1

Probability

Overview

Who will win the next presidential election? What will be the price of a certain stock tomorrow? Will the New York Knicks win the NBA championship next season? There is no definite answer to these questions, because they pertain to *uncertain* phenomena with different possible outcomes. To describe an uncertain phenomenon, we interpret it as a repeatable experiment, which enables us to define the *probability* of different events associated with the phenomenon. 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. Sections 1.5 and 1.6 introduce the key concepts of independence and conditional independence, respectively. Finally, Section 1.7 describes the Monte Carlo method, which enables us to approximate probabilities using computer simulations.

1.1 Intuitive Properties of Probability

In order to define probabilities associated with an uncertain phenomenon, we interpret the phenomenon as an *experiment* with multiple possible outcomes. The set of all possible outcomes 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, there are six possible results that are mutually exclusive (the die cannot land on two numbers at the same time). These six outcomes form the sample space $\Omega := \{1, 2, 3, 4, 5, 6\}$ associated with the die roll. In this case, the sample space is a finite set.

. . . .

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 Example 1.1. 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, then we can set the outcome to equal the sequence of rolls (e.g. if we roll a four, a one, and 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.

. . . .

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

Example 1.3 (Weather in New York: sample space)**.** If we want to model the weather in New York, 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. The sample space containing all possible outcomes is the real line $\Omega := \mathbb{R}$.[1] In this case, the sample space is continuous, and the number of outcomes is uncountable.

Once we have defined the sample space, we quantify the uncertainty about our phenomenon of interest by determining how likely it is for the outcome to belong to different subsets of the sample space. We call these subsets *events*. Events can consist of several outcomes, a single outcome, the whole sample space, or no outcomes at all. An event occurs when the outcome of the experiment belongs to the event, as illustrated by the following examples.

Example 1.4 (Die roll: events)**.** Possible events associated with the sample space in Example 1.1 include:

- Rolling a five, $A := \{5\}$.
- Rolling an even number, $B := \{2, 4, 6\}$.
- 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 depends on the choice of sample space. For example, the event *Rolling twice to obtain a six* contains a single outcome $\{2\}$, if the sample space is Ω_1, and five outcomes ($1 \rightarrow 6, 2 \rightarrow 6, 3 \rightarrow 6, 4 \rightarrow 6, 5 \rightarrow 6$), if the sample space is Ω_2.

Example 1.6 (Weather in New York)**.** If we model the temperature in Washington Square Park at noon and fix the sample space to be the real numbers ($\Omega := \mathbb{R}$), then possible events include:

- The temperature is above $30°$, $A := [30, \infty)$.
- The temperature is equal to $35°$, $B := 35$.
- The temperature is any number, $C := \mathbb{R}$.

If the temperature turns out to be $40°$, 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. Many uncertain phenomena, such as the next presidential election, will occur only once. However, thinking of them as repeatable experiments enables us to quantify our uncertainty about them. 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:

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

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

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

Notice that the probability is between zero and one because the number of times the event occurs must be between zero and the total number of repetitions. This is an informal definition of probability, which enables us to build intuition about its properties. We provide a formal definition in Section 1.2.

When determining the probabilities associated with a sample space, we do not need to assign a probability to every subset of the sample space. In fact, when the sample space is continuous, it is usually not possible to do this in a consistent manner. We refer the interested reader to any textbook on measure theory for more details. In any case, 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 using 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. This event contains all possible outcomes, so it is equal to the sample space Ω. Every time we repeat the experiment, we obtain an outcome that must be in Ω, so by our informal definition of probability,

$$\begin{aligned} \mathrm{P}(\Omega) &= \frac{\text{times } \Omega \text{ occurs}}{\text{total repetitions}} && (1.2) \\ &= \frac{\text{total repetitions}}{\text{total repetitions}} && (1.3) \\ &= 1. && (1.4) \end{aligned}$$

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

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

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

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

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

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

We subtract the probability of the intersection to avoid counting its outcomes twice.

From (1.10) we obtain a formula for the probability of the intersection of two events:

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

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, that is, 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

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

so to compute the probability of the complement of E, we just need to subtract its probability from one, $\mathrm{P}(E^c) = 1 - \mathrm{P}(E)$. An intuitive consequence is that 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 components. First, a sample space containing the mutually exclusive outcomes associated with 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.

The collection of events in a probability space must 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:*

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

1 *If an event belongs to the collection, $A \in \mathcal{C}$, then its complement belongs to the collection, $A^c \in \mathcal{C}$.*
2 *If two events A and B belong to the collection, $A, B \in \mathcal{C}$, then their union belongs to the collection, $A \cup B \in \mathcal{C}$. 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 *The sample space is in the collection, $\Omega \in \mathcal{C}$.*

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*. 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 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 (Die roll: collection of events)**.** A natural choice for the collection of events in our six-sided die example (Example 1.1) is the *power set* of the sample space $\Omega := \{1, 2, 3, 4, 5, 6\}$, which is the set of all $2^6 = 64$ 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 in Definition 1.7, Ω by Condition 3, and the empty set $\emptyset$ by Conditions 1 and 3. This is enough. You can check that the collection $\{\emptyset, A, A^c, \Omega\}$ satisfies Definition 1.7.

. . . .

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 of probability 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, which maps events in $\mathcal{C}$ to a number between 0 and 1, satisfying the following axioms:*

1 *All probabilities are nonnegative,* $\mathrm{P}(A) \geq 0$ *for any event $A \in \mathcal{C}$.*
2 *The probability of the sample space is one,* $\mathrm{P}(\Omega) = 1$.
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 the probability of their union equals the sum of their individual probabilities,*

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

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

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

Axiom 3 in Definition 1.9 implies the formula (1.11) for the probability of the intersection of two events, derived informally in Section 1.1.2.

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

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

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

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,

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

Then, 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 same argument as in Section 1.1.3, so we omit it here.

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 B contains another event A, then the probability of B cannot be smaller than the probability of A.

Lemma 1.12 (Subset of an event). *For any probability measure* P *satisfying the conditions in Definition 1.9, assume that 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 the two disjoint events $A \cap B$ and $A^c \cap B$. Since $A \subseteq B$, $A \cap B = A$, so that by Axiom 3 in Definition 1.9

$$\begin{aligned} \mathrm{P}(B) &= \mathrm{P}(A) + \mathrm{P}(A^c \cap B) && (1.23)\\ &\geq \mathrm{P}(A) && (1.24) \end{aligned}$$

because $\mathrm{P}(A^c \cap B) \geq 0$ by Axiom 1 in Definition 1.9. ■

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 its probability is not defined. Consider, for instance, the collection $\{\emptyset, A, A^c, \Omega\}$ in Example 1.8, where $A := \{5\}$. 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 area of the union of two disjoint two-dimensional (2D) shapes is the sum of their individual areas. This motivates the use of Venn diagrams to visualize probability spaces. In a Venn diagram, the outcomes in the sample space are represented

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

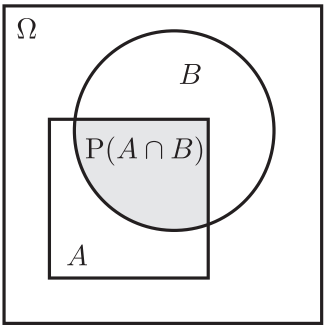

Figure 1.1 Venn diagram of a probability space. The Venn diagram represents a probability space. The sample space Ω is the big square that encompasses everything. The small square and the circle represent two events A and B. Their areas are equal to their respective probabilities. The probability of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

as points in two dimensions. Events are sets of points, depicted as regions delimited by closed curves. The probability of each event is 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, $\mathrm{P}\left(A^c\right) = 1-\theta$. By Axiom 2, $\mathrm{P}\left(\Omega\right) = 1$, which implies $\mathrm{P}\left(\emptyset\right) = 0$, also by Lemma 1.11.

. . . .

Example 1.14 (Die roll: probability measure)**.** As explained in Example 1.8, a reasonable choice for the collection of events associated with 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 (all possible subsets of Ω). However, we can apply a simple strategy: we divide the sample space into a *partition* of events and assign probabilities to the events in the 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 are disjoint and their union equals Ω. We assign a probability to each of them,

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

where $\theta_1, \theta_2, \ldots, \theta_6$ 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, due to Axioms 2 and 3 in Definition 1.9:

$$\sum_{i=1}^{6} \theta_i = \sum_{i=1}^{6} \mathrm{P}(A_i) \tag{1.26}$$

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

$$= \mathrm{P}(\cup_{i=1}^{6} A_i) \tag{1.27}$$
$$= P(\Omega) \tag{1.28}$$
$$= 1. \tag{1.29}$$

Let us assume that this condition is satisfied. Then we are actually done! Any event of the collection in the probability space can be decomposed as a union of events in the partition. Since these events are disjoint, we can add their individual probabilities to compute the probability of the event, leveraging Axiom 3 in Definition 1.9. For instance, the probability of the event *the roll is even* ($\{2, 4, 6\}$) equals

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

Note that for our strategy 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* $\mathcal{C}$*, which satisfies the conditions in Definition 1.7.*
- *A probability measure* P *assigning 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 like this. Instead, we use random variables to define probability spaces implicitly, without worrying about the gory mathematical details. We discuss random variables at length in Chapters 2–6.

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 interested in the probability that an airplane is late, if it rains. We define a probability space where the collection of events contains the event R (*it rains*), the event L (*the airplane is late*), and all their complements, unions, and intersections. Let us assume that we have estimated the following probabilities from past data, as described in Section 1.4:

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

$$\mathrm{P}\left(L \cap R^c\right) = \frac{2}{20}, \qquad \mathrm{P}\left(L^c \cap R^c\right) = \frac{14}{20},$$
$$\mathrm{P}\left(L \cap R\right) = \frac{3}{20}, \qquad \mathrm{P}\left(L^c \cap R\right) = \frac{1}{20}. \tag{1.33}$$

By Axiom 3 in Definition 1.9, the probability that the plane is late equals

$$\mathrm{P}(L) = \mathrm{P}(L \cap R^c) + \mathrm{P}(L \cap R) \tag{1.34}$$
$$= \frac{1}{4} \tag{1.35}$$

because $L = (L \cap R^c) \cup (L \cap R)$, and 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

$$\mathrm{P}(L) = \frac{\text{times airplane is late}}{\text{total repetitions}}, \tag{1.36}$$

where we imagine that the flight is an experiment that can be 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 (1.36) to equal the fraction of late arrivals *out of the times it rains*. This yields the *conditional probability*

$$\mathrm{P}(L \,|\, R) = \frac{\text{times airplane is late and it rains}}{\text{times it rains}}. \tag{1.37}$$

Now, to express this quantity in terms of our known probabilities, we multiply and divide by the total repetitions,

$$\mathrm{P}(L \,|\, R) = \frac{\text{times airplane is late and it rains}}{\text{total repetitions}} \cdot \frac{\text{total repetitions}}{\text{times it rains}} \tag{1.38}$$
$$= \frac{\mathrm{P}(L \cap R)}{\mathrm{P}(R)}. \tag{1.39}$$

Since $\mathrm{P}\left(L \cap R\right) = 3/20$ and $\mathrm{P}(R) = \mathrm{P}(L \cap R) + \mathrm{P}(L^c \cap R) = 1/5$ (by Axiom 3 in Definition 1.9),

$$\mathrm{P}(L \,|\, R) = 3/4. \tag{1.40}$$

The conditional probability that the plane is late, given that it rains, is three times larger than $\mathrm{P}(L)$.

Inspired by our intuitive reasoning, we 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 $\mathrm{P}\left(\cdot \,|\, A\right)$ 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 A with each event in $\mathcal{C}$ (in Exercise 1.1, we check that this satisfies the conditions in Definition 1.7).

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

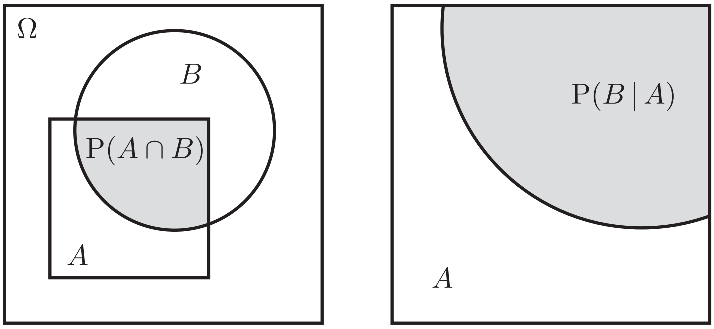

Figure 1.2 Conditional probability. The Venn diagram on the left depicts a probability space where the sample space Ω is a square with area one. The shaded area is the intersection $A \cap B$ of events A (represented by a square) and B (represented by a circle). In order to condition on A, we update 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* the area assigned to A in the Venn diagram 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 $\frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} := \mathrm{P}(B \mid A)$, which is the conditional probability of B given A by Definition 1.16.

All we have left to do is to define the probability measure $\mathrm{P}(\cdot \mid A)$. We could be tempted to use the same probability measure P as in 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 for our new probability space. The probability of the whole sample space would equal $\mathrm{P}(A)$ instead of 1! The problem is that P assigns nonzero probability to A^c, which cannot occur in the new probability space. To correct for this, we divide all the probabilities by $\mathrm{P}(A)$. This normalizes the conditional probabilities and ensures that the conditional probability of A given A equals 1. The resulting definition of conditional probability coincides with our intuitive definition (1.39).

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*

$$\mathrm{P}(B \mid A) := \frac{\mathrm{P}(B \cap A)}{\mathrm{P}(A)}. \tag{1.41}$$

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 axioms in Definition 1.9. Figure 1.2 uses a Venn diagram to illustrate the definition of conditional probability.

1.3.2 The Chain Rule

By Definition 1.16, we can express the probability of the intersection of two events A and B as follows:

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(A)\,\mathrm{P}(B \mid A) && (1.42) \\ &= \mathrm{P}(B)\,\mathrm{P}(A \mid B). && (1.43) \end{aligned}$$

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

In this formula, $\mathrm{P}(A)$ is known as the *prior* probability of A because it describes our uncertainty about A before anything else is revealed. $\mathrm{P}(B \mid A)$ is the *posterior* probability of B because it describes our uncertainty about B once we know that A occurred. Generalizing (1.42) to multiple events yields the *chain rule* that provides a factorization of the probability of the intersection of the events as a product of conditional probabilities.

Theorem 1.17 (Chain rule). *Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space. For any n events $A_1, A_2, \ldots, A_n$ belonging to the collection $\mathcal{C}$,*

$$\mathrm{P}(\cap_i A_i) = \mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid A_1)\,\mathrm{P}(A_3 \mid A_1 \cap A_2) \cdots \mathrm{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}) \tag{1.44}$$

$$= \mathrm{P}(A_1) \prod_{i=2}^{n} \mathrm{P}\left(A_i \mid \cap_{j=1}^{i-1} A_j\right). \tag{1.45}$$

Proof We prove the result for three events A_1, A_2, and A_3. The argument can be easily extended by induction to any countable number of events. We apply (1.43) twice. Setting $A := A_3$ and $B := A_1 \cap A_2$ yields

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1 \cap A_2)\,\mathrm{P}(A_3 \mid A_1 \cap A_2). \tag{1.46}$$

Setting $A := A_2$ and $B := A_1$, this implies

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\,\mathrm{P}(A_2 \mid A_1)\,\mathrm{P}(A_3 \mid A_1 \cap A_2). \tag{1.47}$$

■

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,

$$\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A)\,P(B \mid A)\,P(C \mid A \cap B) \tag{1.48}$$
$$= \mathrm{P}(A)\,P(C \mid A)\,P(B \mid A \cap C) \tag{1.49}$$
$$= \mathrm{P}(B)\,P(A \mid B)\,P(C \mid A \cap B) \tag{1.50}$$
$$= \mathrm{P}(B)\,P(C \mid B)\,P(A \mid B \cap C) \tag{1.51}$$
$$= \mathrm{P}(C)\,P(A \mid C)\,P(B \mid A \cap C) \tag{1.52}$$
$$= \mathrm{P}(C)\,P(B \mid C)\,P(A \mid B \cap C). \tag{1.53}$$

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.

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

1.3.3 Law of Total Probability

Sometimes, estimating the probability of a certain event is more difficult than estimating its probability conditioned on simpler events. The law of total probability, illustrated in Figure 1.3, allows us to compute the probability of the event by combining such conditional probabilities.

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

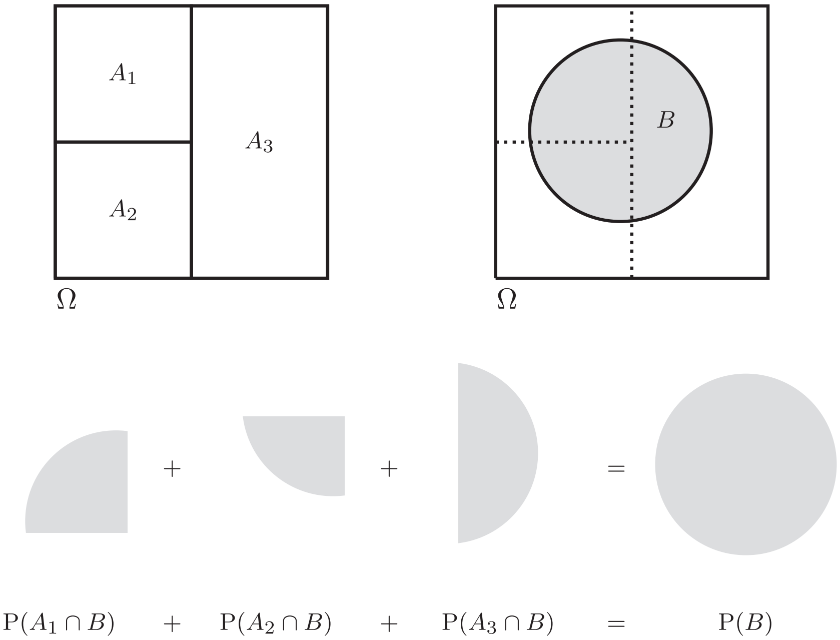
A_1
A_2
A_3
Ω
B
Ω
+
+
=
$\mathrm{P}(A_1 \cap B)$ + $\mathrm{P}(A_2 \cap B)$ + $\mathrm{P}(A_3 \cap B)$ = $\mathrm{P}(B)$

Figure 1.3 The law of total probability. The Venn diagram in the upper left corner shows a partition of Ω consisting of three events A_1, A_2, and A_3. The Venn diagram in the upper right corner shows another event B. B can be decomposed into the union of three smaller events, each equal to its intersection with one of the events of the partition. These events are disjoint, so the sum of the areas of the 2D regions representing them in the Venn diagram is equal to the area of the 2D region representing B, as depicted by the graphic equation underneath the Venn diagrams. This is consistent with Theorem 1.18 that establishes that the probability of B is equal to the sum of the probabilities of $A_1 \cap B$, $A_2 \cap B$, and $A_3 \cap B$.

Theorem 1.18 (Law of total probability). *Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space, and let A_1, $A_2, \ldots \in \mathcal{C}$ be a partition of the sample space Ω (meaning that the events are disjoint and $\cup_i A_i = \Omega$). For any event B belonging to the collection $\mathcal{C}$,*

$$\mathrm{P}(B) = \sum_i \mathrm{P}(B \cap A_i) \tag{1.54}$$

$$= \sum_i \mathrm{P}(A_i)\, \mathrm{P}(B \,|\, A_i)\,. \tag{1.55}$$

Proof Consider the intersections between B and the events in the partition $B \cap A_1$, $B \cap A_2, \ldots$ (illustrated in Figure 1.3). Their union $\cup_i (B \cap A_i)$ is equal to B. To prove this, we show that the two events contain each other. $\cup_i (B \cap A_i) \subseteq B$ because every element of $\cup_i (B \cap A_i)$ is in $B \cap A_i$ for some i, and consequently in B. Conversely, $B \subseteq \cup_i (B \cap A_i)$ because every element of B is in A_i, and consequently in $B \cap A_i$, for one value of i (because $A_1, A_2, \ldots$ form a partition). Since $B \cap A_1, B \cap A_2, \ldots$ are disjoint (because $A_1, A_2, \ldots$ are disjoint), by Axiom 3 in Definition 1.9, and the chain rule (Theorem 1.17),

$$\mathrm{P}(B) = \mathrm{P}(\cup_i (B \cap A_i)) \tag{1.56}$$

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

$$= \sum_i \mathrm{P}(B \cap A_i) \tag{1.57}$$

$$= \sum_i \mathrm{P}(A_i)\,\mathrm{P}(B \mid A_i)\,. \tag{1.58}$$

■

Example 1.19 (Flight delay and rain)**.** Imagine that in our flight delay example, we only have access to the conditional probability that the flight is late given rain and no rain, but we are interested in the probability that the flight is late. Specifically, we know that $\mathrm{P}(L \mid R) = 0.75$ and $\mathrm{P}(L \mid R^c) = 0.125$, and we want to compute $\mathrm{P}(L)$. The events R and R^c are disjoint and cover the whole sample space, so they form a partition of the sample space. Consequently, as long as we know their probabilities, we can apply the law of total probability to obtain the probability that the flight is late. Assuming $\mathrm{P}(R) = 0.2$, by Theorem 1.18,

$$\mathrm{P}(L) = \mathrm{P}(L \mid R)\,\mathrm{P}(R) + \mathrm{P}(L \mid R^c)\,\mathrm{P}(R^c) \tag{1.59}$$

$$= 0.75 \cdot 0.2 + 0.125 \cdot 0.8 = 0.25. \tag{1.60}$$

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

1.3.4 Bayes' Rule

It is important to realize that in general $\mathrm{P}(A \mid B)$ is not necessarily equal to $\mathrm{P}(B \mid A)$. For example, most players in the NBA probably own a basketball: $\mathrm{P}(\text{owns ball} \mid \text{NBA})$ is very high. However, most people who own basketballs (including myself) are not in the NBA: $\mathrm{P}(\text{NBA} \mid \text{owns ball})$ 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 in Figure 1.4. Consequently, in order to compute $\mathrm{P}(A \mid B)$ from $\mathrm{P}(B \mid A)$, we need to take into account the prior probability of each event, as dictated by Bayes' rule.

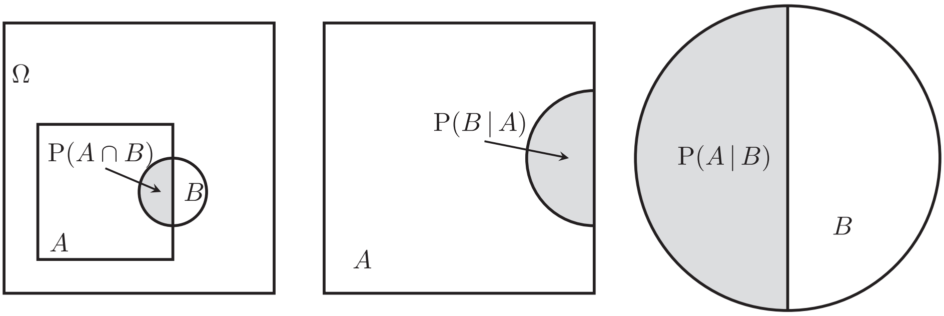

Figure 1.4 $\mathbf{P(B \mid A) \neq P(A \mid B)}$**.** The Venn diagram on the left depicts a probability space where the sample space Ω is a square with area one. The shaded area is the intersection $A \cap B$ of the events A (represented by a square) and B (represented by a circle). In the middle diagram, we condition on A by setting the sample space to equal A, and expanding its area by a factor of $1/\mathrm{P}(A)$ (as in Figure 1.2). In the right diagram, we condition on B by setting the sample space to equal B, and expanding its area by a factor of $1/\mathrm{P}(B)$. Since $\mathrm{P}(A) \neq \mathrm{P}(B)$, the area corresponding to $A \cap B$ is enlarged to different extents in each case, and therefore $\mathrm{P}(A \mid B) \neq \mathrm{P}(B \mid A)$.

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

Theorem 1.20 (Bayes' rule). *For any events A and B in a probability space,*

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A)\,\mathrm{P}(B \mid A)}{\mathrm{P}(B)}, \tag{1.61}$$

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

Proof By the definition of conditional probability (Definition 1.16) and the chain rule (Theorem 1.17),

$$\mathrm{P}(A \mid B) := \frac{\mathrm{P}(A, B)}{\mathrm{P}(B)} \tag{1.62}$$

$$= \frac{\mathrm{P}(A)\,\mathrm{P}(B \mid A)}{\mathrm{P}(B)}. \tag{1.63}$$

■

Example 1.21 (Conditional probability of rain given flight delay)**.** Imagine that the flight in Example 1.19 was late, but 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 is $\mathrm{P}(R) = 0.2$, but since we know the flight was late, we should update the estimate. Applying Bayes' rule and the law of total probability:

$$\mathrm{P}(R \mid L) = \frac{\mathrm{P}(L \mid R)\,\mathrm{P}(R)}{\mathrm{P}(L)} \tag{1.64}$$

$$= \frac{\mathrm{P}(L \mid R)\,\mathrm{P}(R)}{\mathrm{P}(L \mid R)\,\mathrm{P}(R) + \mathrm{P}(L \mid R^c)\,\mathrm{P}(R^c)} \tag{1.65}$$

$$= \frac{0.75 \cdot 0.2}{0.75 \cdot 0.2 + 0.125 \cdot 0.8} = 0.6. \tag{1.66}$$

The posterior probability of rain conditioned on the flight delay is much higher than the prior probability of rain.

....

1.4 Estimating Probabilities from Data

Sections 1.1–1.3 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,*

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

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

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. In Exercise 1.2, we check that the empirical probability is a valid probability measure that satisfies the axioms in Definition 1.9.

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,

$$n_1 := 10, \quad n_2 := 8, \quad n_3 := 18, \quad n_4 := 7, \quad n_5 := 7, \quad n_6 := 10. \tag{1.68}$$

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 each event $A_j := \{j\}$, for j in $\{1,2,3,4,5,6\}$. We denote the data by

$$X := \{x_1, x_2, \dots, x_{60}\}, \tag{1.69}$$

where x_i indicates the value of the ith roll. By Definition 1.22, the empirical probability of A_j is

$$\mathrm{P}_X(A_j) := \frac{1}{60}\sum_{i=1}^{60} 1(x_i = j) \tag{1.70}$$

$$= \frac{n_j}{60}, \tag{1.71}$$

which yields

$$\mathrm{P}_X(A_1) = \frac{10}{60}, \qquad \mathrm{P}_X(A_2) = \frac{8}{60}, \qquad \mathrm{P}_X(A_3) = \frac{18}{60},$$

$$\mathrm{P}_X(A_4) = \frac{7}{60}, \qquad \mathrm{P}_X(A_5) = \frac{7}{60}, \qquad \mathrm{P}_X(A_6) = \frac{10}{60}. \tag{1.72}$$

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 values of the other data. We provide a more formal definition of the i.i.d. assumption in Example 2.18 and Definition 2.23.

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

Table 1.1 ***Empirical probability of a coin toss.*** *The table shows ten different estimates of the probability of heads, 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 the true underlying probability, equal to 0.5.*

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

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 perform the following procedure ten times: We simulate twenty flips from a fair coin, for which the probability of heads is 0.5, and compute the empirical probability of heads. The empirical probability is equal to the true probability only once. In fact, if we use twenty-one flips instead, the empirical probability cannot be correct (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. However, the empirical-probability estimator approximates the true probability with arbitrary precision if the number of data is large enough (under certain reasonable assumptions), 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 (1.39) we define the empirical conditional probability of an event B given another event A as the fraction of times we observe B within the subset of data where A occurs.

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

$$\mathrm{P}_X(B \,|\, A) := \frac{\sum_{i=1}^{n} 1(x_i \in A \cap B)}{\sum_{i=1}^{n} 1(x_i \in A)}, \tag{1.73}$$

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 US House of Representatives using data extracted from Dataset 1. We consider votes 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 consists of the votes on both issues. The sample space contains the four possible outcomes: *Yes-Yes*, *Yes-No*, *No-Yes*, and *No-No*. We define the events B and E to represent positive votes on the budget and on the duty-free exports issue, respectively. Since we do not consider absences or abstentions, B^c and E^c represent negative

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

Table 1.2 ***Voting data from the US House of Representatives.*** *Number of representatives who voted Yes or No on the adoption of the budget resolution, and on duty-free exports, in Dataset 1.*

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

votes. To estimate the probability of B and E, we divide the positive votes for each issue by the total votes, following Definition 1.22:

$$\mathrm{P}(B) = \frac{239}{400} = 0.598, \tag{1.74}$$

$$\mathrm{P}(E) = \frac{172}{400} = 0.43. \tag{1.75}$$

To estimate the conditional probability of E 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 E, following Definition 1.24:

$$\mathrm{P}(E \,|\, B) = \frac{151}{239} = 0.632. \tag{1.76}$$

Similarly,

$$\mathrm{P}(E \,|\, B^c) = \frac{21}{161} = 0.130. \tag{1.77}$$

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 $\mathrm{P}(E)$ is smaller than 1/2. However, if we know that they have voted Yes on the budget, then they are more likely to also vote Yes on the duty-free issue because $\mathrm{P}(E \,|\, B)$ 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 $\mathrm{P}(E^c \,|\, B^c) = 0.870$.

1.5 Independence

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

$$\mathrm{P}\left(A \,|\, B\right) = \mathrm{P}\left(A\right). \tag{1.78}$$

This definition is not valid if $\mathrm{P}\left(B\right) = 0$. We usually use the following definition, which is equivalent to (1.78) by the chain rule (Theorem 1.17), 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*

$$\mathrm{P}\left(A \cap B\right) = \mathrm{P}\left(A\right) \mathrm{P}\left(B\right). \tag{1.79}$$

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

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

We are interested in the following events:

$$\begin{aligned} A &:= \{\text{heads-heads}, \text{heads-tails}\} && (\text{first flip is heads}), && (1.80)\\ B &:= \{\text{heads-heads}, \text{tails-heads}\} && (\text{second flip is heads}), && (1.81)\\ C &:= \{\text{heads-heads}, \text{tails-tails}\} && (\text{flips are the same}). && (1.82) \end{aligned}$$

By Axiom 3 in Definition 1.9,

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(\{\text{heads-heads}\} \cup \{\text{heads-tails}\}) && (1.83)\\ &= \mathrm{P}(\{\text{heads-heads}\}) + \mathrm{P}(\{\text{heads-tails}\}) && (1.84)\\ &= \frac{1}{2} && (1.85) \end{aligned}$$

since the individual events are disjoint. Similarly,

$$\begin{aligned} \mathrm{P}(B) &= \mathrm{P}(\{\text{heads-heads}\} \cup \{\text{tails-heads}\}) = \frac{1}{2}, && (1.86)\\ \mathrm{P}(C) &= \mathrm{P}(\{\text{heads-heads}\} \cup \{\text{tails-tails}\}) = \frac{1}{2}. && (1.87) \end{aligned}$$

By Definition 1.26, A, B, and C are pairwise independent:

$$\begin{aligned} \mathrm{P}(A, B) &= \mathrm{P}(\{\text{heads-heads}\}) = \frac{1}{4} = \mathrm{P}(A)\mathrm{P}(B), && (1.88)\\ \mathrm{P}(A, C) &= \mathrm{P}(\{\text{heads-heads}\}) = \frac{1}{4} = \mathrm{P}(A)\mathrm{P}(C), && (1.89)\\ \mathrm{P}(B, C) &= \mathrm{P}(\{\text{heads-heads}\}) = \frac{1}{4} = \mathrm{P}(B)\mathrm{P}(C). && (1.90) \end{aligned}$$

This makes sense. Revealing the result of the first flip provides no information about the result of the second flip. However, does this imply there is *no dependence between the three events*? Not at all! Notice that $A \cap B \cap C = \{\text{heads-heads}\} = A \cap B$, so the conditional probability of C given $A \cap B$ is

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

$$\begin{aligned} P(C \mid A, B) &= \frac{P(A, B, C)}{P(A, B)} && (1.91) \\ &= \frac{P(\{\text{heads-heads}\})}{P(\{\text{heads-heads}\})} && (1.92) \\ &= 1, && (1.93) \end{aligned}$$

which is definitely not equal to $P(C)$. Indeed, if we know that the first flip is heads and also that the second flip is heads, we can be sure that the two flips are the same! The three events are therefore not independent, despite being pairwise 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* $(\Omega, \mathcal{C}, P)$ *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}, …, A_{i_m}, $\{i_1, i_2, \ldots, i_m\} \subseteq \{1, 2, \ldots, n\}$,

$$P\left(\cap_{j=1}^{m} A_{i_j}\right) = \prod_{j=1}^{m} P\left(A_{i_j}\right). \qquad (1.94)$$

Definition 1.28 guarantees complete independence because if (1.94) holds for all possible subsets of events, then all conditional probabilities of each event A_i conditioned on any subset of the remaining events equal $P(A_i)$. For example,

$$\begin{aligned} P(A_3 \mid A_1, A_2) &= \frac{P(A_1, A_2, A_3)}{P(A_1, A_2)} && (1.95) \\ &= \frac{P(A_1)P(A_2)P(A_3)}{P(A_1)P(A_2)} && (1.96) \\ &= P(A_3). && (1.97) \end{aligned}$$

The following example investigates independence between events 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 E are clearly not independent, since $P(E)$ is very different from $P(E \mid B)$. Here, we repeat the same analysis for two other issues. Voting Yes on an anti-satellite test ban is represented by the event A, and voting Yes on an immigration issue is represented by the event I. Table 1.3 shows the data, which

Table 1.3 ***Voting data from the US House of Representatives.*** *Number of representatives who voted Yes or No on an immigration issue, and on an anti-satellite test ban, in Dataset 1.*

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

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

are also extracted from Dataset 1. We again ignore absences and abstentions. To determine whether the events are independent, we compute the empirical probabilities following Definition 1.22,

$$\mathrm{P}(A, I) = \frac{124}{419} = 0.296, \tag{1.98}$$

$$\mathrm{P}(A) = \frac{237}{419} = 0.566, \tag{1.99}$$

$$\mathrm{P}(I) = \frac{213}{419} = 0.508, \tag{1.100}$$

and verify that

$$\mathrm{P}(A)\mathrm{P}(I) = 0.288 \approx 0.296 = \mathrm{P}(A, I). \tag{1.101}$$

This seems to indicate that the events are almost independent, which is also reflected in the conditional probabilities. By Definition 1.24,

$$\mathrm{P}(A \,|\, I) = \frac{124}{213} = 0.582 \approx 0.566 = \mathrm{P}(A). \tag{1.102}$$

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 equality to hold exactly in (1.101). However, in practice, we *cannot expect to observe exact equality* for empirical probabilities computed from data. The reason is that these probabilities are extremely unlikely to be completely accurate, as discussed in Section 1.4 (see Table 1.1). The following example illustrates this in a situation, 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 Tom Brady won the Super Bowl (top row) and in what years there was at least one Category 5

Table 1.4 ***Tom Brady and Category 5 hurricanes.*** *The table shows in what years Tom Brady won the Super Bowl (top row) and in what years there was at least one Category 5 hurricane in the North Atlantic Ocean (bottom row) between 2002 and 2021.*

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	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗

hurricane in the North Atlantic Ocean (bottom row) between 2002 and 2021. By Definition 1.22, the empirical probability of a hurricane, represented by the event H, is

$$\mathrm{P}(H) = \frac{8}{20} = 0.4. \tag{1.103}$$

We denote the event that Tom Brady wins the Super Bowl by T. Conditioned on this event, by Definition 1.24, the empirical probability of a hurricane is

$$\mathrm{P}(H \,|\, T) = \frac{4}{7} = 0.571, \tag{1.104}$$

which is very different from $\mathrm{P}(H)$. Is this proof that the two events are not independent? No, we simply don't have enough data. In fact, if Brady had won the 2012 Super Bowl and lost in 2017,[2] then $\mathrm{P}(H \,|\, T)$ would equal 0.429, which is almost equal 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 sport news articles have been written with flimsier quantitative evidence.

1.6 Conditional Independence

The dependence between two events in a probability space can change completely, when we condition on a third event. In particular, conditioning may render the two events independent from each other. This is captured by the concept of conditional independence. Two events A and B are conditionally independent given a third event C, if and only if

$$\mathrm{P}\left(B \,|\, A, C\right) = \mathrm{P}\left(B \,|\, C\right), \tag{1.105}$$

where $\mathrm{P}\left(B \,|\, A, C\right) := \mathrm{P}\left(B \,|\, A \cap C\right)$. Intuitively, this means that the probability of B is not affected by whether A occurs or not, *as long as C is known to occur*. The chain rule (Theorem 1.17) holds for the probability measure $\mathrm{P}\left(\cdot \,|\, C\right)$ (as it is a valid probability measure, see Exercise 1.1), so (1.105) is equivalent to

$$\begin{aligned}\mathrm{P}\left(A, B \,|\, C\right) &= \mathrm{P}\left(A \,|\, C\right)\mathrm{P}\left(B \,|\, A, C\right) && (1.106)\\ &= \mathrm{P}\left(A \,|\, C\right)\mathrm{P}\left(B \,|\, C\right). && (1.107)\end{aligned}$$

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*

$$\mathrm{P}\left(A \cap B \,|\, C\right) = \mathrm{P}\left(A \,|\, C\right)\mathrm{P}\left(B \,|\, C\right). \tag{1.108}$$

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}, …, A_{i_m}, $\{i_1, i_2, \ldots, i_m\} \subseteq \{1, 2, \ldots, n\}$,

$$\mathrm{P}\left(\cap_{j=1}^{m} A_{i_j} \,|\, C\right) = \prod_{j=1}^{m} \mathrm{P}\left(A_{i_j} \,|\, C\right). \tag{1.109}$$

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

[2] If you follow American football, you might know that this was very close to happening.

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

Example 1.32 (Conditional independence does not imply independence)**.** Let us consider the probability space in Example 1.19, extended to include the event T, representing that a taxi is available when the flight arrives. Assume that

$$\mathrm{P}(T \mid R) = 0.1, \quad \mathrm{P}(T \mid R^c) = 0.6, \tag{1.110}$$

where R denotes the event that it rains. We model the events L (*flight is late*) and T as conditionally independent given the events R and R^c,

$$\mathrm{P}(T, L \mid R) = \mathrm{P}(T \mid R)\,\mathrm{P}(L \mid R), \tag{1.111}$$
$$\mathrm{P}(T, L \mid R^c) = \mathrm{P}(T \mid R^c)\,\mathrm{P}(L \mid R^c). \tag{1.112}$$

We are assuming that the 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 (Theorem 1.18) and the chain rule (Theorem 1.17), since R and R^c form a partition of the sample space,

$$\mathrm{P}(T) = \mathrm{P}(T, R) + \mathrm{P}(T, R^c) \tag{1.113}$$
$$= \mathrm{P}(T \mid R)\,\mathrm{P}(R) + \mathrm{P}(T \mid R^c)\,\mathrm{P}(R^c) \tag{1.114}$$
$$= 0.1 \cdot 0.2 + 0.6 \cdot 0.8 = 0.5, \tag{1.115}$$
$$\mathrm{P}(T \mid L) = \frac{\mathrm{P}(T, L)}{\mathrm{P}(L)} \tag{1.116}$$
$$= \frac{\mathrm{P}(T, L, R) + \mathrm{P}(T, L, R^c)}{\mathrm{P}(L)} \tag{1.117}$$
$$= \frac{\mathrm{P}(T \mid R)\,\mathrm{P}(L \mid R)\,\mathrm{P}(R) + \mathrm{P}(T \mid R^c)\,\mathrm{P}(L \mid R^c)\,\mathrm{P}(R^c)}{\mathrm{P}(L)}$$
$$= \frac{0.1 \cdot 0.75 \cdot 0.2 + 0.6 \cdot 0.125 \cdot 0.8}{0.25} = 0.3. \tag{1.118}$$

$\mathrm{P}(T)$ and $\mathrm{P}(T \mid L)$ are very different, so the events are *not* independent. This makes intuitive sense. 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 when it rains). Consequently, L provides information about T: If a flight is delayed, taxis are more difficult to find because it is more likely that it rained. We conclude that 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 into our model, which represents a mechanical problem. We model the events M and R as independent, which implies that M and R^c are also independent (see Exercise 1.3),

$$\mathrm{P}(M) = \mathrm{P}(M \mid R) = \mathrm{P}(M \mid R^c). \tag{1.119}$$

In addition, we assume

$$\mathrm{P}(M) = 0.1 \quad \mathrm{P}(L \mid M) = 0.7, \quad \mathrm{P}(L \mid M^c) = 0.2, \quad \mathrm{P}(L \mid R^c, M) = 0.5.$$

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

Now, imagine that we are waiting for a flight on a sunny day, and we are wondering whether there could be a mechanical problem. The fact that it is not raining is of no use to us because M and R^c are independent. Without any further information, the probability of M is 0.1.

Suddenly, they announce that the flight is late. Now, what is the probability that there is a mechanical problem? We may be tempted to answer $\mathrm{P}(M \mid L) = 0.28$, which can be derived by the same reasoning as in Example 1.21. However, the actual conditional probability is $\mathrm{P}(M \mid L, R^c)$ because the information that it is sunny is now relevant. It implies that the rain is not responsible for the delay, so intuitively a mechanical problem should be more likely. Indeed, by the definition of conditional probability (Definition 1.16), the chain rule (Theorem 1.17) and our assumptions,

$$\begin{align}
\mathrm{P}(M \mid L, R^c) &= \frac{\mathrm{P}(L, R^c, M)}{\mathrm{P}(L, R^c)} \tag{1.120}\\
&= \frac{\mathrm{P}(L \mid R^c, M)\,\mathrm{P}(M \mid R^c)\,\mathrm{P}(R^c)}{\mathrm{P}(L \mid R^c)\,\mathrm{P}(R^c)} \tag{1.121}\\
&= \frac{\mathrm{P}(L \mid R^c, M)\,\mathrm{P}(M)}{\mathrm{P}(L \mid R^c)} \tag{1.122}\\
&= \frac{0.5 \cdot 0.1}{0.125} = 0.4, \tag{1.123}
\end{align}$$

which confirms that if the flight is late, a mechanical problem is indeed more likely when it is not raining. Formally, $\mathrm{P}(M \mid L, R^c) \neq \mathrm{P}(M \mid L)$, so the events M and R^c are *not* conditionally independent given the event L. We conclude that 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 E are not independent. Is it possible that the dependence is mainly due to political affiliation? In that case, the two events would be conditionally independent given political affiliation. To investigate this, we incorporate affiliation into our model by defining an event R, which indicates that the politician is a Republican. Conversely, 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, following Definition 1.24,

$$\begin{align}
\mathrm{P}(B, E \mid R) &= \frac{7}{155} = 0.045, \tag{1.124}\\
\mathrm{P}(B \mid R) &= \frac{22}{155} = 0.142, \tag{1.125}\\
\mathrm{P}(E \mid R) &= \frac{14}{155} = 0.090, \tag{1.126}
\end{align}$$

and verify that

$$\mathrm{P}(B \mid R)\mathrm{P}(E \mid R) = 0.013 \tag{1.127}$$

is quite different from $\mathrm{P}(B, E \mid R)$. In addition, the conditional probability

$$\mathrm{P}(B \mid R, E) = \frac{7}{14} = 0.5 \tag{1.128}$$

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

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, in Dataset 1.*

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

is very different from $\mathrm{P}(B \,|\, R)$. We conclude that the events B and E are not conditionally independent given R.

Now, let us condition on the representative being a Democrat. The empirical conditional probabilities (computed from the data on the right of Table 1.5) equal

$$\mathrm{P}(B, E \,|\, R^c) = \frac{144}{245} = 0.588, \tag{1.129}$$

$$\mathrm{P}(B \,|\, R^c) = \frac{217}{245} = 0.886, \tag{1.130}$$

$$\mathrm{P}(E \,|\, R^c) = \frac{158}{245} = 0.645, \tag{1.131}$$

so that

$$\mathrm{P}(B \,|\, R^c)\mathrm{P}(E \,|\, R^c) = 0.571 \approx \mathrm{P}(B, E \,|\, R^c). \tag{1.132}$$

Therefore, B and E are approximately conditionally independent given R^c. This is reflected in the conditional probability

$$\mathrm{P}(B \,|\, R^c, E) = \frac{144}{158} = 0.911, \tag{1.133}$$

which is 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. As illustrated by this example, conditioning on different events can completely change the dependence structure of a probabilistic model.

1.7 The Monte Carlo Method

When performing probabilistic modeling in practice, one quickly comes to a shocking realization: It is often intractable to compute the probability of some events, even if we have all the necessary information! Example 1.36 illustrates this through a probabilistic analysis of a basketball tournament. 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 enormous number of possible results. Unfortunately, such combinatorial explosions are commonplace in probabilistic modeling. The Monte Carlo method provides a pragmatic solution to this problem, inspired in our intuitive definition of probability (1.1): We *simulate* a large number of outcomes and compute the empirical probability of the event of interest.

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

The Monte Carlo method was 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, as reported in Eckhardt (1987):

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

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*,*

$$\mathrm{P}_{\mathrm{MC}}(A) := \frac{\sum_{i=1}^{n} 1(s_i \in A)}{n}, \tag{1.134}$$

where $1(s_i \in A)$ *is an indicator function that is equal to one, if* $s_i \in A$*, and to zero otherwise, for any event* $A \in \mathcal{C}$*.*

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*,*

$$\mathrm{P}_{\mathrm{MC}}(B \mid A) := \frac{\sum_{i=1}^{n} 1(s_i \in A \cap B)}{\sum_{i=1}^{n} 1(s_i \in A)}. \tag{1.135}$$

Example 1.36 (3 × 3 Olympic basketball tournament)**.** The 2020 Tokyo Olympics were the first to include 3 × 3 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 players' performance in the previous 12 months. The left column in Table 1.6 shows the total points of the four players in each team before the tournament, gathered from the official FIBA website.

We begin by using the ranking points to determine the probability of each team beating every other team. Consider two teams A and B. The higher the total sum of the ranking points

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

Table 1.6 ***Predicting the 3 × 3 basketball Olympic tournament.*** *The table shows the probability that each team wins a gold, silver, or bronze medal, or wins the group stage in the 3 × 3 basketball tournament of the 2020 Tokyo Olympics according to the model described in Example 1.36. The probabilities are estimated using* 10^4 *Monte Carlo simulations.*

		Probability of winning (%)			
Country	Ranking points	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

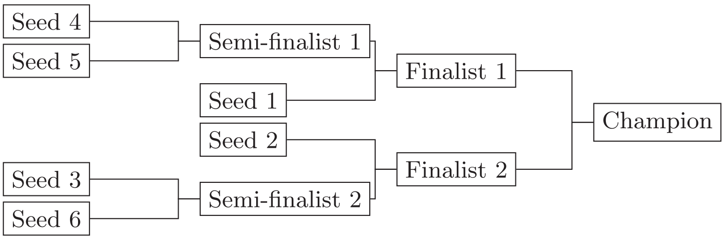

Figure 1.5 3 × 3 basketball bracket in the 2020 Tokyo Olympics. The eight participant teams were seeded according to the group stage. The first and second teams qualified directly for the semi-finals. The third, fourth, fifth, and sixth teams played the quarter-finals. The two last teams were eliminated.

of A with respect to B, the more likely A is to win a game between them. Consequently, a reasonable estimate for the probability that team A beats team B is

$$\mathrm{P}(\text{team A beats team B}) = \frac{\text{ranking points of A}}{\text{ranking points of A } + \text{ ranking points of B}}. \quad (1.136)$$

This yields, for example,

$$\mathrm{P}(\text{Belgium beats Poland}) = \frac{664381}{664381 + 654908} \quad (1.137)$$

$$= 0.504. \quad (1.138)$$

Our estimator 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 derive 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 $\binom{8}{2} = 28$ games. The results determined the seeding for a playoff bracket with six more games: The five games in Figure 1.5, 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 requires considering all 2^{34} possible results of

Table 1.7 ***Conditioning on a rare event.*** *The table shows predictions for the 3 × 3 basketball tournament in the 2020 Tokyo Olympics conditioned on the event that Serbia is eliminated in the group stage, according to the model described in Example 1.36. Each column shows the results for a different number of Monte Carlo simulations. Serbia was a heavy favorite, so they are unlikely to be eliminated, as is apparent from the last row, which shows the number of simulations where this happens.* 10^4 *total simulations are too few, but for* 10^6 *we observe enough relevant simulations for the conditional-probability estimates to be accurate.*

	Probability of gold conditioned on the event *Serbia does not reach bracket* (%)		
Country	10^4 sims	10^6 sims	10^7 sims
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
Sims where Serbia does not reach bracket	70	5,539	55,719

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 makes exact computation completely impossible. For example, March Madness (the American college basketball championship) has sixty-seven games, which results in more than 10^{20} possible results, and the Wimbledon tennis tournament or the Premier League soccer championship have even more games.

Fortunately, the Monte Carlo method enables us to approximate our probabilities of interest. Following Definition 1.35, we repeatedly simulate the tournament using the probabilities in (1.136) and then compute the fraction of outcomes for which each event of interest occurs. Table 1.6 shows the results. Our model suggests that Latvia and Serbia are heavy favorites. Out of the 10^4 simulations of the tournament, they each won the gold medal about 40% of the time. Overall, the predictions of the model are quite reasonable. In the actual tournament, Latvia ended up winning gold, beating ROC in the final. Serbia won bronze, beating Belgium in the bronze-medal game.

The accuracy of the Monte Carlo method depends on the number of simulations that we perform. For example, an event with probability 0.01 only occurs in (approximately) 1 out of every 100 simulations, so we better consider at least a few hundred, or ideally a few thousand, simulations in order to estimate its probability. In practice, it is crucial to quantify the uncertainty associated with the probability estimates obtained via the Monte Carlo method, 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 consider the problem of predicting the tournament results if Serbia happens to be eliminated in the group stage. Following Definition 1.35, we simulate the tournament 10^4 times. Then, we select the outcomes in which Serbia ends up seventh or eighth during the group stage (resulting in

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

elimination), and compute the fraction of these outcomes that are in each of the events of interest (e.g. Latvia wins the gold medal). If we don't pay attention, we could be fooled into thinking that this yields an accurate approximation. After all, we are using 10^4 simulations. However, the probabilities are estimated based exclusively on the subset of simulations in which Serbia drops out after the group stage, but this occurs only seventy times! Consequently, the estimated conditional probabilities, reported in Table 1.7, are not very precise. Increasing the number of simulations to 10^6, yields more than $5,000$ relevant instances in which Serbia is eliminated in the group stage, resulting in very different estimates. For instance, the conditional probability of Belgium winning gold drops from 10% to 6.5%. Further increasing the number simulations to 10^7 barely changes the estimates, suggesting that the approximation is accurate.

. . . .

Exercises

1.1 (Conditional probability space) Let $(\Omega, \mathcal{C}, \mathrm{P})$ be a probability space, and let A be an event in the collection $\mathcal{C}$, such that $\mathrm{P}(A) \neq 0$. As explained in Section 1.3, in order to condition on A, we define $\mathcal{C}_A$ as the collection containing the intersection of A with all the events in $\mathcal{C}$:

$$\mathcal{C}_A = \{A \cap S \colon S \in \mathcal{C}\}. \tag{1.139}$$

If we consider a new sample space $\Omega_A := A$, prove that $\mathcal{C}_A$ is a valid collection, and also that the conditional probability measure

$$\mathrm{P}_A\left(S \cap A\right) := \frac{\mathrm{P}\left(S \cap A\right)}{\mathrm{P}\left(A\right)}, \tag{1.140}$$

where $S \in \mathcal{C}$, is a valid probability measure on $\mathcal{C}_A$.

1.2 (Empirical probability measure) We have available n data points $x_1, \ldots, x_n$ taking values in a discrete set Ω. We define a probability space where the sample space is Ω and the collection of events is the power set of Ω. The probability measure is defined following Definition 1.22. For each subset $S \subseteq \Omega$,

$$\mathrm{P}(S) := \frac{1}{n}\sum_{i=1}^{n} 1(x_i \in S), \tag{1.141}$$

where $1(x_i \in S)$ is an indicator function that is equal to one if $x_i \in S$ and to zero otherwise. As an example, suppose the data are coin flips where heads are represented by 1 and tails by 0, so $\Omega := \{0, 1\}$. If $n = 10$ and the data are 6 heads and 4 tails, then

$$\mathrm{P}(\emptyset) = 0, \quad \mathrm{P}(\{1\}) = 0.6, \quad \mathrm{P}(\{0\}) = 0.4, \quad \text{and} \quad \mathrm{P}(\{0, 1\}) = 1. \tag{1.142}$$

Prove that this is a valid probability measure.

1.3 (Independence and complements) Prove that if two events A and B in the same probability space are independent, then A^c and B are also independent.

1.4 (Conditional independence and complements) If two events A and B in the same probability space are conditionally independent given another event C in the same probability space, are A and B also conditionally independent given C^c? Prove that they are or provide a counterexample.

1.5 (Partition and independence) Show that events in a partition cannot be independent. Assume that every event in the partition has nonzero probability.

1.6 (Conditional probability and complements) Let A and B be two events in the same probability space. If $\mathrm{P}\left(A \,|\, B\right) = 1$, is it true that $\mathrm{P}\left(B^c | A^c\right) = 1$?

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

1.7 (The Linda problem) In (Tversky and Kahneman, 1983) Amos Tversky and Daniel Kahneman provided the following description to a group of survey respondents:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

They then asked the respondents which of the following options they considered more likely:

- Linda is a bank teller.
- Linda is a bank teller and is active in the feminist movement.

Most respondents chose the second option. Show that this contradicts the axioms of probability.

1.8 (Quiz) A school teacher gives a weekly quiz to her students consisting of two questions. The following table shows the questions that a student answered correctly (✓) or wrong (✗):

Week	1	2	3	4	5	6	7	8	9	10
Question 1	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
Question 2	✗	✗	✓	✓	✓	✗	✗	✓	✓	✗

Use the empirical-probability estimator to estimate the conditional probability that the student gets the second question right given that (1) they got the first question right, and (2) they got the first question wrong. Does this suggest that the answers to both questions could be independent?

1.9 (Baby name) Anna is having a baby, which will be a boy or a girl with probability $1/2$. When the baby is born, she will call her aunt Margaret to tell her whether the baby is a boy or a girl. Margaret is a bit deaf; she will misunderstand and think the baby is the wrong sex with probability 0.2. Then Margaret will tell her neighbor Bob, who is also a bit deaf. He will misunderstand what Margaret says (i.e., he will think that the baby is the opposite sex of what she says) with probability 0.1.

a What is the probability that both Margaret and Bob think that the baby is a girl?
b If Bob thinks the baby is a girl, what is the probability that he is right?

1.10 (Cake) Milena is preparing a birthday cake. To finish on time she requires some help, so she asks Scott and Antonis for help. If nobody helps, she won't finish on time. If both of them help, she finishes on time. If only one helps (no matter who), she finishes on time with probability 0.5. The probability that Scott helps is 0.4. The probability that Antonis helps is 0.8. They decide to help or not independently from each other.

a What is the conditional probability that Scott helps if we know that Milena finishes on time?
b What is the conditional probability that Scott helps given that Antonis helps and Milena finishes on time? Is the event *Scott helps* conditionally independent from the event *Antonis helps* given the event *Milena finishes on time*?

1.11 (Baby sleep) A babysitter is taking care of a baby. They give her some food and put her to sleep. Assume the following:

- The probability that the food is bad is 0.1.
- If a baby eats food that is bad, they will wake up in the middle of the night. If the food is not bad, they may still wake up (with a probability that depends on whether they are good or bad sleepers).
- All babies can be classified into *good sleepers* or *bad sleepers*. The probability that a baby that is a *good sleeper* wakes up in the middle of the night is 0.1. The probability for a baby that is a *bad sleeper* is 0.8.
- A baby is a *good sleeper* with probability 0.6.

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press

Answer the following questions indicating any (reasonable) assumptions you make about independence between events.

a What is the probability that the baby wakes up in the middle of the night?

b If the baby wakes up in the middle of the night, what is the probability that the food is bad?

c Compute the conditional probability that the food is bad given that a good sleeper wakes up.

d Under our assumptions, are the events *baby is a good sleeper* and *food is bad* conditionally independent given the event *wakes up in the middle of the night*? Justify your answer mathematically and explain it intuitively.

1.12 (COVID-19 tests) A company with 10 employees decides to test them for COVID-19 before they go back to work in person. From available data, they determine that the probability of each employee being ill is 0.01. The employees have not been in contact with each other for a while, so the events *employee i is ill*, for $1 \leq i \leq 10$, are modeled as independent. If an employee is ill, the test is positive with probability 0.98. If they are not ill, the test is positive with probability 0.05.

a Is it reasonable to model the events *test i is positive*, for $1 \leq i \leq 10$, as mutually independent? From now on, model them as mutually independent whether you think it is reasonable or not, and also model the set of complements of these events as mutually independent.

b The company tests all employees. What is the probability that there is at least one positive test?

c If there is at least one positive test, what is the probability that nobody is ill? If you make any independence or conditional independence assumptions, please explain why you think they are reasonable.

1.13 (Boxing championship) In the boxing world championship, the challengers Manny and Saul will first fight one another, and the winner will fight the reigning champion Floyd. The probability that Manny beats Saul is only 0.4, but he has a higher probability of beating Floyd due to their different fighting styles. The probability that Manny and Saul beat Floyd are 0.25 and 0.1, respectively. We assume there cannot be any ties.

a If Floyd loses, what is the conditional probability that Manny becomes champion?

b Estimate the conditional probability that Manny becomes champion given that Floyd loses, from the following 20 simulations, generated according to our assumptions (F means that Floyd won, M that Manny won and S that Saul won):

Simulation	1	2	3	4	5	6	7	8	9	10
Challenger fight	S	M	S	M	M	S	M	M	S	S
Championship fight	F	F	F	M	M	F	F	F	F	F

Simulation	11	12	13	14	15	16	17	18	19	20
Challenger fight	M	M	M	S	M	S	S	S	S	S
Championship fight	F	F	F	F	M	F	F	F	F	S

Why is it not surprising that you obtain a different answer?

1.14 (Videogame) In a videogame, Silvio faces three fighters: first Honda, then Zangief, and then Blanka. He must defeat all three fighters to win the game. The probability that they defeat each fighter is 0.8 for Honda, 0.5 for Zangief and 0.4 for Blanka. If Silvio loses a fight, he faces the same fighter again, but if he loses the second fight, the game stops and he loses. We assume all fights are independent.

a What is the probability that the player wins the game?

b Use the following independent simulated fights to obtain independent simulations of the game:

Honda: L W L W W W W W L W W W L L W W W W W L
Zangief: W L L L W L W L W W L W L L W L W L W W
Blanka: L W L L L L W L W W L L L W L L L L L L

W indicates that Silvio wins the fight, L that he loses. Report the results of the simulated games, and use them to obtain a Monte Carlo estimate of the probability that Silvio wins the game. Is the result accurate? How can you improve it?

1.15 (Rare event) If we apply the Monte Carlo method to estimate the probability of an event A, how many independent simulations should we perform to make sure that the estimated probability is nonzero with probability at least 0.99? Derive the number of simulations as a function of the probability of the event $\mathrm{P}(A)$, and compute it for $\mathrm{P}(A) := 0.01$.

1.16 (Streak of heads) In this problem, we consider the problem of testing whether a randomly generated sequence is truly random. A certain computer program is supposed to generate independent fair coin flips. When you try it out, you are surprised that it contains long streaks of heads. In particular, you generate a sequence of length 200, which turns out to contain a sequence of 8 heads in a row.

a Compute the probability that the longest streak of heads that you observe has length x for $x \in \{1, 2, 3, 4, 5\}$ when you flip a fair coin 5 times, and the flips are independent.

b Estimate these probabilities using the Monte Carlo method.

c What is the estimated probability that the longest streak of heads has length 8 or more for 200 flips? Is the sequence of 8 ones evidence that the program may not be generating truly random sequences?

https://doi.org/10.1017/9781009180108.003 Published online by Cambridge University Press