1

Probability

Overview

Who will win the next presidential election? What will be the price of Tesla stock tomorrow? Will the New York Knicks win the NBA championship next season? There is no definite answer to these questions, because they are associated to *uncertain* phenomena with different possible outcomes. We can describe such phenomena by determining the *probability* of each possible outcome, which quantifies how likely it is for that particular outcome to occur. This simple idea is a fundamental underpinning of statistics and data science. In Section 1.1 we provide an intuitive definition of probability and describe its main properties. Building upon this intuition, Section 1.2 introduces the mathematical framework of probability spaces. Section 1.3 defines conditional probability, which allows us to update probabilities when additional information is revealed. In Section 1.4 we explain how to estimate probabilities from data. Section 1.5 and 1.6 introduce the key concepts of independence and conditional independence, respectively. Finally, Section 1.7 discusses how to compute probabilities using computer simulations.

1.1 Intuitive Properties Of Probability

In order to define probabilities associated to a certain uncertain phenomenon, we interpret it as an *experiment* with multiple possible outcomes. The set of all possible outcomes associated to a particular phenomenon is called the *sample space*, usually denoted by Ω. As the following examples show, the sample space can be discrete or continuous.

Example 1.1 (Die roll: Sample space)**.** If we roll a six-sided die once, there are six possible results that are mutually exclusive (the die cannot land on two numbers at the same time). In order to describe the die roll probabilistically, we define the outcome as the number that the die lands on. The sample space is the finite set $\Omega := \{1, 2, 3, 4, 5, 6\}$.

Example 1.2 (Rolling a die until it lands on a six: Sample space)**.** Imagine that we roll a six-sided die repeatedly until it lands on a six. Modeling the outcomes for this situation is not as straightforward as in the previous example. If we are just interested in the number of rolls that occur, we can set the outcome to equal that number. In that case, the sample space is the set of natural numbers $\Omega_1 := \mathbb{N}$.

If we are interested in the actual values of the rolls before the final six occurs, then we can set the outcome to equal the sequence of roll results ending in six (e.g. if we roll a four, then a one and finally a six, the outcome is $4 \to 1 \to 6$). The sample space Ω_2 is then the (infinite) set of all such sequences. Either way, the sample space is discrete, but countably infinite.

Example 1.3 (Weather in New York tomorrow: Sample space)**.** If we want to model the weather in New York tomorrow, then there are a lot of choices to make! To simplify matters, let us assume that we are only interested in the temperature in Washington Square Park at noon. We define the outcome to be that temperature, represented as a real number, so the sample space is the real line $\Omega := \mathbb{R}$.* In this case, the sample space is continuous; the number of possible outcomes is uncountable.

Once we have defined the sample space, we describe our phenomenon of interest by determining how likely sets of outcomes are. We call these sets of outcomes *events*. Events can consist of several outcomes, a single outcome, the whole sample space, or no outcomes at all (although this is not a very interesting event). We say that an event occurs if the outcome of the experiment belongs to it, as illustrated by the following examples.

Example 1.4 (Die roll: Events)**.** Possible events associated to the sample space in Example 1.1 include rolling a five $A := \{5\}$, rolling an even number $B := \{2, 4, 6\}$, or rolling any number $C := \{1, 2, 3, 4, 5, 6\}$. If the roll is a four, then events B and C occur, but A does not.

Example 1.5 (Rolling a die until it lands on a six: Events)**.** In Example 1.2 the structure of the events depend on the choice of sample space. For example, the event *Rolling twice to obtain a six* is an event with a single outcome $\{2\}$ if the sample space is Ω_1, but contains five outcomes ($1 \to 6$, $2 \to 6$, $3 \to 6$, $4 \to 6$ and $5 \to 6$) if the sample space is Ω_2.

Example 1.6 (Weather in New York tomorrow)**.** If we choose to just model the temperature in Washington Square Park at noon and fix $\Omega := \mathbb{R}$, then possible events include temperatures above 30 degrees $A := [30, \infty)$, a temperature exactly equal to 35 $B := 35$, or any temperature $C := \mathbb{R}$. If the temperature turns out to be 40 degrees, then A and C occur, but B does not.

In order to quantify how likely an event is, we assign it a number, which we call a *probability*. The key idea behind the concept of probability is to interpret the uncertain phenomenon of interest as an experiment, which *can be repeated over and over*. Of course, this is just an abstraction. The next presidential election will

*Strictly speaking, temperatures cannot be lower than absolute zero, but we use the whole real line for convenience.

happen only once. However, thinking of it as a repeatable experiment enables us to reason about it probabilistically. The probability $\mathrm{P}(A)$ of an event A represents the fraction of times that the event occurs (i.e. the outcome of the experiment belongs to the event) when we repeat the experiment an arbitrarily large number of times:

$$\mathrm{P}(\text{event}) := \frac{\text{times event occurs}}{\text{total repetitions}}. \tag{1.1}$$

Notice that the probability is between zero and one, because the number of times the event occurs ranges between zero and the total number of repetitions. This is a non-rigorous, intuitive definition of probability. We provide a formal definition in Section 1.2.

When determining the probabilities associated to a sample space, we do not need to assign a probability to every possible event. In fact, when the sample space is continuous, it may not be possible to assign a probability to every possible subset in a consistent manner. This is a mathematical issue with no practical implications, which has an interesting connection to the Banach-Tarski paradox. We refer the interested reader to any textbook on measure theory for more details. However, we definitely want to assign probabilities to *some* events. In the remainder of this section, we discuss what these events should be, and derive their associated probabilities from our informal definition of probability.

1.1.1 Probability Of The Sample Space

We should definitely assign a probability to the event that *anything at all* happens. The event that contains all possible outcomes is the sample space Ω itself. Any time we repeat the experiment, we obtain an outcome that must be in Ω, so by our informal definition

$$\mathrm{P}(\Omega) = \frac{\text{times } \Omega \text{ occurs}}{\text{total repetitions}} \tag{1.2}$$

$$= \frac{\text{total repetitions}}{\text{total repetitions}} \tag{1.3}$$

$$= 1. \tag{1.4}$$

Therefore the probability assigned to the sample space should always equal one.

1.1.2 Probability Of Unions And Intersections Of Events

If we assign a probability to two events, we should also assign probabilities to their union and intersection. The union of two events is the event that *either* of them occurs. The intersection of two events is the event that *both* of them occur simultaneously. We begin by considering *disjoint* events, which are events that do not have any outcomes in common, so their intersection is empty. In Example 1.4 the events A and B are disjoint because no outcome is in both sets, but A and

C are not disjoint because 5 belongs to both of them. If two events D_1 and D_2 are disjoint, our informal definition of probability implies

$$\begin{aligned} \mathrm{P}\left(D_1 \cup D_2\right) &= \frac{\text{times } D_1 \text{ or } D_2 \text{ occur}}{\text{total repetitions}} && (1.5)\\ &= \frac{\text{times } D_1 \text{ occurs} + \text{times } D_2 \text{ occurs}}{\text{total repetitions}} && (1.6)\\ &= \frac{\text{times } D_1 \text{ occurs}}{\text{total repetitions}} + \frac{\text{times } D_2 \text{ occurs}}{\text{total repetitions}} && (1.7)\\ &= \mathrm{P}\left(D_1\right) + \mathrm{P}\left(D_2\right). && (1.8) \end{aligned}$$

Therefore, the probability of the union of disjoint events should equal the sum of their individual probabilities.

If two events E_1 and E_2 are not disjoint, then their intersection is not empty. As a result, according to our informal definition, the probability of their union equals

$$\begin{aligned} \mathrm{P}\left(E_1 \cup E_2\right) &= \frac{\text{times } E_1 \text{ or } E_2 \text{ occur}}{\text{total repetitions}} && (1.9)\\ &= \frac{\text{times } E_1 \text{ occurs} + \text{times } E_2 \text{ occurs} - \text{times } E_1 \text{ and } E_2 \text{ occur}}{\text{total repetitions}} \\ &= \frac{\text{times } E_1 \text{ occurs}}{\text{total repetitions}} + \frac{\text{times } E_2 \text{ occurs}}{\text{total repetitions}} - \frac{\text{times } E_1 \text{ and } E_2 \text{ occur}}{\text{total repetitions}} \\ &= \mathrm{P}\left(E_1\right) + \mathrm{P}\left(E_2\right) - \mathrm{P}\left(E_1 \cap E_2\right). && (1.10) \end{aligned}$$

This makes sense: we need to subtract the probability of the intersection in order not to count its outcomes twice.

From (1.10) we obtain an expression for the probability of the intersection of two events:

$$\mathrm{P}\left(E_1 \cap E_2\right) = \mathrm{P}\left(E_1\right) + \mathrm{P}\left(E_2\right) - \mathrm{P}\left(E_1 \cup E_2\right). \tag{1.11}$$

1.1.3 Probability Of The Complement Of An Event

If we assign a probability to an event, we should also assign a probability to its complement, i.e. to the event *not* occurring. Mathematically, the complement is the set of all the outcomes that are *not* in the event. In Example 1.4 the complement of A is $\{1, 2, 3, 4, 6\}$ and the complement of B is $\{1, 3, 5\}$. For any event E, the union of E and its complement E^c is equal to the whole sample space (every outcome is either in E or in its complement). In addition, E and E^c are disjoint by definition (no outcome can be in both events). By our informal definition of probability, this implies

$$\begin{aligned} \mathrm{P}\left(E\right) + \mathrm{P}\left(E^c\right) &= \mathrm{P}\left(E \cup E^c\right) && (1.12)\\ &= \mathrm{P}(\Omega) && (1.13)\\ &= 1, && (1.14) \end{aligned}$$

so to compute the probability of the complement of E, we just need to subtract its probability from one, $\mathrm{P}\left(E^c\right) = 1 - \mathrm{P}\left(E\right)$. Intuitively, if an event is very likely (probability close to one), its complement should be unlikely (probability close to zero), and vice versa.

1.2 Mathematical Definition Of Probability

In this section we present the mathematical framework of probability spaces, which allows us to characterize uncertain phenomena using probabilities. A probability space has three elements. First, a sample space containing the mutually exclusive outcomes associated to the phenomenon. Second, a collection containing the events that are assigned probabilities. Third, a probability measure, which is a function that assigns a probability to each event in the collection. In order for a collection to be valid, it needs to satisfy the conditions in the following definition.

Definition 1.7 (Collection of events)**.** *When defining a probability space based on a sample space Ω, we assign probabilities to a collection of events (a set of subsets of Ω) denoted by $\mathcal{C}$ such that:*

1 *If an event $A \in \mathcal{C}$ then $A^c \in \mathcal{C}$.*
2 *If the events $A, B \in \mathcal{C}$, then $A \cup B \in \mathcal{C}$. This also holds for infinite sequences; if $A_1, A_2, A_3, \ldots \in \mathcal{C}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{C}$.*
3 *$\Omega \in \mathcal{C}$.*

A collection satisfying Definition 1.7 is called a σ-algebra in mathematical jargon, which may sound somewhat intimidating. However, the definition just implements the intuitive properties discussed in Section 1.1. If we assign probabilities to certain events, then *we should also assign probabilities to their complements, unions, and intersections*. Note that although the definition does not mention intersections explicitly, it implies that intersections of events in $\mathcal{C}$ also belong to $\mathcal{C}$. This follows from the fact that $A \cap B = (A^c \cup B^c)^c$ (a simple consequence of De Morgan's laws) combined with Conditions 1 and 2. The empty set $\emptyset$ always belongs to a valid collection because it is the complement of Ω. The simplest possible collection satisfying the conditions is $\{\Omega, \emptyset\}$, but this is not a very interesting collection; usually we want to consider more events.

Example 1.8 (Collections of events for a single six-sided die roll)**.** A natural collection of events for the six-sided die example is the *power set* of the sample space $\Omega := \{1, 2, 3, 4, 5, 6\}$, which is the set of all 2^6 subsets of Ω. However, other choices are possible. For example, we may want to consider the smallest possible collection containing the event $A := \{5\}$. In that case, the collection must also contain $A^c = \{1, 2, 3, 4, 6\}$ by Condition 1, Ω by Condition 3, and the empty set $\emptyset$ by Conditions 1 and 3. This is enough. You can check that the collection $\{\emptyset, A, A^c, \Omega\}$ is a valid collection for any choice of the event A.

Once we have defined a sample space and a corresponding collection of events,

the final ingredient to define a probability space is a probability measure that assigns probabilities to the events in the collection. The probability measure must satisfy the following axioms, which encode the intuitive properties derived in Section 1.1.

Definition 1.9 (Probability measure)**.** *Given a sample space Ω, let $\mathcal{C}$ be a collection of events satisfying the conditions in Definition 1.7. A probability measure* P *is a function that maps events in $\mathcal{C}$ to a number between 0 and 1, and satisfies the following axioms:*

1 $\mathrm{P}(A) \geq 0$ *for any event* $A \in \mathcal{C}$.
2 $\mathrm{P}(\Omega) = 1$.
3 If the events $A_1, A_2, \ldots, A_n \in \mathcal{C}$ *are disjoint (i.e.* $A_i \cap A_j = \emptyset$ *for* $i \neq j$*) then*

$$\mathrm{P}\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}\left(A_i\right). \tag{1.15}$$

Similarly, for a countably infinite sequence of disjoint events $A_1, A_2, \ldots \in \mathcal{C}$

$$\mathrm{P}\left(\lim_{n\to\infty} \cup_{i=1}^{n} A_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}\left(A_i\right). \tag{1.16}$$

Axiom 1 ensures that we cannot have negative probabilities, which would contradict the intuitive definition (1.1). Axiom 2 assigns a probability of one to the whole sample space. Axiom 3 determines that the probability of the union of disjoint events must equal the sum of their individual probabilities. Axiom 3 implies the identity relating the probability of the union and intersection of non-disjoint events, which we derived in Section 1.1.2.

Lemma 1.10 (Probability of unions and intersections of events)**.** *For any probability measure* P *satisfying the axioms in Definition 1.9, and any events A and B in the corresponding collection of events,*

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{1.17}$$

Proof First we decompose A into the union of $A \cap B$ and $A \cap B^c$, which are disjoint events, so that by Axiom 3 in Definition 1.9,

$$\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c). \tag{1.18}$$

Similarly,

$$\mathrm{P}(B) = \mathrm{P}(A \cap B) + \mathrm{P}(A^c \cap B). \tag{1.19}$$

Finally, we decompose $A \cup B$ into the union of $A \cap B$, $A \cap B^c$ and $A^c \cap B$, which are all disjoint so that

$$\begin{aligned} \mathrm{P}(A \cup B) &= \mathrm{P}(A \cap B^c) + \mathrm{P}(A^c \cap B) + \mathrm{P}(A \cap B) && (1.20)\\ &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B), && (1.21) \end{aligned}$$

where the last equality follows from (1.18) and (1.19). ■

The formula for the probability of the complement of an event derived in Section 1.1.3 is also a direct consequence of Definition 1.9. The proof follows from the argument in Section 1.1.3.

Lemma 1.11 (Probability of the complement of an event)**.** *For any probability measure* P *satisfying the conditions in Definition 1.9, and any event* A*,*

$$\mathrm{P}(A^c) = 1 - \mathrm{P}(A). \tag{1.22}$$

Another important consequence of Definition 1.9 is that if an event A contains another event B, then the probability of A cannot be smaller than the probability of B.

Lemma 1.12 (Subset of an event)**.** *For any probability measure* P *satisfying the conditions in Definition 1.9, assume there exist two events* A *and* B *in the corresponding collection of events such that* $A \subseteq B$*. Then* $\mathrm{P}(A) \leq \mathrm{P}(B)$*.*

Proof We can express B as the union of two disjoint events $A \cap B$ and $A^c \cap B$. Since $A \subseteq B$, $A \cap B = A$, so that by Axiom 2 in Definition 1.9

$$\mathrm{P}(B) = \mathrm{P}(A) + \mathrm{P}(A^c \cap B) \tag{1.23}$$

$$\geq \mathrm{P}(A), \tag{1.24}$$

because by Axiom 1 $\mathrm{P}(A^c \cap B) \geq 0$. ■

A caveat to Lemma 1.12, is that it is possible for a subset of an event in the collection to *not* belong to the collection, which means that their probability is not defined. Consider for instance the collection $\{\emptyset, A, A^c, \Omega\}$ in Example 1.8 where $A := \{5\}$. Then the event $\{2\}$ is a subset of A^c, but it does not belong to the collection, so there is no probability assigned to it.

Probability measures have similar properties to other measures such as mass, length, area, or volume. For example, the mass of the union of two disjoint objects is the sum of their individual masses. This motivates the use of Venn diagrams to visualize probability spaces. In a Venn diagram, the outcomes in the sample space are represented as points in two dimensions. Events are subsets of points, usually depicted as regions delimited by closed curves. The probability of each event is then equal to the area of the corresponding region. The region representing the sample space must have area one and contain all outcomes. Figure 1.1 shows an example.

Example 1.13 (Simple probability measure)**.** Consider the collection of events $\{\emptyset, A, A^c, \Omega\}$, where A is an arbitrary event. To define a valid probability measure, we just need to assign a probability to A, $\mathrm{P}(A) = \theta$. The probability θ can be any number between zero and one. Once that value is fixed, the probabilities of the remaining events are determined by the conditions in Definition 1.9. By Lemma 1.11, $\mathrm{P}(A^c) = 1 - \theta$. By Axiom 2, $\mathrm{P}(\Omega) = 1$, which in turn implies $\mathrm{P}(\emptyset) = 0$ also by Lemma 1.11.

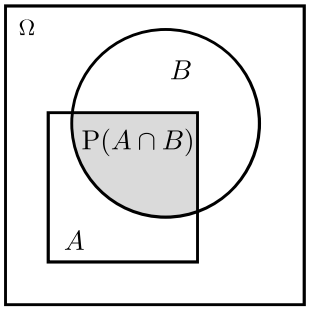

Figure 1.1 Venn diagram of a probability space. Venn diagram representing a probability space. The sample space Ω is the big square that contains everything else. The small square and the circle represent two events A and B, respectively. Their areas are equal to their respective probabilities. The probability of their intersection $A \cap B$ is equal to the area of the intersection between the two shapes, which is shaded.

Example 1.14 (Die roll)**.** As explained in Example 1.8, a reasonable choice for the collection of events associated to the single six-sided die roll is the power set of the sample space $\Omega := \{1, 2, 3, 4, 5, 6\}$. At first it may seem daunting to define the probability measure given that there are 64 events in the collection. However, there is a simple strategy to go about this: divide the sample space into the smallest possible components that belong to the collection, and assign probabilities to these components.

To make this approach more precise, we need to introduce the concept of a *partition*. A partition of the sample space Ω is any collection of disjoint sets $A_1, A_2, \ldots$ that covers Ω, meaning that $\Omega = \cup_i A_i$. In this case, we choose $A_i := \{i\}$, for i in $\{1, 2, 3, 4, 5, 6\}$. These six events clearly cover Ω and are disjoint. We then assign a probability to each of them,

$$P(A_i) = \theta_i, \tag{1.25}$$

where θ_1, θ_2, …, θ_6 are numbers between zero and one. The careful reader may have noticed that these numbers cannot be completely arbitrary. The sum of the probabilities must equal one,

$$\begin{aligned} \sum_{i=1}^{6} \theta_i &= \sum_{i=1}^{6} \mathrm{P}(A_i) && (1.26)\\ &= \mathrm{P}(\cup_{i=1}^{6} A_i) && (1.27)\\ &= P(\Omega) && (1.28)\\ &= 1. && (1.29) \end{aligned}$$

Let us assume that this condition is satisfied. Then we are actually done! We have implicitly defined the probability of any event in the collection. The reason

is that any event can be decomposed as a union of events in the partition, and since these are disjoint we can just add them up to compute the probability of the event. For instance, the probability of the event *the roll is even* ($\{2, 4, 6\}$) equals

$$\begin{aligned} \mathrm{P}\left(\{2, 4, 6\}\right) &= \mathrm{P}\left(\cup_{i \in \{2,4,6\}} A_i\right) && (1.30) \\ &= \sum_{i \in \{2,4,6\}} \mathrm{P}\left(A_i\right) && (1.31) \\ &= \theta_2 + \theta_4 + \theta_6. && (1.32) \end{aligned}$$

For this to work, the partition needs to be granular enough. The events $\{1\}$ and $\{2, 3, 4, 5, 6\}$ are also a partition of Ω, but we cannot express $\{2, 4, 6\}$ as a union of events in this partition.

We have now rigorously defined all the elements of a probability space. This yields the following formal definition.

Definition 1.15 (Probability space)**.** *A probability space is a triple* $(\Omega, \mathcal{C}, \mathrm{P})$ *consisting of:*

- *A sample space* Ω*, which contains all possible outcomes of the experiment.*
- *A collection of events* $\mathcal{C}$*, containing subsets of* Ω*, which satisfies the conditions in Definition 1.7.*
- *A probability measure* P *that assigns probabilities to the events in* $\mathcal{C}$*, which satisfies the axioms in Definition 1.9.*

At this point, you may feel that this probability-space business sounds pretty complicated. We have explained how to choose a sample space, a collection of events, and a probability measure for a very simple example (the single die roll), and even that was not very straightforward. Imagine doing it for more complex phenomena! The good news is that in practice we never construct probability spaces in this way. Instead, we use random variables, which enable us to define probability spaces implicitly, without worrying about the gory mathematical details. We discuss random variables in the following chapters.

1.3 Conditional Probability

1.3.1 Definition

Conditional probability is a crucial concept in probabilistic modeling. It allows us to update models when additional information is revealed. Imagine that we are investigating how the punctuality of flight arrivals is affected by rain. In particular, we want to determine the probability that an airplane is late if it rains. We define a probability space where the collection of events contains the events R (indicating that it rains), the event L (indicating that the airplane is

late), and all their complements, unions and intersections. Then, we estimate the corresponding probability measure from data of past flights:

$$\mathrm{P}\left(L \cap R^c\right) = \frac{2}{20}, \qquad \mathrm{P}\left(L^c \cap R^c\right) = \frac{14}{20}, \tag{1.33}$$

$$\mathrm{P}\left(L \cap R\right) = \frac{3}{20}, \qquad \mathrm{P}\left(L^c \cap R\right) = \frac{1}{20}. \tag{1.34}$$

Since any event in the collection can be represented as a union of some of these four events, the probabilities completely define the probability measure of the probability space. We explain how to actually estimate probabilities from data in Section 1.4.

The probability that the plane is late equals

$$\mathrm{P}(L) = \mathrm{P}(L \cap R^c) + \mathrm{P}(L \cap R) \tag{1.35}$$

$$= \frac{1}{4} \tag{1.36}$$

because the events $L \cap R^c$ and $L \cap R$ are disjoint. However, this is not the probability we are interested in! According to our intuitive definition of probability in (1.1), we can interpret $\mathrm{P}(L)$ as

$$\mathrm{P}(L) = \frac{\text{times airplane is late}}{\text{total repetitions}}, \tag{1.37}$$

where we imagine that the flight is an experiment that can repeated many times. This is not what we want. Our goal is to determine the probability that the plane is late *if it rains*, which can be captured by modifying (1.37) to equal the fraction of late arrivals *out of the times it rains*. This yields the *conditional probability*

$$\mathrm{P}(L \,|\, R) = \frac{\text{times airplane is late and it rains}}{\text{times it rains}}. \tag{1.38}$$

Now, to express this quantity in terms of the probability measure of our probability space, we multiply and divide by the total repetitions,

$$\mathrm{P}(L \,|\, R) = \frac{\text{times airplane is late and it rains}}{\text{total repetitions}} \cdot \frac{\text{total repetitions}}{\text{times it rains}} \tag{1.39}$$

$$= \frac{\mathrm{P}(L \cap R)}{\mathrm{P}(R)}. \tag{1.40}$$

Since $\mathrm{P}(R) = \mathrm{P}(L \cap R) + \mathrm{P}(L^c \cap R) = 1/5$,

$$\mathrm{P}(L \,|\, R) = \frac{3}{4}, \tag{1.41}$$

which is three times larger than $\mathrm{P}(L)$.

Inspired by this example, let us define conditional probability more formally. Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space, and let $A \in \mathcal{C}$ be an event with nonzero probability. In order to condition on A, we build a new probability space $(\Omega_A, \mathcal{C}_A, \mathrm{P}\left(\cdot \,|\, A\right))$ that preserves the properties of the original probability space as much as possible, but where *all outcomes are in* A. We denote the new probability measure

$\mathrm{P}(\cdot \mid A)$ to indicate that we are conditioning on A. In the new probability space, every outcome belongs to A, so it is natural to set the new sample space equal to A, $\Omega_A := A$. If an outcome in the new probability space belongs to an event B in $\mathcal{C}$, then it must lie in $A \cap B$. We therefore define the new collection of events $\mathcal{C}_A$ as the collection containing the intersections of the events in $\mathcal{C}$ with A (in Exercise 1.1 we check that this satisfies the conditions in Definition 1.7).

All we have left is to define the probability measure $\mathrm{P}(\cdot \mid A)$. We could be tempted to just use the probability measure P of the original probability space. Any event in $\mathcal{C}_A$ is of the form $A \cap B$ for some $B \in \mathcal{C}$, so it also belongs to $\mathcal{C}$ and is assigned the probability $\mathrm{P}(A \cap B)$ by P. However, this does not yield a valid probability measure: The probability of the whole sample space would equal $\mathrm{P}(A)$ instead of one! The problem is that we are underestimating the probabilities because P assigns nonzero probability to A^c, which cannot occur in the new probability space. To correct for this, we divide by $P(A)$ to normalize all the probabilities in the new probability space. This yields a formal definition that coincides with our intuitive definition in (1.40).

Definition 1.16 (Conditional probability)**.** *Let A and B be events in a probability space $(\Omega, \mathcal{C}, \mathrm{P})$, and assume $\mathrm{P}(A) \neq 0$. The conditional probability of B given A is defined as*

$$\mathrm{P}(B \mid A) := \frac{\mathrm{P}(B \cap A)}{\mathrm{P}(A)}. \tag{1.42}$$

Defined in this way, $\mathrm{P}_A(B \cap A) := \mathrm{P}(\cdot \mid A)$ is a valid probability measure for the probability space $(A, \mathcal{C}_A, \mathrm{P}_A)$ (in Exercise 1.1 we check that it satisfies the conditions in Definition 1.9). Figure 1.2 uses a Venn diagram to provide a visualization of conditional probability for a simple example.

1.3.2 The Chain Rule

Conditional probabilities can be used to compute the intersection of several events in a structured way. By Definition 1.16, we can express the probability of the intersection of two events $A, B \in \mathcal{C}$ as follows,

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(A)\,\mathrm{P}(B \mid A) && (1.43)\\ &= \mathrm{P}(B)\,\mathrm{P}(A \mid B)\,. && (1.44) \end{aligned}$$

In this formula, $\mathrm{P}(A)$ is known as the *prior* probability of A, as it captures the information we have about A before anything else is revealed. Analogously, $\mathrm{P}(A \mid B)$ is known as the *posterior* probability. Generalizing (1.43) to a sequence of events gives the *chain rule*, which provides a factorization of the probability of the intersection of multiple events as a product of conditional probabilities.

Theorem 1.17 (Chain rule)**.** *Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space and $A_1, A_2, \ldots$*

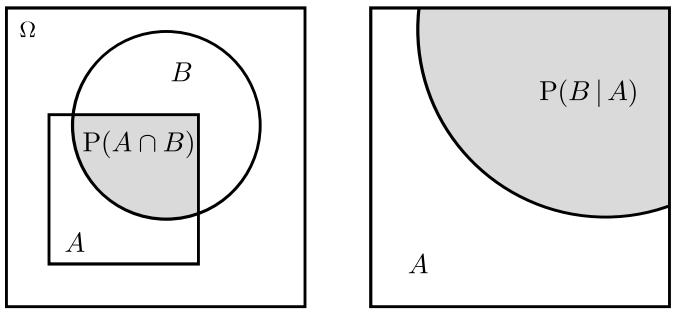

Figure 1.2 Visualizing conditional probability. The diagram on the left depicts a probability space where the sample space Ω is a square with area one. The shaded area is the intersection $A \cap B$ of events A (represented by a square) and B (represented by a circle). In order to condition on A, we update the probability space as shown on the right. We set the sample space to equal A, and discard the rest of Ω. In addition, we *blow up* A by a factor of $1/\mathrm{P}(A)$ to ensure that the new sample space has unit area. This increases the area assigned to $A \cap B$ from $\mathrm{P}(A \cap B)$ to $\mathrm{P}(B \mid A)$.

a collection of events in $\mathcal{C}$*,*

$$\mathrm{P}\left(\cap_i A_i\right) = \mathrm{P}\left(A_1\right) \mathrm{P}\left(A_2 \mid A_1\right) \mathrm{P}\left(A_3 \mid A_1 \cap A_2\right) \cdots \tag{1.45}$$

$$= \prod_i \mathrm{P}\left(A_i \mid \cap_{j=1}^{i-1} A_j\right). \tag{1.46}$$

Proof We prove the result for three events A, B, and C. The argument can be easily extended by induction to any (countable) number of events. From the definition of conditional probability,

$$\mathrm{P}\left(C \mid A \cap B\right) = \frac{\mathrm{P}\left(A \cap B \cap C\right)}{\mathrm{P}\left(A \cap B\right)} \tag{1.47}$$

$$= \frac{\mathrm{P}\left(B \cap C \mid A\right)}{\mathrm{P}\left(B \mid A\right)}, \tag{1.48}$$

where we have multiplied and divided by $\mathrm{P}(A)$ to obtain (1.48). This means that $P\left(B \cap C \mid A\right) = \mathrm{P}\left(B \mid A\right) \mathrm{P}\left(C \mid A \cap B\right)$, so by (1.44) applied to the events A and $B \cap C$,

$$\mathrm{P}\left(A \cap B \cap C\right) = \mathrm{P}\left(A\right) P\left(B \cap C \mid A\right) \tag{1.49}$$

$$= \mathrm{P}\left(A\right) \mathrm{P}\left(B \mid A\right) \mathrm{P}\left(C \mid A \cap B\right). \tag{1.50}$$

■

Note that the order in which we condition when applying the chain rule is *completely arbitrary*. For example, for three events A, B, and C we have six

possible factorizations, which include

$$\begin{aligned} \mathrm{P}\left(A \cap B \cap C\right) &= \mathrm{P}\left(A\right) P\left(B \,|\, A\right) P\left(C \,|\, A \cap B\right) && (1.51)\\ &= \mathrm{P}\left(B\right) P\left(C \,|\, B\right) P\left(A \,|\, B \cap C\right) && (1.52)\\ &= \mathrm{P}\left(C\right) P\left(A \,|\, C\right) P\left(B \,|\, A \cap C\right). && (1.53) \end{aligned}$$

In probabilistic modeling (and homework problems) it is often crucial to choose the order wisely in order to exploit the information that we have available.

In order to alleviate notation, in the rest of the book we often use a comma instead of $\cap$ to describe intersections of events. For example, we write $\mathrm{P}\left(A, B, C\right)$ instead of $\mathrm{P}\left(A \cap B \cap C\right)$.

1.3.3 Law Of Total Probability

Sometimes, estimating the probability of a certain event directly may be more challenging than estimating its probability conditioned on simpler events. The law of total probability, illustrated in Figure 1.3, allows us to pool these conditional probabilities together to compute the probability of the event.

Theorem 1.18 (Law of total probability)**.** *Let* $(\Omega, \mathcal{C}, \mathrm{P})$ *be a probability space and let the collection of disjoint sets* A_1, A_2, $\ldots \in \mathcal{C}$ *be any partition of* Ω*. For any event* $B \in \mathcal{C}$

$$\begin{aligned} \mathrm{P}\left(B\right) &= \sum_i \mathrm{P}\left(B \cap A_i\right) && (1.54)\\ &= \sum_i \mathrm{P}\left(A_i\right) \mathrm{P}\left(B \,|\, A_i\right). && (1.55) \end{aligned}$$

Proof This is an immediate consequence of the chain rule and Axiom 3 in Definition 1.9, since $B = \cup_i (B \cap A_i)$ and the events $B \cap A_i$ are disjoint. ■

Example 1.19 (Flight delay and rain)**.** Consider the problem of estimating the probability that a flight will be late tomorrow using the probabilities in (1.34). We have determined that $\mathrm{P}\left(L \,|\, R\right) = 0.75$ (see (1.41)). The same argument yields $\mathrm{P}\left(L \,|\, R^c\right) = 0.125$. After checking out a weather website, we determine that the chance of rain tomorrow is $1/5$, so $\mathrm{P}\left(R\right) = 0.2$. Now, how can we integrate all of this information? The events R and R^c are disjoint and cover the whole sample space, so they form a partition of the sample space. We apply the law of total probability to determine

$$\begin{aligned} \mathrm{P}\left(L\right) &= \mathrm{P}\left(L \,|\, R\right) \mathrm{P}\left(R\right) + \mathrm{P}\left(L \,|\, R^c\right) \mathrm{P}\left(R^c\right) && (1.56)\\ &= 0.75 \cdot 0.2 + 0.125 \cdot 0.8 = 0.25. && (1.57) \end{aligned}$$

The probability that the flight is delayed is $1/4$.

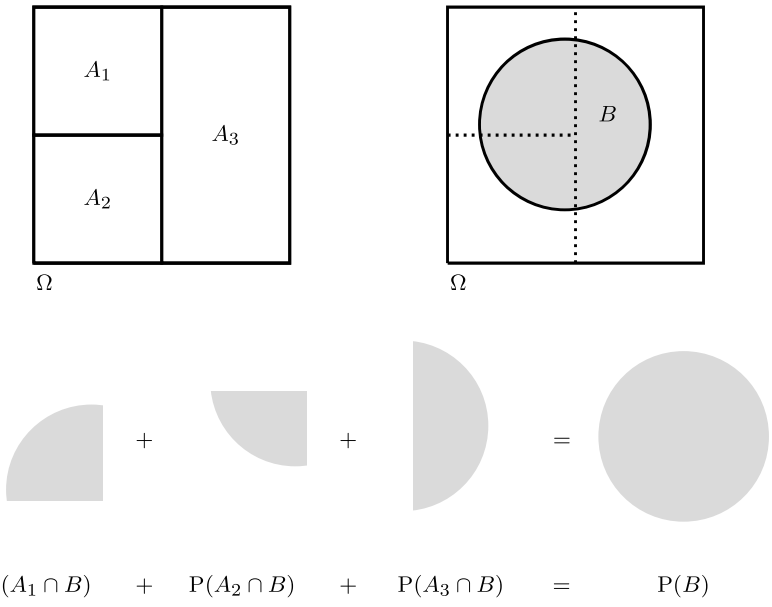

Figure 1.3 The law of total probability. The Venn diagram on the upper left shows a partition of Ω consisting of three events A_1, A_2, and A_3. The upper right diagram shows another event B. B can be decomposed into three events (components), each corresponding to the intersection of B with one of the events of the partition. These components are disjoint, so the sum of the areas of the 2D sets representing them is equal to the area of the 2D set representing B, as depicted below the Venn diagrams. Analogously, the probability of B is equal to the sum of the probabilities of $A_1 \cap B$, $A_2 \cap B$, and $A_3 \cap B$.

1.3.4 Bayes' Rule

It is important to realize that in general $\mathrm{P}(A \,|\, B) \neq \mathrm{P}(B \,|\, A)$. For example, most players in the NBA probably own a basketball: $\mathrm{P}(\text{owns ball} \,|\, \text{NBA})$ is very high. However, most people that own basketballs (including myself) are not in the NBA: $\mathrm{P}(\text{NBA} \,|\, \text{owns ball})$ is very low. The reason is that the prior probabilities are very different: $\mathrm{P}(\text{NBA})$ is much smaller than $\mathrm{P}(\text{owns ball})$. This is illustrated by a simple example in Figure 1.4. However, it is possible to *invert* conditional probabilities. We can compute $\mathrm{P}(A \,|\, B)$ from $\mathrm{P}(B \,|\, A)$, as long as we take into account the priors, by applying Bayes' rule.

Theorem 1.20 (Bayes' rule)**.** *For any events A and B in a probability space $(\Omega, \mathcal{C}, \mathrm{P})$*

$$\mathrm{P}(A \,|\, B) = \frac{\mathrm{P}(A)\,\mathrm{P}(B \,|\, A)}{\mathrm{P}(B)}, \tag{1.58}$$

as long as $\mathrm{P}(B) > 0$.

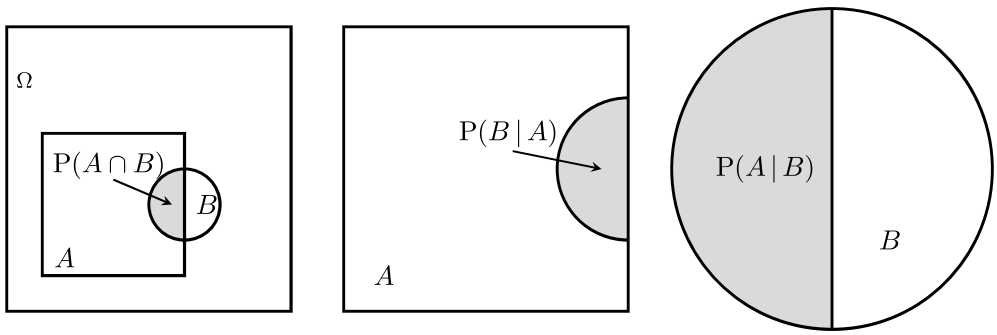

Figure 1.4 $\mathbf{P(A \mid B) \neq P(B \mid A)}$. The Venn diagram on the left depicts a probability space where the sample space Ω is a square with area one. The shaded area is the intersection $A \cap B$ of events A (represented by a square) and B (represented by a circle). In the middle, we condition on A by setting the sample space to equal A, and enlarging by a factor of $1/\mathrm{P}(A)$. On the right, we condition on B by setting the sample space to equal B, and enlarging it by a factor of $1/\mathrm{P}(B)$. Since $\mathrm{P}(A) \neq \mathrm{P}(B)$, $A \cap B$ is enlarged to different extents in each case, and therefore $\mathrm{P}(A \mid B) \neq \mathrm{P}(B \mid A)$.

Proof The result is a direct consequence of the definition of conditional probability and the chain rule. ■

Example 1.21 (Estimating the probability of rain from flight delay)**.** Imagine that the flight in Example 1.19 was finally late and you don't know whether it rained or not, because you spent the day indoors studying probability spaces. You decide to use your newly-acquired knowledge to estimate the probability that it rained. The prior probability of rain was 0.2, but since we know the flight was late, we should update the estimate. Applying Bayes' rule and the law of total probability:

$$\mathrm{P}(R \mid L) = \frac{\mathrm{P}(L \mid R)\,\mathrm{P}(R)}{\mathrm{P}(L)} \tag{1.59}$$

$$= \frac{\mathrm{P}(L \mid R)\,\mathrm{P}(R)}{\mathrm{P}(L \mid R)\,\mathrm{P}(R) + \mathrm{P}(L \mid R^c)\,\mathrm{P}(R^c)} \tag{1.60}$$

$$= \frac{0.75 \cdot 0.2}{0.75 \cdot 0.2 + 0.125 \cdot 0.8} = 0.6. \tag{1.61}$$

As expected, the posterior probability of rain is higher once we condition on the flight delay.

1.4 Estimating Probabilities From Data

The previous sections describe the machinery of probability spaces, which provides a set of rules to define and manipulate probabilities. Now, we ask a question

that takes us beyond probability theory into the realm of statistics and data science: *How do we estimate the probability of an event from data?*

In statistics, a rule for estimating a certain quantity of interest is called an *estimator*. In order to design an estimator for the probability of an event, we seek inspiration in our intuitive definition of probability (1.1). Assume that we have access to a dataset where each data point can be modeled as an outcome in a probability space. Since the probability of an event represents the fraction of times the event occurs, it seems natural to use the observed fraction of occurrences as an estimate of the probability.

Definition 1.22 (Empirical probability)**.** *Let Ω denote a sample space, and A an event within that sample space, $A \subseteq \Omega$. Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a dataset with values in Ω. The empirical probability of A is defined as the fraction of elements of X that belong to A,*

$$\mathrm{P}_X(A) := \frac{1}{n} \sum_{i=1}^{n} 1(x_i \in A), \tag{1.62}$$

where $1(x_i \in A)$ is an indicator function that is equal to one if $x_i \in A$ and to zero otherwise.

In words, the empirical probability of an event is the fraction of times we observe it in the data. It follows directly from the definition that this is a valid probability measure (see Exercise 1.2).

Example 1.23 (Unfair die)**.** In books about probability, six-sided dice are often assumed to be fair, meaning that there is an equal chance of rolling every number. However, this may not be the case for real dice. My daughter has a toy six-sided die, which I suspect is not fair. In order to resolve this question scientifically, I rolled it 60 times and recorded the results. Let n_j, $j \in \{1, 2, 3, 4, 5, 6\}$, denote the number of times that a roll with value j was observed. According to the data,

$$n_1 := 10, \quad n_2 := 8, \quad n_3 := 18, \quad n_4 := 7, \quad n_5 := 7, \quad n_6 := 10. \tag{1.63}$$

Following Example 1.14, we model the die roll using a probability space where the collection of events is the power set of the outcomes, and we define the probability measure by assigning a probability to the events $A_j := \{j\}$, for j in $\{1, 2, 3, 4, 5, 6\}$. Using empirical probabilities to estimate the measure from the data $X := \{x_1, x_2, \ldots, x_{60}\}$, where x_i indicates the value of the ith roll, yields

$$\mathrm{P}_X(A_j) := \frac{1}{60} \sum_{i=1}^{60} 1(x_i = j) \tag{1.64}$$

$$= \frac{n_j}{60}, \tag{1.65}$$

Table 1.1 ***Empirical probability of a coin toss.*** *The table shows ten different estimates of the probability of heads for a single coin toss. Each estimate was obtained by simulating a fair coin flip twenty times and then computing the empirical probability as described in Definition 1.22. Most of the empirical probabilities are different from 1/2.*

Heads (out of 20)	15	13	10	9	9	8	9	9	12	8
Empirical probability	0.75	0.65	0.5	0.45	0.45	0.4	0.45	0.45	0.6	0.4

so that

$$\mathrm{P}_X(A_1) = \frac{10}{60}, \qquad \mathrm{P}_X(A_2) = \frac{8}{60}, \qquad \mathrm{P}_X(A_3) = \frac{18}{60},$$
$$\mathrm{P}_X(A_4) = \frac{7}{60}, \qquad \mathrm{P}_X(A_5) = \frac{7}{60}, \qquad \mathrm{P}_X(A_6) = \frac{10}{60}. \tag{1.66}$$

From the results it looks like the die may not be fair, in line with my suspicions. In Chapter 10 we evaluate this conjecture rigorously using the framework of hypothesis testing.

When computing empirical probabilities, we interpret each data point as the result of repeating an experiment that represents the phenomenon of interest. Mathematically, we assume that the data are *independent and identically distributed* (i.i.d.), which means that the value of each data point only depends on the corresponding probability, and not on the value of the remaining data. We provide a more formal definition of the i.i.d. assumption in Example 2.18 and Definition 2.23.

In most cases, the i.i.d. assumption is just an approximation, but even if it were to hold exactly, empirical probabilities cannot be expected to be completely accurate. This is illustrated in Table 1.1, where we compute empirical probabilities in an idealized situation where the true underlying probabilities are known. We repeatedly simulate twenty flips from a fair coin (i.e. the probability of heads is 0.5), and compute the empirical probability of heads. The empirical probability is only correct once out of the ten repetitions. In fact, if we use twenty-one flips instead, we are guaranteed to never be right (we would need to observe ten and a half heads). This is our first encounter with a fundamental challenge in statistical estimation: estimates based on finite data are almost never exact. Fortunately, under certain reasonable assumptions, the empirical-probability estimator approximates the true probability of the event of interest with arbitrary precision, as long as it is computed from sufficient data, as established in Theorem 9.24 (see also Example 2.18).

Empirical probabilities can also be used to estimate conditional probabilities from data. Inspired by (1.40) we define the empirical conditional probability of an event B given another event A as the ratio between the observed simultaneous occurrences of A and B, and the observed occurrence of A. In other words, we

select the data points in A, and then compute the empirical probability of B from this reduced dataset.

Definition 1.24 (Empirical conditional probability)**.** *Let Ω denote a sample space, and A and B events within that sample space, $A, B \subseteq \Omega$. Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a dataset with values in Ω. The empirical conditional probability of B given A is the fraction of the elements of X in A that also belong to B,*

$$\mathrm{P}_X(B \mid A) := \frac{\sum_{i=1}^{n} 1(x_i \in A \cap B)}{\sum_{i=1}^{n} 1(x_i \in A)}, \tag{1.67}$$

where $1(x_i \in S)$ is an indicator function that is equal to one if $x_i \in S$ and to zero otherwise, for any event $S \subseteq \Omega$.

Example 1.25 (House of Representatives: Empirical probabilities)**.** In this example, we model the voting behavior of congressmen in the U.S. House of Representatives using Dataset 1, which consists of votes from 1984 on two issues: adoption of the budget resolution and duty-free exports. Table 1.2 shows the voting records. For simplicity we ignore absences and abstentions. We would like to understand the relationship between the two issues. If a representative votes Yes for the budget, are they more likely to vote Yes for duty-free exports? To answer such questions we build a probabilistic model, in which the voting process is interpreted as a repeatable experiment.

The outcome of the experiment is the votes on both issues, so the sample space just contains four possible outcomes: *Yes-Yes*, *Yes-No*, *No-Yes*, and *No-No*. We define the events B and D to represent positive votes on the budget and on the duty-free issue, respectively. Since we do not consider absences or abstentions, B^c and D^c represent negative votes. To estimate the probability of B and D we divide the positive votes for each issue by the total votes, following Definition 1.22:

$$\mathrm{P}(B) = \frac{239}{400} = 0.598, \tag{1.68}$$

$$\mathrm{P}(D) = \frac{172}{400} = 0.43. \tag{1.69}$$

To estimate the conditional probability of D given B, we only consider outcomes in B (i.e. representatives who voted Yes on the budget) and compute what fraction of them that are also in D,

$$\mathrm{P}(D \mid B) = \frac{151}{239} = 0.632. \tag{1.70}$$

Similarly,

$$\mathrm{P}(D \mid B^c) = \frac{21}{161} = 0.130. \tag{1.71}$$

Our analysis shows that if we know nothing about a representative, they are slightly more likely to vote No on the duty-free issue, because $\mathrm{P}(D)$ is smaller than $1/2$. However, if we know that they have voted Yes on the budget, then they

Table 1.2 ***Voting data from the U.S. House of Representatives.*** *Number of representatives who voted Yes or No on the adoption of the budget resolution, and on duty-free exports.*

		Duty-free exports	
		Yes	No
Budget	Yes	151	88
	No	21	140

are more likely to also vote Yes on the duty-free issue, because $\mathrm{P}(D \,|\, B)$ is larger than $1/2$. If we know that they voted No on the budget, then they are very likely to also vote No on the duty-free issue, because $\mathrm{P}(D^c \,|\, B^c) = 0.870$.

1.5 Independence

Conditional probabilities quantify the extent to which the knowledge of the occurrence of a certain event affects the probability of another event. In some cases, it makes no difference: the events are *independent*. More formally, events A and B are independent if and only if

$$\mathrm{P}\left(A \,|\, B\right) = \mathrm{P}\left(A\right). \tag{1.72}$$

This definition is not valid if $\mathrm{P}\left(B\right) = 0$. We usually use the following definition, which is equivalent but can also be applied when the probability of one of the events is zero.

Definition 1.26 (Independence of two events)**.** *Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space. Two events $A, B \in \mathcal{C}$ are independent if and only if*

$$\mathrm{P}\left(A \cap B\right) = \mathrm{P}\left(A\right)\mathrm{P}\left(B\right). \tag{1.73}$$

The following example shows that when we consider more than two events, pairwise independence does not necessarily imply a lack of dependence between the events.

Example 1.27 (Two coin flips)**.** Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space representing two fair coin flips. The sample space Ω contains four outcomes: *heads-heads*, *heads-tails*, *tails-heads*, and *tails-tails*. The collection $\mathcal{C}$ is the power set (all possible subsets) of Ω. The probability measure assigns

$$\mathrm{P}(\{\text{heads-heads}\}) = \mathrm{P}(\{\text{heads-tails}\}) = \mathrm{P}(\{\text{tails-heads}\}) = \mathrm{P}(\{\text{tails-tails}\}) = \frac{1}{4}.$$

We are interested in the following events:

$$A := \{\text{heads-heads}, \text{heads-tails}\} \qquad \text{(first flip is heads)}, \tag{1.74}$$
$$B := \{\text{heads-heads}, \text{tails-heads}\} \qquad \text{(second flip is heads)}, \tag{1.75}$$
$$C := \{\text{heads-heads}, \text{tails-tails}\} \qquad \text{(flips are the same)}. \tag{1.76}$$

We have

$$\mathrm{P}(A) = \mathrm{P}(\{\text{heads-heads}\} \cup \{\text{heads-tails}\}) = \frac{1}{2}, \tag{1.77}$$
$$\mathrm{P}(B) = \mathrm{P}(\{\text{heads-heads}\} \cup \{\text{tails-heads}\}) = \frac{1}{2}, \tag{1.78}$$
$$\mathrm{P}(C) = \mathrm{P}(\{\text{heads-heads}\} \cup \{\text{tails-tails}\}) = \frac{1}{2}. \tag{1.79}$$

The three events are pairwise independent:

$$\mathrm{P}(A, B) = \mathrm{P}(\{\text{heads-heads}\}) = \frac{1}{4} = \mathrm{P}(A)\mathrm{P}(B), \tag{1.80}$$
$$\mathrm{P}(A, C) = \mathrm{P}(\{\text{heads-heads}\}) = \frac{1}{4} = \mathrm{P}(A)\mathrm{P}(C), \tag{1.81}$$
$$\mathrm{P}(B, C) = \mathrm{P}(\{\text{heads-heads}\}) = \frac{1}{4} = \mathrm{P}(B)\mathrm{P}(C). \tag{1.82}$$

This makes sense. Revealing the result of the first flip provides no information about the result of the second flip. Does this imply there is no dependence between the three events? Not at all. The conditional probability of C given $A \cap B$ is

$$\mathrm{P}(C \,|\, A, B) = \frac{\mathrm{P}(A, B, C)}{\mathrm{P}(A, B)} \tag{1.83}$$
$$= \frac{\mathrm{P}(\{\text{heads-heads}\})}{\mathrm{P}(\{\text{heads-heads}\})} \tag{1.84}$$
$$= 1, \tag{1.85}$$

which is definitely not equal to $\mathrm{P}(C)$. The three events are therefore not independent.

Motivated by this example, we extend the definition of independence to more than two events.

Definition 1.28 (Mutual independence of multiple events)**.** *Let* $(\Omega, \mathcal{C}, \mathrm{P})$ *be a probability space. The events* $A_1, A_2, \ldots, A_n \in \mathcal{C}$ *are mutually independent if and only if for any possible subset of* m *events* $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$, $\{i_1, i_2, \ldots, i_m\} \subseteq \{1, 2, \ldots, n\}$,

$$\mathrm{P}\left(\cap_{j=1}^{m} A_{i_j}\right) = \prod_{j=1}^{m} \mathrm{P}\left(A_{i_j}\right). \tag{1.86}$$

If (1.86) holds for all possible subsets, then all conditional probabilities of

Table 1.3 ***Voting data from the U.S. House of Representatives.*** *Number of representatives who voted Yes or No on immigration, and on an anti-satellite test ban.*

		Immigration	
		Yes	No
Anti-satellite test ban	Yes	124	113
	No	89	93

A_i conditioned on any intersection of the remaining events equals $\mathrm{P}(A_i)$. For example,

$$\mathrm{P}(A_3 \,|\, A_1, A_2) = \frac{\mathrm{P}(A_1, A_2, A_3)}{\mathrm{P}(A_1, A_2)} \tag{1.87}$$

$$= \frac{\mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3)}{\mathrm{P}(A_1)\mathrm{P}(A_2)} \tag{1.88}$$

$$= \mathrm{P}(A_3). \tag{1.89}$$

The following example investigates independence using real data.

Example 1.29 (House of Representatives: Vote dependence)**.** Based on the empirical probabilities computed in Example 1.25, the events B and D are clearly not independent, since $\mathrm{P}(D) \neq \mathrm{P}(D \,|\, B)$. Here, we repeat the same analysis for two other issues, anti-satellite test ban (A) and immigration (I). To determine whether the events are independent, we compute the empirical probabilities

$$\mathrm{P}(A, I) = \frac{124}{419} = 0.296, \tag{1.90}$$

$$\mathrm{P}(A) = \frac{237}{419} = 0.566, \tag{1.91}$$

$$\mathrm{P}(I) = \frac{213}{419} = 0.508, \tag{1.92}$$

and verify that

$$\mathrm{P}(A)\mathrm{P}(I) = 0.288 \approx 0.296 = \mathrm{P}(A, I). \tag{1.93}$$

This seems to indicate that the events are almost independent. This is reflected in the conditional probabilities:

$$\mathrm{P}(A \,|\, I) = \frac{124}{213} \tag{1.94}$$

$$= 0.582 \approx \mathrm{P}(A). \tag{1.95}$$

In our model, the probability that a representative voted Yes on the anti-satellite test ban barely changes if we find out that they voted Yes on the immigration issue.

You may be a bit uneasy about our conclusion in Example 1.29. Strictly speaking, the events A and I are not independent, because this requires the equality

Table 1.4 ***Tom Brady and Category 5 hurricanes.*** *The table shows in what years between 2001 and 2020 Tom Brady won the Super Bowl (top row) and there was at least one Category 5 hurricane in the North Atlantic Ocean (bottom row).*

Year	02	03	04	05	06	07	08	09	10	11
Brady wins	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
Hurricane	✗	✓	✓	✓	✗	✓	✗	✗	✗	✗

Year	12	13	14	15	16	17	18	19	20	21
Brady wins	✗	✗	✗	✓	✗	✓	✗	✓	✗	✓
Hurricane	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗

in Eq. (1.93) to hold exactly. However, this will never happen when we use empirical probabilities computed from real data, because they are not usually not completely accurate, as discussed in Section 1.4. The following example illustrates these spurious dependencies for a case where we are pretty sure that the events are independent.

Example 1.30 (Tom Brady and Category 5 hurricanes)**.** Table 1.4 shows in what years between 2001 and 2020 Tom Brady won the Super Bowl (top row) and there was at least one Category 5 hurricane in the North Atlantic Ocean (bottom row). The empirical probability of a hurricane, represented by the event H, in any given year is

$$\mathrm{P}(H) = \frac{8}{20} = 0.4. \tag{1.96}$$

The empirical probability of a hurricane conditioned on the event that Tom Brady wins the Super Bowl, denoted by T is

$$\mathrm{P}(H \,|\, T) = \frac{4}{7} = 0.571. \tag{1.97}$$

If Brady wins again next year, should we begin emergency preparations in the Caribbean? Probably not. The dependence is caused by our limited number of data. In fact, if Brady had, for instance, won the 2012 Super Bowl and lost in 2017∗, then $\mathrm{P}(H \,|\, T)$ would equal 0.429, which is very close to $\mathrm{P}(H)$. In Example 10.25 we examine these data from the perspective of hypothesis testing. This example may seem a bit silly, but many a sport news article has been written with flimsier quantitative evidence.

∗If you follow American football, you might remember that in both cases this was very close to happening.

1.6 Conditional Independence

The dependence between two events in a probability space can change completely if we condition on a third event. To understand why, we define conditional independence, which is an important concept in probabilistic modeling and machine learning. Two events A and B are conditionally independent given a third event C if and only if

$$\mathrm{P}\left(A \,|\, B, C\right) = \mathrm{P}\left(A \,|\, C\right), \tag{1.98}$$

where $\mathrm{P}\left(A \,|\, B, C\right) := \mathrm{P}\left(A \,|\, B \cap C\right)$. Intuitively, this means that the probability of A is not affected by whether B occurs or not, *as long as* C *occurs*.

Definition 1.31 (Conditional independence)**.** *Let* $(\Omega, \mathcal{C}, \mathrm{P})$ *be a probability space. Two events* $A, B \in \mathcal{C}$ *are conditionally independent given a third event* $C \in \mathcal{C}$ *if and only if*

$$\mathrm{P}\left(A \cap B \,|\, C\right) = \mathrm{P}\left(A \,|\, C\right) \mathrm{P}\left(B \,|\, C\right). \tag{1.99}$$

The events $A_1, A_2, \ldots, A_n \in \mathcal{C}$ *are mutually conditionally independent given another event* C *if and only if for any possible subset of* m *events* A_{i_1}, A_{i_2}, $\ldots$, A_{i_m}, $\{i_1, i_2, \ldots, i_m\} \subseteq \{1, 2, \ldots, n\}$,

$$\mathrm{P}\left(\cap_{j=1}^{m} A_{i_j} \,|\, C\right) = \prod_{j=1}^{m} \mathrm{P}\left(A_{i_j} \,|\, C\right). \tag{1.100}$$

The following examples show that independence does not imply conditional independence or vice versa.

Example 1.32 (Conditional independence does not imply independence)**.** Let us consider the probability space in Example 1.21, extended to include the event that a taxi is available when the flight arrives. From past data we determine that

$$\mathrm{P}\left(T \,|\, R\right) = 0.1, \quad \mathrm{P}\left(T \,|\, R^c\right) = 0.6, \tag{1.101}$$

where T denotes the event of finding a taxi. We model the events L (flight is late) and T as conditionally independent given the events R (rain) and R^c (no rain),

$$\mathrm{P}\left(T, L \,|\, R\right) = \mathrm{P}\left(T \,|\, R\right) \mathrm{P}\left(L \,|\, R\right), \tag{1.102}$$

$$\mathrm{P}\left(T, L \,|\, R^c\right) = \mathrm{P}\left(T \,|\, R^c\right) \mathrm{P}\left(L \,|\, R^c\right). \tag{1.103}$$

We are assuming that availability of taxis is unrelated to flight delay, as long as we know whether it rains or not. Does this imply that they are also unrelated *if we don't know whether it rains*? More formally, are T and R independent?

They are not. By the law of total probability,

$$\begin{aligned}
\mathrm{P}(T) &= \mathrm{P}(T,R) + \mathrm{P}(T,R^c) && (1.104)\\
&= \mathrm{P}(T|R)\,\mathrm{P}(R) + \mathrm{P}(T|R^c)\,\mathrm{P}(R^c) && (1.105)\\
&= 0.1\cdot 0.2 + 0.6\cdot 0.8 = 0.5, && (1.106)\\
\mathrm{P}(T\,|\,L) &= \frac{\mathrm{P}(T,L,R) + \mathrm{P}(T,L,R^c)}{\mathrm{P}(L)} && (1.107)\\
&= \frac{\mathrm{P}(T\,|\,R)\,\mathrm{P}(L\,|\,R)\,\mathrm{P}(R) + \mathrm{P}(T\,|\,R^c)\,\mathrm{P}(L\,|\,R^c)\,\mathrm{P}(R^c)}{\mathrm{P}(L)}\\
&= \frac{0.1\cdot 0.75\cdot 0.2 + 0.6\cdot 0.125\cdot 0.8}{0.25} = 0.3. && (1.108)
\end{aligned}$$

$\mathrm{P}(T) \neq \mathrm{P}(T\,|\,L)$ so the events are *not* independent. The events L and T are connected through R. L provides information about R (if a flight is delayed, then it is more likely that it rained) and R provides information about T (taxis are more difficult to find if it rains). Consequently, L provides information about T. Conditional independence does not imply independence.

Example 1.33 (Independence does not imply conditional independence)**.** Flight delays are sometimes caused by mechanical problems in the airplane. We incorporate another event M to our model, which represents a mechanical problem in the plane. From past data, we obtain

$$\mathrm{P}(M) = \mathrm{P}(M\,|\,R) = \mathrm{P}(M\,|\,R^c) = 0.1, \tag{1.109}$$

so we conclude that the events M (*mechanical problem*) and R (*rain*) are independent. In addition, we estimate

$$\mathrm{P}(L\,|\,M) = 0.7, \quad \mathrm{P}(L\,|\,M^c) = 0.2, \quad \mathrm{P}(L\,|\,R^c, M) = 0.5.$$

Now, imagine that we are waiting for a flight, and we are wondering whether there could be a mechanical problem. Without any further information, the probability is 0.1. It is a sunny day, but this is of no help because according to our assumptions the events M and R are independent.

Suddenly they announce that the flight is late. Now, what is the probability that his plane had a mechanical problem? At first thought you might apply Bayes' rule to compute $\mathrm{P}(M\,|\,L) = 0.28$ as in Example 1.21. However, you are not using the fact that it is sunny. This means that the rain was not responsible for the delay, so intuitively a mechanical problem should be more likely. Indeed,

$$\begin{aligned}
\mathrm{P}(M\,|\,L,R^c) &= \frac{\mathrm{P}(L,R^c,M)}{\mathrm{P}(L,R^c)} && (1.110)\\
&= \frac{\mathrm{P}(L\,|\,R^c,M)\,\mathrm{P}(R^c)\,\mathrm{P}(M)}{\mathrm{P}(L\,|\,R^c)\,\mathrm{P}(R^c)} \quad \text{(by the chain rule)}\\
&= \frac{0.5\cdot 0.1}{0.125} = 0.4. && (1.111)
\end{aligned}$$

Table 1.5 ***Voting and political affiliation.*** *Number of Republicans (left) and Democrats (right) who voted Yes or No on the adoption of the budget resolution, and on duty-free exports.*

Republicans		Duty-free exports	
		Yes	No
Budget	Yes	7	15
	No	7	126

Democrats		Duty-free exports	
		Yes	No
Budget	Yes	144	73
	No	14	14

Since $\mathrm{P}\left(M \,|\, L, R^c\right) \neq \mathrm{P}\left(M \,|\, L\right)$ the events M and R are *not* conditionally independent given the event L. Independence does not imply conditional independence.

Example 1.34 (House of Representatives: Conditioning on political affiliation)**.** A key factor that determines how politicians vote in congress is political affiliation. In Example 1.25 we observe that the events B and D are not independent. Is it possible that the dependence is mainly due to political affiliation? If this were the case, then the two events would be conditionally independent given political affiliation. In order to investigate this, we incorporate affiliation by definining an event R, which indicates that the candidate is a Republican (R^c means that they are a Democrat).

From the data on the left of Table 1.5, we compute the empirical conditional probabilities given R,

$$\mathrm{P}(B, D \,|\, R) = \frac{7}{155} = 0.045, \tag{1.112}$$

$$\mathrm{P}(B \,|\, R) = \frac{22}{155} = 0.142, \tag{1.113}$$

$$\mathrm{P}(D \,|\, R) = \frac{14}{155} = 0.090, \tag{1.114}$$

and verify that

$$\mathrm{P}(B \,|\, R)\mathrm{P}(D \,|\, R) = 0.013 \tag{1.115}$$

is quite different from $\mathrm{P}(B, D \,|\, R)$. Consequently, the conditional probability

$$\mathrm{P}(B \,|\, R, D) = \frac{7}{14} = 0.5 \tag{1.116}$$

is very different from $\mathrm{P}(B \,|\, R)$. The events B and D are not conditionally independent given R. Does this mean that our hypothesis is completely wrong? Not completely.

We now condition on the representative being a Democrat. The empirical conditional probabilities (computed from the data on the right of Table 1.5) equal

$$\mathrm{P}(B, D \,|\, R^c) = \frac{144}{245} = 0.588, \tag{1.117}$$

$$\mathrm{P}(B \,|\, R^c) = \frac{217}{245} = 0.886, \tag{1.118}$$

$$\mathrm{P}(D \,|\, R^c) = \frac{158}{245} = 0.645, \tag{1.119}$$

so that

$$\mathrm{P}(B \,|\, R^c)\mathrm{P}(D \,|\, R^c) = 0.571 \approx \mathrm{P}(B, D \,|\, R^c), \tag{1.120}$$

B and D seem approximately conditionally independent given R^c. This is reflected in the conditional probability

$$\mathrm{P}(B \,|\, R^c, D) = \frac{144}{158} \tag{1.121}$$

$$= 0.911, \tag{1.122}$$

which is very close to $\mathrm{P}(B \,|\, R^c)$. According to the data, if we are interested in whether a Democrat has voted Yes on the budget, then knowing that they voted Yes on the duty-free exports provides very little information. This is not the case if we do not know the affiliation of the representative or if they are a Republican. This shows that conditioning (or not) on different events can completely change the dependence structure of a probability space.

1.7 The Monte Carlo Method

When applying probabilistic analysis in practice, one quickly comes to a realization that can be somewhat shocking: even if we have all the necessary information, it is often intractable to compute the probability of some events! We illustrate this through a probabilistic analysis of a basketball tournament in Example 1.36 below. Even if we know the probability of any team beating any other team, computing the probability that a team wins the tournament requires keeping track of an impossibly large number of possible results. Unfortunately, such combinatorial explosions are commonplace in probabilistic modeling. The Monte Carlo methods provides a pragmatic solution to this problem, inspired in our intuitive definition of probability (1.1): we generate a large number of simulated outcomes and compute the empirical probability of the event of interest.

Monte Carlo methods were developed in the context of nuclear-weapon research in the 1940s, pioneered by Stanislaw Ulam and John von Neumann. The name *Monte Carlo* was a code name inspired by the Monte Carlo Casino in Monaco. Ulam came up with the idea motivated by a game of cards. In his own words:*

The first thoughts and attempts I made to practice (the Monte Carlo Method)

*See http://en.wikipedia.org/wiki/Monte_Carlo_method#History

were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays.

This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later, I described the idea to John von Neumann, and we began to plan actual calculations.

Definition 1.35 (Monte Carlo method for estimating the probability of an event)**.** *Given a probability space* $(\Omega, \mathcal{C}, \mathrm{P})$*, let us assume that we can repeatedly generate outcomes from* Ω *according to the probability measure* P*. To approximate the probability of any event* A *in the collection* $\mathcal{C}$*, we:*

1 *Generate* n *simulated outcomes:* $s_1, s_2, \ldots, s_n \in \Omega$*.*
2 *Compute the fraction of the outcomes in* A*,*

$$\mathrm{P}_{MC}(A) := \frac{\sum_{i=1}^{n} 1(s_i \in A)}{n}, \tag{1.123}$$

where $1(x_i \in S)$ *is an indicator function that is equal to one if* $s_i \in S$ *and to zero otherwise, for any event* $S \subseteq \Omega$*.*

Similarly, to approximate the conditional probability of any event $B \in \mathcal{C}$ *conditioned on* A*, we:*

1 *Generate* n *simulated outcomes:* $s_1, s_2, \ldots, s_n \in \Omega$*.*
2 *Compute the fraction of the outcomes in* A *that are also in* B*,*

$$\mathrm{P}_{MC}(B \,|\, A) := \frac{\sum_{i=1}^{n} 1(s_i \in A \cap B)}{\sum_{i=1}^{n} 1(s_i \in A)}. \tag{1.124}$$

Example 1.36 (3x3 Olympic basketball tournament)**.** The 2020 Tokyo Olympics were the first to include 3x3 basketball. Eight teams participated: Belgium, China, Japan, Latvia, the Netherlands, Poland, the Russian Olympic Committee (ROC), and Serbia. Here, we imagine that the tournament has not happened yet, and we want to estimate the probability of each participant winning a gold, silver or bronze medal based on the ranking points of each individual player. These ranking points reflect the performance of each player in the previous 12 months. The left column in Table 1.6 shows the total points of the four players in each team.

We begin by using the ranking points to estimate the probability of each team

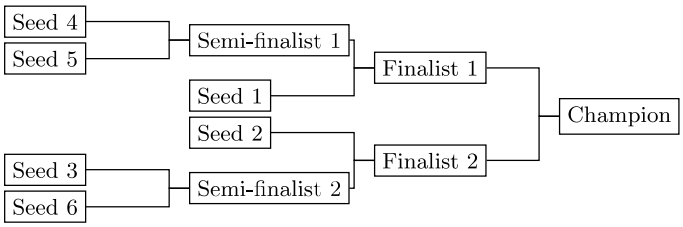

Figure 1.5 3x3 basketball bracket in the 2021 Tokyo Olympics. The eight participant teams were seeded according to the group stage. The first and second teams classified directly for the semi-finals. The teams ranked between third and sixth played the quarter-finals. The two last teams were eliminated.

beating every other team. The estimated probability that team A beats team B is

$$\mathrm{P}(\text{team A beats team B}) = \frac{\text{total points of A}}{\text{total points of A} + \text{total points of B}}. \tag{1.125}$$

For example,

$$\mathrm{P}(\text{Belgium beats Poland}) = 0.504, \tag{1.126}$$

$$\mathrm{P}(\text{China beats Serbia}) = 0.106, \tag{1.127}$$

$$\mathrm{P}(\text{Latvia beats the Netherlands}) = 0.794. \tag{1.128}$$

This is a simple heuristic that can probably be improved, but let us assume that we are happy with it. Now, how do we use these probabilities to compute the probability that a team wins the gold, silver or bronze medal?

We need to take into account the logistics of the tournament, which consisted of a group stage followed by playoffs. In the group stage, the eight participant teams played each other once, for a total of 28 games. The results determined the seeding for a playoff bracket described in Figure 1.5 with 6 more games (the 5 games in the figure and the bronze-medal game). In order to compute the probability that a team wins a medal, we need to sum the probabilities of all the ways in which this can happen. This essentially requires considering all 2^{34} possible results of the group stage and the bracket, which are more than ten billion! With modern computing this is not intractable, but would take some time. However, in many practical situations the number of possibilities quickly gets out of hand. For example, March Madness (the American college basketball championship) has 67 games with more than 10^{20} possible outcomes, and Wimbledon or the Premier League have even more games. In such cases, exact computation is impossible.

In order to approximate our probabilities of interest applying the Monte Carlo method in Definition 1.35, we repeatedly simulate the tournament using the probabilities in (1.125) and then compute the fraction of outcomes for which the event of interest occurs. Table 1.6 shows the results. Our model suggests that Latvia and Serbia were heavy favorites. Out of the 10^4 simulations of the tournament,

Table 1.6 ***Predicting the 3x3 basketball Olympic tournament.*** *The table shows the probability that each team wins a gold, silver or bronze medal, or wins the group stage based on the total ranking points of the players in each team before the tournament. The probabilities are estimated using* 10^4 *Monte Carlo simulations, as explained in Example 1.36.*

Country	Ranking points	Probability of winning (%)			
		Gold	Silver	Bronze	Group
Serbia	2,997,304	43.2	27.1	19.6	43.3
Latvia	2,959,152	42.0	28.0	18.9	42.9
ROC	970,438	6.3	14.9	18.9	5.6
Netherlands	768,134	3.6	10.3	14.4	3.2
Belgium	664,381	2.2	8.5	11.4	2.4
Poland	654,908	2.2	7.7	11.3	2.1
China	356,522	0.3	1.7	3.1	0.4
Japan	334,018	0.2	1.7	2.5	0.2

they each won about 40% of the time. Interestingly, their probability of winning the group stage (rightmost column of Table 1.6) is slightly higher than that of winning the gold medal. This makes sense: a single lost game in the bracket results in elimination, which favors upsets. In fact, in the actual tournament Serbia won the group stage undefeated, but lost to ROC in the semi-finals. Overall, the predictions of the model were quite reasonable. Latvia ended up winning gold, beating ROC in the final, and Serbia won bronze by beating Belgium in the bronze-medal game.

The number of simulations that we perform is obviously critical for the accuracy of the Monte Carlo method. If we are interested in an event with probability 0.01, we better perform at least a hundred simulations, otherwise chances are we won't observe it at all! For practical applications, it is therefore crucial to quantify the uncertainty associated with the probability estimates obtained via Monte Carlo simulation, which can be achieved using confidence intervals as explained in Example 9.46.

When approximating conditional probabilities, the number of *relevant* simulations can easily dwindle if the event we are considering is rare. To illustrate this, we repeat our predictions conditioning on the event that Serbia is eliminated in the group stage. Following Definition 1.35 we simulate the tournament 10^4 times. We then select the outcomes in which Serbia ends up 7th or 8th, and compute the fraction of those outcomes that are in each of the events of interest. If we don't pay attention, we could be fooled into thinking that this yields an accurate approximation because we are using 10^4 simulations. However, Serbia is eliminated in just 70 of them! Consequently, the estimated conditional probabilities, reported in Table 1.7, are not very precise. Increasing the number of simulations to 10^6, results in very different estimates. For instance, the conditional probability of Belgium winning gold drops from 10% to 6.5%. For $n := 10^6$, we obtain

Table 1.7 ***Conditioning on a rare event.*** *Predictions for the 3x3 basketball tournament in the 2021 Tokyo Olympics conditioned on the event that Serbia, one of the heavy favorites, is eliminated in the group phase. This happens with low probability, as is apparent from the last row, which shows the number of simulations in which the event occurs. We need at least one million total simulations in order to ensure that we observe enough instances for the conditional-probability estimates to be accurate.*

Country	Probability of gold conditioned on the event *Serbia does not reach bracket* (%)		
	10^4 runs	10^6 runs	10^7 runs
Latvia	68.6	63.5	63.4
ROC	10.0	13.3	13.2
Netherlands	7.1	8.5	8.6
Belgium	10.0	6.5	6.3
Poland	4.3	6.2	6.1
China	0	1.2	1.3
Japan	0	0.8	1.1
Serbia	0	0	0
Runs where Serbia does not reach bracket	70	5,539	55,719

more than $5,000$ relevant instances, which is sufficient to achieve a reasonable accuracy; increasing n to 10^7 barely changes the estimate.