $n_0 = 40$ and $n_1 = 60$. The maximum-likelihood estimator of the parameter θ is

$$\theta_{\text{ML}} = \arg\max_\theta \log \mathcal{L}_X(\theta) \tag{2.72}$$

$$= \arg\max_\theta n_1 \log \theta + n_0 \log(1-\theta). \tag{2.73}$$

The derivative and second derivative of the log-likelihood function equal

$$\frac{\mathrm{d} \log \mathcal{L}_X(\theta)}{\mathrm{d}\theta} = \frac{n_1}{\theta} - \frac{n_0}{1-\theta}, \tag{2.74}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_X(\theta)}{\mathrm{d}\theta^2} = -\frac{n_1}{\theta^2} - \frac{n_0}{(1-\theta)^2} < 0 \quad \text{for all } \theta \in [0,1]. \tag{2.75}$$

The function is concave, as the second derivative is negative. This is good news, because it means that there cannot be different local maxima. The maximum is at the point where the first derivative equals zero, namely

$$\theta_{\text{ML}} = \frac{n_1}{n_0 + n_1}. \tag{2.76}$$

The estimator is the fraction of samples that equal one. This is equivalent to estimating θ using the empirical probability of observing a one.

We are now ready to fit the parametric model we derived in Example 2.19 to the data in Example 2.14.

Example 2.27 (Fitting the parametric model for free-throw streaks)**.** Let $X := \{x_1, x_2, \ldots, x_n\}$ contain the lengths of the 377 observed free-throw streaks (here,

$n := 377$). Since $p_\theta(a) = \theta^a(1-\theta)$, the log-likelihood equals

$$\log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i) \tag{2.77}$$

$$= \sum_{i=1}^{n} \log\left(\theta^{x_i}(1-\theta)\right) \tag{2.78}$$

$$= \sum_{i=1}^{n} \left(x_i \log\theta + \log(1-\theta)\right) \tag{2.79}$$

$$= \left(\sum_{i=1}^{n} x_i\right) \log\theta + n\log(1-\theta) \tag{2.80}$$

$$= n_{\text{made}} \log\theta + n_{\text{missed}} \log(1-\theta), \tag{2.81}$$

where $n_{\text{made}} = \sum_{i=1}^{n} x_i$ is the number of made free throws and $n_{\text{missed}} = n$ is the number of missed free throws. The log-likelihood is exactly the same as the one from the Bernoulli model in Example 2.26. This makes sense: if we focus on individual free throws instead of on streaks of made free throws, our assumptions imply that the data are realizations of i.i.d. Bernoulli random variables with parameter θ. By the same argument as in Example 2.26, the maximum-likelihood estimator for θ equals the fraction of made free throws,

$$\theta_{\text{ML}} = \frac{n_{\text{made}}}{n_{\text{missed}} + n_{\text{made}}} \tag{2.82}$$

$$= 0.875. \tag{2.83}$$

The corresponding pmf is shown in Figure 2.12, which also shows the fits corresponding to two other values of θ. The maximum-likelihood estimate achieves a much better approximation to the empirical pmf of the observed data. This is not a surprise, since maximum-likelihood estimation yields the parametric pmf according to which the data are most likely.

It is instructive to study the likelihood function and the maximum-likelihood estimator when our modeling assumptions hold completely (i.e. the data are i.i.d. samples from a known parametric model with unknown parameters). To this end, we generate synthetic simulated datasets of i.i.d. samples from the geometric model derived in Example 2.27 fixing the parameter to $\theta_{\text{true}} := 0.875$. The log-likelihood functions and maximum-likelihood estimates are shown in Figure 2.13. The log-likelihood changes depending on the observed realizations, but its maximum tends to be close to the true value of the parameter (the maximum-likelihood estimates are 0.873, 0.879, and 0.867). Compared to the empirical pmf, the parametric estimate achieves a much better approximation to the true pmf. Of course, the comparison is not fair: are assuming that we know exactly how the data were generated, which is never the case in practice. The next section provides a more realistic comparison of both estimators.

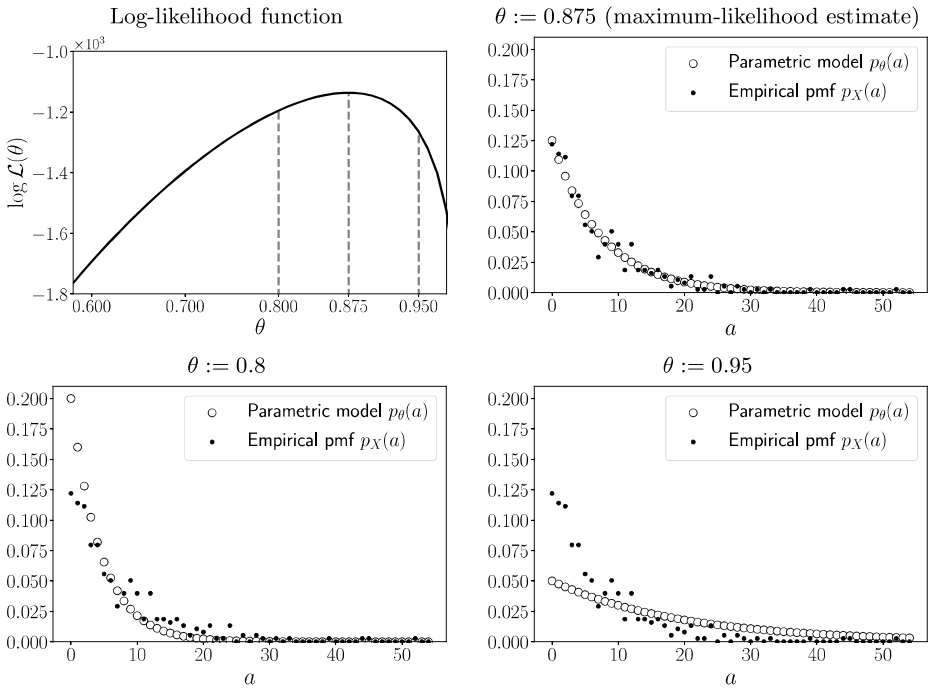

Figure 2.12 Parametric model for consecutive made free throws. The graphs at the top show the log-likelihood function of the parametric model derived in Example 2.19 (top left) and the corresponding fit to the data in Example 2.14 obtained via maximum-likelihood estimation (top right). The two bottom graphs show the fits corresponding to two other choices of the parameter θ, which produce much worse fits.

2.5 Comparing Parametric And Nonparametric Models

We have described two alternative approaches to estimate pmfs of discrete random variables from data. The first strategy is to approximate the pmf directly using empirical probabilities, as explained in Section 2.2. This is known as *nonparametric* estimation, in contrast to the second approach, where we fit a predetermined parametric distribution to the data, as explained in Section 2.4. In this section, we discuss the advantages and disadvantages of the two strategies.

Let us focus on the free-throw dataset from Example 2.14. We could be tempted to evaluate the pmf models based on how well they fit the data. To assess the fit of a given pmf estimate p_{est}, we compare the estimated probability $p_{\text{est}}(\ell)$ to the corresponding empirical probability of observing a streak of length ℓ. According to this metric, the empirical-pmf estimator is amazing; it produces a perfect fit! Of course, this is too good to be true. As we already mentioned in Section 2.2, the empirical pmf seems to capture some spurious patterns that are probably caused

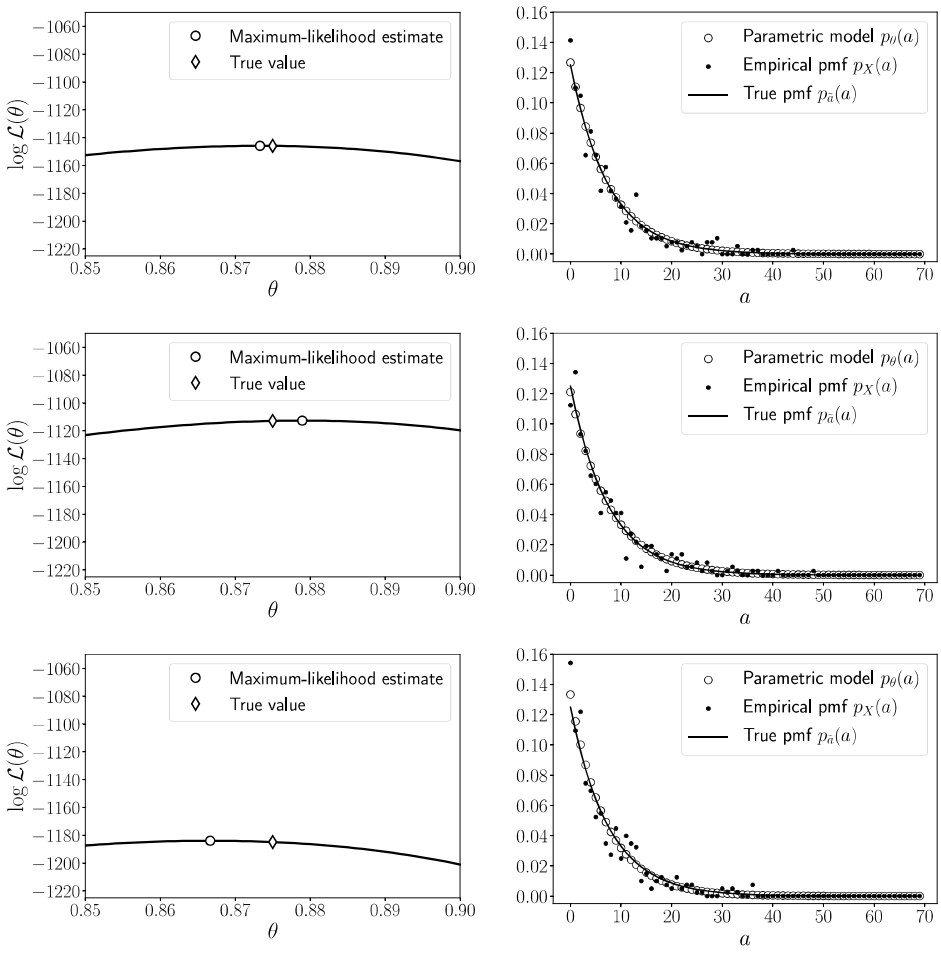

Figure 2.13 Maximum-likelihood estimation applied to simulated data. Log-likelihood function (left column) and corresponding maximum-likelihood fit (right column) of the parametric model derived in Example 2.19 applied to 3,015 i.i.d. samples simulated according to the model with a $\theta_{\text{true}} := 0.875$ (referred to as *true value* on the graphs). On the right we also show the true pmf (black curve).

by the limited number of data. The problem here is that we are using *the same data to build the model and evaluate it*. This is like testing a student who has seen the answers already; they can mindlessly repeat them to you without having learned anything. Instead, we need to evaluate models by checking whether they *generalize* to *held-out data*, which have *not* been used to fit them.

In order to perform rigorous evaluation of a statistical estimator, we select a subset of the available data beforehand and reserve it exclusively for evaluation.

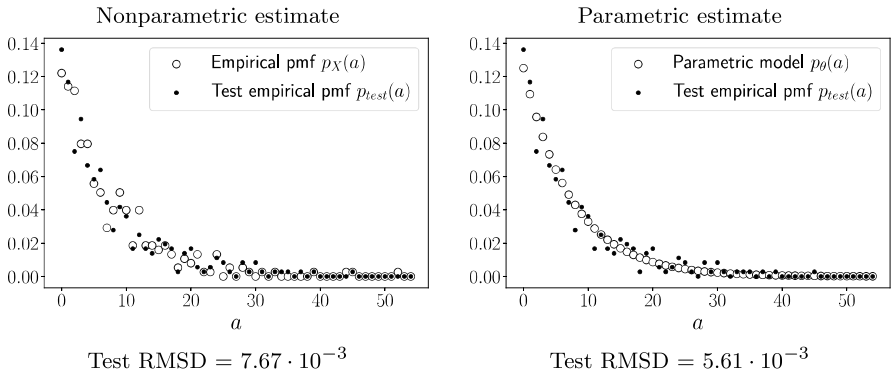

Figure 2.14 Parametric vs. nonparametric model for free-throw streaks. Probability of Kevin Durant making a certain number of consecutive free throws before missing, obtained by fitting a nonparametric empirical-pmf model (left, see Example 2.14) and a parametric geometric model (right, see Example 2.19). Both models are fit on a training set consisting of the first 3,015 free throws in Durant's NBA career. In order to evaluate the pmf estimates, we compare them to the empirical pmf of a test set corresponding to the following 3,015 free throws. The parametric model approximates the test pmf better in terms of RMSD.

We call this the *test set*. The data used to fit the model is known as the *training set* (*training* a model means fitting it to data). In Example 2.14, we use the first 3,015 free throws in Durant's NBA career as a training set to fit a nonparametric empirical-pmf model. In Example 2.19, we use a parametric pmf based on the geometric distribution to model the same data. We now use the following 3,015 free throws as a test set to evaluate the generalization ability of both models. As a quantitative metric, we compute the root mean square deviation (RMSD) between each estimated pmf p_{est} and the empirical pmf of the test data p_{test}:

$$\text{RMSD}(p_{\text{est}}) := \sqrt{\frac{1}{L}\sum_{\ell=0}^{L}(p_{\text{est}}(\ell) - p_{\text{test}}(\ell))^2}, \tag{2.84}$$

where L is the maximum streak length, which we set equal to 55. The nonparametric model has a test RMSD of $7.67 \cdot 10^{-3}$. The parametric model has a smaller test RMSD error: $5.61 \cdot 10^{-3}$. Figure 2.14 shows both estimated pmfs superimposed on the empirical pmf of the test data.

As we had suspected, the fluctuations in the nonparametric empirical-pmf estimate (e.g. a streak of length 7 having lower probability than a streak of length 12) are not reproduced in the test data. This suggests that they are capturing *noise*, a term we use to describe unpredictable structure in the data which does not carry any useful information. In contrast, the meaningful information that we are trying to extract from the data is called the *signal*. Estimators that fit the noise in the training data are said to *overfit*, because they are approximating

the training set too closely. A training error that is much better than the test error is a symptom of overfitting. Here, the training error of the nonparametric empirical-pmf estimate is zero. In contrast, the training error of the parametric model is $5.46 \cdot 10^{-3}$, which is very similar to the test error, indicating that it generalizes well.

As a rule of thumb, complicated models with a large number of parameters are more prone to overfitting than simpler models, because they have more flexibility to fit noisy fluctuations in the training data. Even though the empirical pmf is called a *nonparametric* estimator, it also requires estimating some *parameters* from the data, namely each entry of the pmf. In the case of the free-throw data, the number of parameters to be estimated is 55.* In contrast, the parametric model only requires estimating one parameter! This constrains the shape of the pmf, so that it cannot overfit the noise in the training data.

Parametric models are not always superior to nonparametric models. The assumptions underlying parametric models never hold exactly. This is not very problematic when the training data are limited. In fact, as illustrated by the free-throw data, it can be a crucial advantage: the assumptions in Example 2.19 yield a model that is simple enough to avoid overfitting. However, if the training data are more plentiful, inaccurate assumptions may result in an overly simplistic model, which is not able to capture meaningful complex structure. This is known as *underfitting*. The following example evaluates a nonparametric and parametric model fit with training sets of different sizes. As the number of available training data increases, the relative performance of the nonparametric model improves, eventually surpassing the performance of the parametric model, as shown in Figure 2.16.

Example 2.28 (Call center)**.** In this example we analyze Dataset 3, which consists of telephone data recorded at the call center of an anonymous bank in Israel in 1999. Our goal is to model the distribution of the number of calls that arrive between 6 am and 7 am on weekdays. We model the number of calls as a discrete random variable and estimate its pmf using the empirical pmf and a parametric Poisson model. The maximum number of observed calls is 24, so the nonparametric empirical-pmf estimator actually consists of 24 parameters. In contrast, the Poisson model has a single parameter, which we fit via maximum-likelihood estimation.

Let X denote the training data, which correspond to telephone-call counts extracted from the training set. The likelihood is

$$\mathcal{L}_X(\lambda) = \prod_{i=1}^{n} p_{\tilde{x}}(x_i) \tag{2.85}$$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \tag{2.86}$$

*The maximum possible streak length to 55, so the possible lengths go from zero to 55. From these 56 numbers, only 55 need to be estimated from the data because the pmf must sum to one.

so the log-likelihood equals

$$\log \mathcal{L}_X(\lambda) = \sum_{i=1}^{n} \left(x_i \log \lambda - \lambda - \log(x_i!) \right). \tag{2.87}$$

The derivative and second derivative of the log-likelihood are

$$\frac{\mathrm{d} \log \mathcal{L}_X(\lambda)}{\mathrm{d}\lambda} = \sum_{i=1}^{n} \frac{x_i}{\lambda} - 1, \tag{2.88}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_X(\lambda)}{\mathrm{d}\lambda^2} = -\sum_{i=1}^{n} \frac{x_i}{\lambda^2} < 0. \tag{2.89}$$

The function is concave, as the second derivative is negative. The maximum is consequently at the point where the first derivative equals zero, namely

$$\lambda_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{2.90}$$

which is just the average number of calls. This is consistent with the interpretation of the parameter as a *rate* in our derivation of the Poisson pmf (see Example 2.21).

In order to evaluate the nonparametric and parametric pmf estimators, we build training sets of different sizes. The training sets consist of the number of calls received during weekdays between 6 am and 7 am in a period from month 1 (January) to month m, where we vary m from one (just January) to six (January to June). We use the remaining months (July to December) as a test set. Figure 2.15 shows the estimated pmfs superimposed to the empirical pmf of the test set. Figure 2.16 shows the test RMSD, as defined in (2.84), of the nonparametric and parametric estimates. When $m = 1$, the number of data is around 20, which is very small compared to the number of parameters of the empirical pmf. As a result, the nonparametric model overfits the training data, resulting in a test RMSD that is double that of the Poisson model. As we increase the number of months used for training, the nonparametric model improves substantially, eventually overtaking the parametric model. This is quite typical: parametric models (with reasonable assumptions) tend to perform well when the training data are limited, whereas nonparametric models can be extremely effective if we have enough training data to exploit their flexibility without overfitting.

Let us take a closer look at the pmf estimates in Figure 2.15 to illustrate the advantages and disadvantages of nonparametric and parametric approaches. It turns out that January has an unusual number of days when there are no calls, possibly due to holidays when the bank is closed. When the training set only contains January (top row in Figure 2.15) the empirical pmf consequently assigns a high probability to zero. This results in a large test error, because the fraction of data with zero calls in the test set is much lower (perhaps because there are less holidays between July and December). As more data are used in the estimate, the empirical pmf assigns less and less probability to zero, because the additional training months contain less days with zero calls.

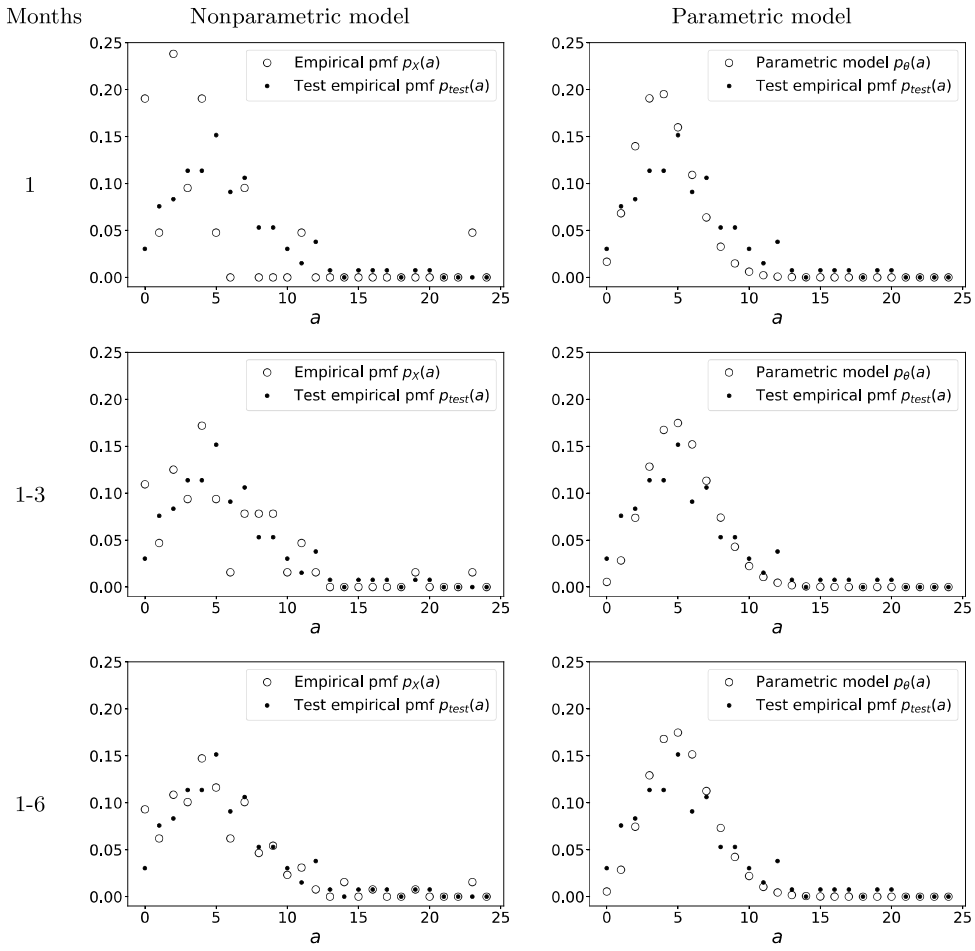

Figure 2.15 Parametric vs. nonparametric model for calls arriving at a bank. The left column shows a nonparametric estimate of the pmf corresponding to the empirical pmf of the training data. The right column shows a parametric estimator obtained via maximum-likelihood estimation based on a Poisson model. Each row corresponds to a training set containing a different number of months: January (first row), January to March (second row), and January to June (third row). The estimates are compared to the empirical pmf of the test data, which contains the months from July to December.

In contrast, the parametric model does not overfit the unusual number of days with zero calls even when we only use January to train the model. The reason is that the shape of the Poisson pmf cannot assign a large probability to zero and at the same time approximate the rest of the observed empirical probabilities. In this case, this results in effective generalization to the test data, but it could also have

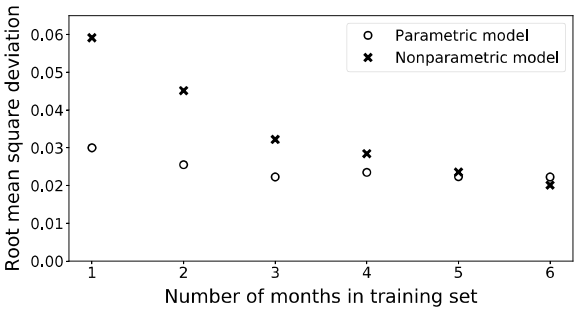

Figure 2.16 Comparison of parametric and nonparametric models for different amounts of data. Test RMSD of the parametric estimator based on a Poisson model and the nonparametric estimator based on the empirical pmf. The training sets contain data from month 1 (January) to month m, where we vary m from one (just January) to six (January to June). The test set contains data from July to December. The parametric approach is clearly superior when the number of training data is small. However, the nonparametric estimator improves dramatically as we increase the training data, eventually overtaking the parametric estimator.

hurt it. In fact, in Figure 2.15 we can see that the Poisson model systematically underestimates the probability of zero calls. This would have been even more apparent if there had been a month in the test data with a lot of holidays (for example, January 2000). The problem is that the observed data violates the assumptions underlying the Poisson model, which does not take into account holidays. To address this underfitting issue, we cannot just use more training data; the Poisson model is too simple to account for the additional structure. Instead, we need to use a more complex model.