1

Axioms of Probability Theory

Probability theory is the branch of mathematics that models and studies random phenomena. Although randomness has been the object of much interest over many centuries, the theory only reached maturity with *Kolmogorov's axioms*[1] in the 1930s [195].

As a mathematical theory founded on Kolmogorov's axioms, *Probability Theory* is essentially uncontroversial at this point. However, the notion of probability (i.e., chance) remains somewhat controversial. We will adopt here the frequentist notion of probability [193], which defines the chance that a particular experiment results in a given outcome as the limiting frequency of this event as the experiment is repeated an increasing number of times. The problem of giving probability a proper definition as it concerns real phenomena is discussed in [67] (with a good dose of humor).

1.1 Elements of Set Theory

Kolmogorov's formalization of probability relies on some basic notions of *Set Theory*.

A *set* is simply an abstract collection of 'objects', sometimes called *elements* or *items*. Let Ω denote such a set. A *subset* of Ω is a set made of elements that belong to Ω. In what follows, a set will be a subset of Ω.

We write $\omega \in \mathcal{A}$ when the element ω belongs to the set $\mathcal{A}$. And we write $\mathcal{A} \subset \mathcal{B}$ when set $\mathcal{A}$ is a subset of set $\mathcal{B}$. This means that $\omega \in \mathcal{A} \Rightarrow \omega \in \mathcal{B}$. A set with only one element ω is denoted $\{\omega\}$ and is called a *singleton*. Note that $\omega \in \mathcal{A} \Leftrightarrow \{\omega\} \subset \mathcal{A}$. The *empty set* is defined as a set with no elements and is denoted $\varnothing$. By convention, it is included in any other set.

Problem 1.1 Prove that $\subset$ is transitive, meaning that if $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{C}$, then $\mathcal{A} \subset \mathcal{C}$.

[1] Named after Andrey Kolmogorov (1903–1987).

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

The following are some basic set operations.

- *Intersection and disjointness* The intersection of two sets $\mathcal{A}$ and $\mathcal{B}$ is the set with all the elements belonging to both $\mathcal{A}$ and $\mathcal{B}$, and is denoted $\mathcal{A} \cap \mathcal{B}$. $\mathcal{A}$ and $\mathcal{B}$ are said to be *disjoint* if $\mathcal{A} \cap \mathcal{B} = \varnothing$.
- *Union* The union of two sets $\mathcal{A}$ and $\mathcal{B}$ is the set with elements belonging to $\mathcal{A}$ or $\mathcal{B}$, and is denoted $\mathcal{A} \cup \mathcal{B}$.
- *Set difference and complement* The set difference of $\mathcal{B}$ minus $\mathcal{A}$ is the set with elements those in $\mathcal{B}$ that are not in $\mathcal{A}$, and is denoted $\mathcal{B} \setminus \mathcal{A}$. It is sometimes called the complement of $\mathcal{A}$ in $\mathcal{B}$. The complement of $\mathcal{A}$ in the whole set Ω is often denoted $\mathcal{A}^{\mathsf{c}}$.
- *Symmetric set difference* The symmetric set difference of $\mathcal{A}$ and $\mathcal{B}$ is defined as the set with elements either in $\mathcal{A}$ or in $\mathcal{B}$, but not in both, and is denoted $\mathcal{A} \triangle \mathcal{B}$.

Sets and set operations can be visualized using a *Venn diagram*. See Figure 1.1 for an example.

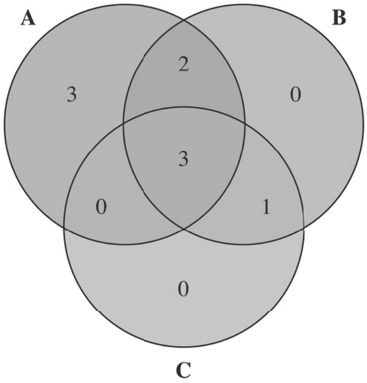

Figure 1.1 A Venn diagram helping visualize the sets $\mathcal{A} = \{1,2,4,5,6,7,8,9\}$, $\mathcal{B} = \{2,3,4,5,7,9\}$, and $\mathcal{C} = \{3,4,5,9\}$. The numbers shown in the figure represent the size of each subset. For example, the intersection of these three sets contains 3 elements, since $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C} = \{4,5,9\}$.

Problem 1.2 Prove that $\mathcal{A} \cap \varnothing = \varnothing$, $\mathcal{A} \cup \varnothing = \mathcal{A}$, and $\mathcal{A} \setminus \varnothing = \mathcal{A}$. What is $\mathcal{A} \triangle \varnothing$?

Problem 1.3 Prove that the complement is an involution, i.e., $(\mathcal{A}^{\mathsf{c}})^{\mathsf{c}} = \mathcal{A}$.

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

Problem 1.4 Show that the set difference operation is not symmetric in the sense that $\mathcal{B} \smallsetminus \mathcal{A} \neq \mathcal{A} \smallsetminus \mathcal{B}$ in general. In fact, prove that $\mathcal{B} \smallsetminus \mathcal{A} = \mathcal{A} \smallsetminus \mathcal{B}$ if and only if $\mathcal{A} = \mathcal{B} = \varnothing$.

Proposition 1.5. *The following are true:*

(i) *The intersection operation is commutative, meaning* $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$, *and associative, meaning* $(\mathcal{A} \cap \mathcal{B}) \cap \mathcal{C} = \mathcal{A} \cap (\mathcal{B} \cap \mathcal{C})$. *The same is true for the union operation.*

(ii) *The intersection operation is distributive over the union operation, meaning* $(\mathcal{A} \cup \mathcal{B}) \cap \mathcal{C} = (\mathcal{A} \cap \mathcal{C}) \cup (\mathcal{B} \cap \mathcal{C})$.

(iii) *It holds that* $(\mathcal{A} \cap \mathcal{B})^{\mathsf{c}} = \mathcal{A}^{\mathsf{c}} \cup \mathcal{B}^{\mathsf{c}}$. *More generally,* $\mathcal{C} \smallsetminus (\mathcal{A} \cap \mathcal{B}) = (\mathcal{C} \smallsetminus \mathcal{A}) \cup (\mathcal{C} \smallsetminus \mathcal{B})$.

We thus may write $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$ and $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$, that is, without parentheses, as there is no ambiguity. More generally, for a collection of sets $\{\mathcal{A}_i : i \in I\}$, where I is some index set, we can therefore refer to their intersection and union, denoted

$$\text{(intersection)} \quad \bigcap_{i \in I} \mathcal{A}_i, \qquad \text{(union)} \quad \bigcup_{i \in I} \mathcal{A}_i\ .$$

Remark 1.6 For the reader seeing these operations for the first time, it can be useful to think of $\cap$ and $\cup$ in analogy with the product $\times$ and sum $+$ operations on the integers. In that analogy, $\varnothing$ plays the role of 0.

Problem 1.7 Prove Proposition 1.5. In fact, prove the following identities:

$$\Big(\bigcup_{i \in I} \mathcal{A}_i\Big) \cap \mathcal{B} = \bigcup_{i \in I} (\mathcal{A}_i \cap \mathcal{B}),$$

and

$$\Big(\bigcup_{i \in I} \mathcal{A}_i\Big)^{\mathsf{c}} = \bigcap_{i \in I} \mathcal{A}_i^{\mathsf{c}}, \qquad \text{as well as} \qquad \Big(\bigcap_{i \in I} \mathcal{A}_i\Big)^{\mathsf{c}} = \bigcup_{i \in I} \mathcal{A}_i^{\mathsf{c}},$$

for any collection of sets $\{\mathcal{A}_i : i \in I\}$ and any set $\mathcal{B}$.

1.2 Outcomes and Events

Having introduced some elements of Set Theory, we use some of these concepts to define a probability experiment and its possible outcomes.

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

1.2.1 Outcomes and the Sample Space

In the context of an *experiment*, all the possible *outcomes* are gathered in a *sample space*, denoted Ω henceforth. In mathematical terms, the sample space is a set and the outcomes are elements of that set.

Example 1.8 (Flipping a coin) Suppose that we flip a coin three times in sequence. Assuming the coin can only land heads (H) or tails (T), the sample space Ω consists of all possible ordered sequences of length 3, which in lexicographic order can be written as

$$\Omega = \Big\{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\Big\}.$$

Example 1.9 (Drawing from an urn) Suppose that we draw two balls from an urn in sequence. Assume the urn contains red (R), green (G), and (B) blue balls. If the urn contains at least two balls of each color, or if at each trial the ball is returned to the urn, the sample space Ω consists of all possible ordered sequences of length 2, which in the RGB order can be written as

$$\Omega = \Big\{\text{RR}, \text{RG}, \text{RB}, \text{GR}, \text{GG}, \text{GB}, \text{BR}, \text{BG}, \text{BB}\Big\}. \tag{1.1}$$

If the urn (only) contains one red ball, one green ball, and two or more blue balls, and a ball drawn from the urn is not returned to the urn, the number of possible outcomes is reduced and the resulting sample space is now

$$\Omega = \Big\{\text{RG}, \text{RB}, \text{GR}, \text{GB}, \text{BR}, \text{BG}, \text{BB}\Big\}.$$

Problem 1.10 What is the sample space when we flip a coin five times? More generally, can you describe the sample space, in words and/or mathematical language, corresponding to an experiment where the coin is flipped n times? What is the size of that sample space?

Problem 1.11 Consider an experiment that consists in drawing two balls from an urn that contains red, green, blue, and yellow balls. However, yellow balls are ignored, in the sense that if such a ball is drawn then it is discarded. How does that change the sample space compared to Example 1.9?

While in the previous examples the sample space is finite, the following is an example where it is (countably) infinite.

Example 1.12 (Flipping a coin until the first heads) Consider an experiment where we flip a coin repeatedly until it lands heads. The sample space in this case is

$$\Omega = \Big\{\text{H}, \text{TH}, \text{TTH}, \text{TTTH}, \dots\Big\}.$$

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

Problem 1.13 Describe the sample space when the experiment consists in drawing repeatedly without replacement from an urn with red, green, and blue balls, three of each color, until a blue ball is drawn.

Remark 1.14 A sample space is in fact only required to contain all possible outcomes. For instance, in Example 1.9 we may always take the sample space to be (1.1) even though in the second situation that space contains outcomes that will never arise.

1.2.2 Events

Events are subsets of Ω that are of particular interest. We say that an event *happens* when the experiment results in an outcome that belongs to the event.

Example 1.15 In the context of Example 1.8, consider the event that the second toss results in heads. As a subset of the sample space, this event is defined as

$$\mathcal{E} = \Big\{\text{HHH, HHT, THH, THT}\Big\}.$$

Example 1.16 In the context of Example 1.9, consider the event that the two balls drawn from the urn are of the same color. This event corresponds to the set

$$\mathcal{E} = \Big\{\text{RR, GG, BB}\Big\}.$$

Example 1.17 In the context of Example 1.12, the event that the number of total tosses is even corresponds to the set

$$\mathcal{E} = \Big\{\text{TH, TTTH, TTTTTH}, \dots\Big\}.$$

Problem 1.18 In the context of Example 1.8, consider the event that at least two tosses result in heads. Describe this event as a set of outcomes.

1.2.3 Collection of Events

Recall that we are interested in particular subsets of the sample space Ω and that we call these ‘events’. Let Σ denote the collection of events. We assume throughout that Σ satisfies the following conditions:

- The entire sample space is an event, meaning

$$\Omega \in \Sigma. \tag{1.2}$$

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

- The complement of an event is an event, meaning

$$\mathcal{A} \in \Sigma \;\Rightarrow\; \mathcal{A}^{\mathsf{c}} \in \Sigma. \tag{1.3}$$

- A countable union of events is an event, meaning

$$\mathcal{A}_1, \mathcal{A}_2, \dots \in \Sigma \;\Rightarrow\; \bigcup_{i \geq 1} \mathcal{A}_i \in \Sigma. \tag{1.4}$$

A collection of subsets that satisfies these conditions is called a *σ-algebra*.[2]

Problem 1.19 Suppose that Σ is a σ-algebra. Show that $\varnothing \in \Sigma$ and that a countable intersection of subsets of Σ is also in Σ.

From now on, Σ will denote a σ-algebra over Ω unless otherwise specified. (Note that such a σ-algebra always exists: an example is $\{\varnothing, \Omega\}$.) The pair (Ω, Σ) is then called a *measurable space*.

Remark 1.20 (The power set) The *power set* of Ω, often denoted 2^{Ω}, is the collection of all its subsets. (Problem 1.49 provides a motivation for this name and notation.) The power set is trivially a σ-algebra. In the context of an experiment with a discrete sample space, it is customary to work with the power set as σ-algebra, because this can always be done without loss of generality (Chapter 2). When the sample space is not discrete, the situation is more complex and the σ-algebra needs to be chosen with more care (Section 3.2).

1.3 Probability Axioms

Before observing the result of an experiment, we speak of the probability that an event will happen. The Kolmogorov axioms formalize this assignment of probabilities to events. This has to be done carefully so that the resulting theory is both coherent and useful for modeling randomness.

A *probability distribution* (aka *probability measure*) on (Ω, Σ) is any real-valued function $\mathbb{P}$ defined on Σ satisfying the following properties or axioms:[3]

- *Non-negativity*

$$\mathbb{P}(\mathcal{A}) \geq 0, \quad \forall \mathcal{A} \in \Sigma.$$

- *Unit measure*

$$\mathbb{P}(\Omega) = 1.$$

[2] This refers to the algebra of sets presented in Section 1.1.

[3] Throughout, we will often use ‘distribution’ or ‘measure’ as shorthand for ‘probability distribution’.

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

- *Additivity on disjoint events* For any discrete collection of disjoint events $\{\mathcal{A}_i : i \in I\}$,

$$\mathbb{P}\Big(\bigcup_{i \in I} \mathcal{A}_i\Big) = \sum_{i \in I} \mathbb{P}(\mathcal{A}_i). \tag{1.5}$$

A triplet $(\Omega, \Sigma, \mathbb{P})$ with Ω a sample space (a set), Σ a σ-algebra over Ω, and $\mathbb{P}$ a distribution on Σ, is called a *probability space*. We consider such a triplet in what follows.

Problem 1.21 Show that $\mathbb{P}(\varnothing) = 0$ and that

$$0 \le \mathbb{P}(\mathcal{A}) \le 1, \quad \mathcal{A} \in \Sigma.$$

Thus, although nominally a probability distribution takes values in $\mathbb{R}_+$, in fact it takes values in $[0, 1]$.

Proposition 1.22 (Law of Total Probability)**.** *For any two events $\mathcal{A}$ and $\mathcal{B}$,*

$$\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{A} \cap \mathcal{B}) + \mathbb{P}(\mathcal{A} \cap \mathcal{B}^{\mathsf{c}}). \tag{1.6}$$

Problem 1.23 Prove Proposition 1.22 using the 3rd axiom.

The 3rd axiom applies to events that are disjoint. The following is a corollary that applies more generally. (In turn, this result implies the 3rd axiom.)

Proposition 1.24 (Law of Addition)**.** *For any two events $\mathcal{A}$ and $\mathcal{B}$, not necessarily disjoint,*

$$\mathbb{P}(\mathcal{A} \cup \mathcal{B}) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}). \tag{1.7}$$

In particular,

$$\mathbb{P}(\mathcal{A}^{\mathsf{c}}) = 1 - \mathbb{P}(\mathcal{A}), \tag{1.8}$$

and,

$$\mathcal{A} \subset \mathcal{B} \;\Rightarrow\; \mathbb{P}(\mathcal{B} \setminus \mathcal{A}) = \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A}). \tag{1.9}$$

Proof We first observe that we can get (1.9) from the fact that $\mathcal{B}$ is the disjoint union of $\mathcal{A}$ and $\mathcal{B} \setminus \mathcal{A}$ and an application of the 3rd axiom.

We now use this to prove (1.7). We start from the disjoint union

$$\mathcal{A} \cup \mathcal{B} = (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{B} \setminus \mathcal{A}) \cup (\mathcal{A} \cap \mathcal{B}).$$

Applying the 3rd axiom yields

$$\mathbb{P}(\mathcal{A} \cup \mathcal{B}) = \mathbb{P}(\mathcal{A} \setminus \mathcal{B}) + \mathbb{P}(\mathcal{B} \setminus \mathcal{A}) + \mathbb{P}(\mathcal{A} \cap \mathcal{B}).$$

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

Then $\mathcal{A} \setminus \mathcal{B} = \mathcal{A} \setminus (\mathcal{A} \cap \mathcal{B})$, and applying (1.9), we get

$$\mathbb{P}(\mathcal{A} \setminus \mathcal{B}) = \mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}),$$

and exchanging the roles of $\mathcal{A}$ and $\mathcal{B}$,

$$\mathbb{P}(\mathcal{B} \setminus \mathcal{A}) = \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}).$$

After some cancellations, we obtain (1.7), which then immediately implies (1.8). □

Problem 1.25 (Uniform distribution) Suppose that Ω is finite. For $\mathcal{A} \subset \Omega$, define $\mathbb{U}(\mathcal{A}) = |\mathcal{A}|/|\Omega|$, where $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A}$. Show that $\mathbb{U}$ is a probability distribution on Ω (equipped with its power set, as usual).

1.4 Inclusion-Exclusion Formula

The inclusion-exclusion formula is an expression for the probability of a discrete union of events. We start with some basic inequalities that are directly related to the inclusion-exclusion formula and useful on their own.

Boole's Inequality

Also know as the *union bound*, this inequality[4] is arguably one of the simplest, yet also one of the most useful, inequalities of Probability Theory.

Problem 1.26 (Boole's inequality) Prove that for any countable collection of events $\{\mathcal{A}_i : i \in I\}$,

$$\mathbb{P}\Big(\bigcup_{i \in I} \mathcal{A}_i\Big) \le \sum_{i \in I} \mathbb{P}(\mathcal{A}_i). \tag{1.10}$$

Note that the right-hand side can be larger than 1 or even infinite. [One possibility is to use a recursion on the number of events, together with Proposition 1.24, to prove the result for any finite number of events. Then pass to the limit to obtain the result as stated.]

Bonferroni's Inequalities

These inequalities[5] comprise Boole's inequality. For two events, we saw the Law of Addition (Proposition 1.24), which is an exact expression for the probability of their union. Consider now three events $\mathcal{A}, \mathcal{B}, \mathcal{C}$. Boole's

[4] Named after George Boole (1815–1864).

[5] Named after Carlo Emilio Bonferroni (1892–1960).

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

inequality (1.10) gives

$$\mathbb{P}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}) \leq \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{C}).$$

The following provides an inequality in the other direction.

Problem 1.27 Show that

$$\begin{aligned}\mathbb{P}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}) \geq \mathbb{P}(\mathcal{A}) &+ \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{C}) \\ &- \mathbb{P}(\mathcal{A} \cap \mathcal{B}) - \mathbb{P}(\mathcal{B} \cap \mathcal{C}) - \mathbb{P}(\mathcal{C} \cap \mathcal{A}).\end{aligned}$$

[Drawing a Venn diagram will prove useful.]

In the proof, one typically proves first the identity

$$\begin{aligned}\mathbb{P}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}) = \mathbb{P}(\mathcal{A}) &+ \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{C}) \\ &- \mathbb{P}(\mathcal{A} \cap \mathcal{B}) - \mathbb{P}(\mathcal{B} \cap \mathcal{C}) - \mathbb{P}(\mathcal{C} \cap \mathcal{A}) \\ &+ \mathbb{P}(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}),\end{aligned}$$

which generalizes the Law of Addition to three events.

Proposition 1.28 (Bonferroni's inequalities)**.** *Consider any collection of events $\mathcal{A}_1, \dots, \mathcal{A}_n$, and define*

$$S_k := \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{P}(\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_k}).$$

Then

$$\mathbb{P}(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n) \leq \sum_{j=1}^{k} (-1)^{j-1} S_j, \quad k \textit{ odd};$$

$$\mathbb{P}(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n) \geq \sum_{j=1}^{k} (-1)^{j-1} S_j, \quad k \textit{ even}.$$

Problem 1.29 Write down all of Bonferroni's inequalities for the case of four events $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4$.

Inclusion-Exclusion Formula

The last Bonferroni inequality (at $k = n$) is in fact an equality, the so-called *inclusion-exclusion formula*,

$$\mathbb{P}(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n) = \sum_{j=1}^{n} (-1)^{j-1} S_j. \tag{1.11}$$

(In particular, the last inequality in Problem 1.29 is an equality.)

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

1.5 Conditional Probability and Independence

1.5.1 Conditional Probability

Conditioning on an event $\mathcal{B}$ restricts the sample space to $\mathcal{B}$. In other words, although the experiment might yield other outcomes, conditioning on $\mathcal{B}$ focuses the attention on the outcomes that made $\mathcal{B}$ happen. In what follows we assume that $\mathbb{P}(\mathcal{B}) > 0$.

Problem 1.30 Show that $\mathbb{Q}$, defined for $\mathcal{A} \in \Sigma$ as $\mathbb{Q}(\mathcal{A}) = \mathbb{P}(\mathcal{A} \cap \mathcal{B})$, is a probability distribution if and only if $\mathbb{P}(\mathcal{B}) = 1$.

To define a bona fide probability distribution we renormalize $\mathbb{Q}$ to have total mass equal to 1 (required by the 2nd axiom) as follows:

$$\mathbb{P}(\mathcal{A} \,|\, \mathcal{B}) = \frac{\mathbb{P}(\mathcal{A} \cap \mathcal{B})}{\mathbb{P}(\mathcal{B})}, \quad \text{for } \mathcal{A} \in \Sigma.$$

We call $\mathbb{P}(\mathcal{A} \,|\, \mathcal{B})$ the *conditional probability* of $\mathcal{A}$ given $\mathcal{B}$.

Problem 1.31 Show that $\mathbb{P}(\cdot \,|\, \mathcal{B})$ is indeed a probability distribution on Ω.

Problem 1.32 In the context of Example 1.8, assume that any outcome is equally likely. Then what is the probability that the last toss lands heads if the previous tosses landed heads? Answer that same question when the coin is tossed $n \geq 2$ times, with n arbitrary and possibly large. [Regardless of n, the answer is $1/2$.]

The conclusions of Problem 1.32 may surprise some readers. And indeed, conditional probabilities can be rather unintuitive. We will come back to Problem 1.32, which is an example of the *Gambler's Fallacy*. Here is another famous example.

Example 1.33 (Monty Hall Problem) This problem is based on a television show in the US called *Let's Make a Deal* and named after its longtime presenter, Monty Hall. The following description is taken from a *New York Times* article [189]:

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the other doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to take the switch?

Not many problems in probability are discussed in the *New York Times*, to say the least. This problem is so simple to state and the answer so counter-

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

intuitive that it generated quite a controversy (read the article). The problem can mislead anyone, including professional mathematicians, let alone the layperson appearing on television!

There is an entire book on the Monty Hall Problem [154]. The textbook [84] discusses this problem among other paradoxes arising when dealing with conditional probabilities.

1.5.2 Independence

Two events $\mathcal{A}$ and $\mathcal{B}$ are said to be *independent* if knowing that $\mathcal{B}$ happens does not change the chances (i.e., the probability) that $\mathcal{A}$ happens. This is formalized by saying that the probability of $\mathcal{A}$ conditional on $\mathcal{B}$ is equal to its (unconditional) probability, or in formula,

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B}) = \mathbb{P}(\mathcal{A}). \tag{1.12}$$

The wording in English would imply a symmetric relationship, and it is indeed the case that (1.12) is equivalent to $\mathbb{P}(\mathcal{B} \mid \mathcal{A}) = \mathbb{P}(\mathcal{B})$. The following equivalent definition of independence makes the symmetry transparent.

Proposition 1.34. *Two events $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if*

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{A})\,\mathbb{P}(\mathcal{B}). \tag{1.13}$$

The identity (1.13) is often taken as a definition of independence.

Problem 1.35 Show that any event that never happens (i.e., having zero probability) is independent of any other event. In particular, $\varnothing$ is independent of any event.

Problem 1.36 Show that any event that always happens (i.e., having probability one) is independent of any other event. In particular, Ω is independent of any event.

The identity (1.13) only applies to independent events. However, it can be generalized as follows. (Note the parallel with the Law of Addition (1.7).)

Problem 1.37 (Law of Multiplication) Prove that, for any events $\mathcal{A}$ and $\mathcal{B}$,

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{A} \mid \mathcal{B})\,\mathbb{P}(\mathcal{B}). \tag{1.14}$$

Problem 1.38 (Independence and disjointness) The notions of independence and disjointness are often confused by the novice, even though they are very different. For example, show that two disjoint events are

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

independent only when at least one of them either never happens or always happens.

Problem 1.39 Combine the Law of Total Probability (1.6) and the Law of Multiplication (1.14) to get

$$\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{A} \mid \mathcal{B})\, \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{A} \mid \mathcal{B}^{\mathsf{c}})\, \mathbb{P}(\mathcal{B}^{\mathsf{c}}) \tag{1.15}$$

Problem 1.40 Suppose we draw without replacement from an urn with r red balls and b blue balls. At each stage, every ball remaining in the urn is equally likely to be picked. Use (1.15) to derive the probability of drawing a blue ball on the 3rd trial.

1.5.3 Mutual Independence

One may be interested in several events at once. Some events, $\mathcal{A}_i, i \in I$, are said to be *mutually independent* (or *jointly independent*) if

$$\mathbb{P}(\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_k}) = \mathbb{P}(\mathcal{A}_{i_1}) \times \cdots \times \mathbb{P}(\mathcal{A}_{i_k}),$$
$$\text{for any } k\text{-tuple } 1 \le i_1 < \cdots < i_k \le r.$$

They are said to be *pairwise independent* if

$$\mathbb{P}(\mathcal{A}_i \cap \mathcal{A}_j) = \mathbb{P}(\mathcal{A}_i)\, \mathbb{P}(\mathcal{A}_j), \quad \text{for all } i \neq j.$$

Obviously, mutual independence implies pairwise independence. The reverse implication is false, as the following counter-example shows.

Problem 1.41 Consider the uniform distribution on

$$\big\{(0,0,0), (0,1,1), (1,0,1), (1,1,0)\big\}.$$

Let $\mathcal{A}_i$ be the event that the ith coordinate is 1. Show that these events are pairwise independent but not mutually independent.

The following generalizes the Law of Multiplication (1.14). It is sometimes referred to as the *Chain Rule*.

Proposition 1.42 (General Law of Multiplication)**.** *For any collection of events, $\mathcal{A}_1, \dots, \mathcal{A}_r$,*

$$\mathbb{P}(\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_r) = \prod_{k=1}^{r} \mathbb{P}\big(\mathcal{A}_k \mid \mathcal{A}_1 \cap \cdots \cap \mathcal{A}_{k-1}\big). \tag{1.16}$$

For example, for any events $\mathcal{A}, \mathcal{B}, \mathcal{C}$,

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}) = \mathbb{P}(\mathcal{C} \mid \mathcal{A} \cap \mathcal{B})\, \mathbb{P}(\mathcal{B} \mid \mathcal{A})\, \mathbb{P}(\mathcal{A}).$$

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

Problem 1.43 In the same setting as Problem 1.32, show that the result of the tosses are mutually independent. That is, define $\mathcal{A}_i$ as the event that the ith toss results in heads and show that $\mathcal{A}_1, \dots, \mathcal{A}_n$ are mutually independent. In fact, show that the distribution is the uniform distribution (Problem 1.25) if and only if the tosses are fair and mutually independent.

1.5.4 Bayes Formula

The *Bayes formula*[6] can be used to "turn around" a conditional probability.

Proposition 1.44 (Bayes formula)**.** *For any two events $\mathcal{A}$ and $\mathcal{B}$,*

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B}) = \frac{\mathbb{P}(\mathcal{B} \mid \mathcal{A})\, \mathbb{P}(\mathcal{A})}{\mathbb{P}(\mathcal{B})}. \tag{1.17}$$

Proof By (1.14),

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{A} \mid \mathcal{B})\, \mathbb{P}(\mathcal{B}),$$

and also

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{B} \mid \mathcal{A})\, \mathbb{P}(\mathcal{A}),$$

which yield the result when combined. □

The denominator in (1.17) is sometimes expanded using (1.15) to get

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B}) = \frac{\mathbb{P}(\mathcal{B} \mid \mathcal{A})\, \mathbb{P}(\mathcal{A})}{\mathbb{P}(\mathcal{B} \mid \mathcal{A})\, \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B} \mid \mathcal{A}^{\mathsf{c}})\, \mathbb{P}(\mathcal{A}^{\mathsf{c}})}. \tag{1.18}$$

This form is particularly useful when $\mathbb{P}(\mathcal{B})$ is not directly available.

Problem 1.45 Suppose we draw without replacement from an urn with r red balls and b blue balls. What is the probability of drawing a blue ball on the 1st trial when drawing a blue ball on the 2nd trial?

Base Rate Fallacy

Consider a medical test for the detection of a rare disease. There are two types of mistakes that the test can make:

- *False positive* when the test is positive even though the subject does not have the disease;
- *False negative* when the test is negative even though the subject has the disease.

[6] Named after Thomas Bayes (1701–1761).

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

Let α denote the probability of a false positive; $1 - \alpha$ is sometimes called the *sensitivity*. Let β denote the probability of a false negative; $1 - \beta$ is sometimes called the *specificity*. For example, the study reported in [143] evaluates the sensitivity and specificity of several HIV tests.

Suppose that the incidence of a certain disease is π, meaning that the disease affects a proportion π of the population of interest. A person is chosen at random from the population and given the test, which turns out to be positive. What are the chances that this person actually has the disease? Ignoring the *base rate* (i.e., the disease's prevalence) would lead one to believe these chances to be $1 - \beta$. This is an example of the *Base Rate Fallacy*.

Indeed, define the events

$$\mathcal{A} = \text{'the person has the disease'},$$
$$\mathcal{B} = \text{'the test is positive'}.$$

Thus, our goal is to compute $\mathbb{P}(\mathcal{A} \mid \mathcal{B})$. Because the person was chosen at random from the population, we know that $\mathbb{P}(\mathcal{A}) = \pi$. We know the test's sensitivity, $\mathbb{P}(\mathcal{B}^c \mid \mathcal{A}^c) = 1 - \alpha$, and its specificity, $\mathbb{P}(\mathcal{B} \mid \mathcal{A}) = 1 - \beta$. Plugging this into (1.18), we get

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B}) = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}. \tag{1.19}$$

Mathematically, the Base Rate Fallacy arises from confusing $\mathbb{P}(\mathcal{A} \mid \mathcal{B})$ (which is what we want) with $\mathbb{P}(\mathcal{B} \mid \mathcal{A})$. We saw that the former depends on the latter *and* on the base rate $\mathbb{P}(\mathcal{A})$.

Problem 1.46 Show that $\mathbb{P}(\mathcal{A} \mid \mathcal{B}) = \mathbb{P}(\mathcal{B} \mid \mathcal{A})$ if and only if $\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{B})$.

Example 1.47 (Finding terrorists) In a totally different setting, Sageman [160] makes the point that a system for identifying terrorists, even if 99% accurate, cannot be ethically deployed on an entire population.

Fallacies in the Courtroom

Suppose that in a trial for murder in the US, some blood of type O- was found on the crime scene, matching the defendant's blood type. That blood type has a prevalence of about 1% in the US.[7] This leads the prosecutor to conclude that the suspect is guilty with 99% chance. But this is an example of the *Prosecutor's Fallacy*.

[7] https://redcrossblood.org/learn-about-blood/blood-types.html

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

In terms of mathematics, the error is the same as in the Base Rate Fallacy. In practice, the situation is even worse here because it is not even clear how to define the base rate. (Certainly, the base rate cannot be the unconditional probability that the defendant is guilty.)

In the same hypothetical setting, the defense could argue that, assuming the crime took place in a city with a population of about half a million, the defendant is only one among five thousand people in the region with the same blood type and that therefore the chances that he is guilty are $1/5000 = 0.02\%$. The argument is actually correct if there is no other evidence and it can be argued that the suspect was chosen more or less uniformly at random from the population. Otherwise, in particular if the latter is doubtful, this is is an example of the *Defendant's Fallacy*.

Example 1.48 *People v. Collins* is a robbery case[8] that took place in Los Angeles, California in 1968. A witness had seen a Black male with a beard and mustache together with White female with a blonde ponytail fleeing in a yellow car. The Collins (a married couple) exhibited all these attributes. The prosecutor argued that the chances of another couple matching the description were 1 in 12000000. This lead to a conviction. However, the California Supreme Court overturned the decision. This was based on the questionable computations of the base rate as well as the fact that the chances of another couple in the Los Angeles area (with a population in the millions) matching the description were much higher.

For more on the use of statistics in the courtroom, see [187].

1.6 Additional Problems

Problem 1.49 Show that if $|\Omega| = N$, then the collection of all subsets of Ω (including the empty set) has cardinality 2^N. This motivates the notation 2^Ω for this collection and also its name, as it is often called the power set of Ω.

Problem 1.50 Let $\{\Sigma_i, i \in I\}$ denote a family of σ-algebras over a set Ω. Prove that $\bigcap_{i \in I} \Sigma_i$ is also a σ-algebra over Ω.

Problem 1.51 Let $\{\mathcal{A}_i, i \in I\}$ denote a family of subsets of a set Ω. Show that there is a unique smallest (in terms of inclusion) σ-algebra over Ω that contains each of these subsets. This σ-algebra is said to be *generated* by the family $\{\mathcal{A}_i, i \in I\}$.

[8] https://courtlistener.com/opinion/1207456/people-v-collins/

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press

Problem 1.52 (General Base Rate Fallacy) Assume that the same diagnostic test is performed on m individuals to detect the presence of a certain pathogen. Due to variation in characteristics, the test performed on Individual i has sensitivity $1 - \alpha_i$ and specificity $1 - \beta_i$. Assume that a proportion π of these individuals have the pathogen. Show that (1.19) remains valid as the probability that an individual chosen uniformly at random has the pathogen given that the test is positive, with $1 - \alpha$ defined as the average sensitivity and $1 - \beta$ defined as the average specificity, meaning $\alpha = \frac{1}{m} \sum_{i=1}^{m} \alpha_i$ and $\beta = \frac{1}{m} \sum_{i=1}^{m} \beta_i$.

https://doi.org/10.1017/9781108779197.004 Published online by Cambridge University Press