1

Probability and counting

Luck. Coincidence. Randomness. Uncertainty. Risk. Doubt. Fortune. Chance. You've probably heard these words countless times, but chances are that they were used in a vague, casual way. Unfortunately, despite its ubiquity in science and everyday life, probability can be deeply counterintuitive. If we rely on intuitions of doubtful validity, we run a serious risk of making inaccurate predictions or overconfident decisions. The goal of this book is to introduce probability as a logical framework for quantifying uncertainty and randomness in a principled way. We'll also aim to strengthen intuition, both when our initial guesses coincide with logical reasoning and when we're not so lucky.

1.1 Why study probability?

Mathematics is the logic of certainty; probability is the logic of uncertainty. Probability is extremely useful in a wide variety of fields, since it provides tools for understanding and explaining variation, separating signal from noise, and modeling complex phenomena. To give just a small sample from a continually growing list of applications:

1. *Statistics*: Probability is the foundation and language for statistics, enabling many powerful methods for using data to learn about the world.

2. *Physics*: Einstein famously said "God does not play dice with the universe", but current understanding of quantum physics heavily involves probability at the most fundamental level of nature. Statistical mechanics is another major branch of physics that is built on probability.

3. *Biology*: Genetics is deeply intertwined with probability, both in the inheritance of genes and in modeling random mutations.

4. *Computer science*: Randomized algorithms make random choices while they are run, and in many important applications they are simpler and more efficient than any currently known deterministic alternatives. Probability also plays an essential role in studying the performance of algorithms, and in machine learning and artificial intelligence.

5. *Meteorology*: Weather forecasts are (or should be) computed and expressed in terms of probability.

6. *Gambling*: Many of the earliest investigations of probability were aimed at answering questions about gambling and games of chance.

7. *Finance*: At the risk of redundancy with the previous example, it should be pointed out that probability is central in quantitative finance. Modeling stock prices over time and determining "fair" prices for financial instruments are based heavily on probability.

8. *Political science*: In recent years, political science has become more and more quantitative and statistical, with applications such as analyzing surveys of public opinion, assessing gerrymandering, and predicting elections.

9. *Medicine*: The development of randomized clinical trials, in which patients are randomly assigned to receive treatment or placebo, has transformed medical research in recent years. As the biostatistician David Harrington remarked, "Some have conjectured that it could be the most significant advance in scientific medicine in the twentieth century.... In one of the delightful ironies of modern science, the randomized trial 'adjusts' for both observed and unobserved heterogeneity in a controlled experiment by introducing chance variation into the study design." [16]

10. *Life*: Life is uncertain, and probability is the logic of uncertainty. While it isn't practical to carry out a formal probability calculation for every decision made in life, thinking hard about probability can help us avert some common fallacies, shed light on coincidences, and make better predictions.

Probability provides procedures for principled problem-solving, but it can also produce pitfalls and paradoxes. For example, we'll see in this chapter that even Gottfried Wilhelm von Leibniz and Sir Isaac Newton, the two people who independently discovered calculus in the 17th century, were not immune to basic errors in probability. Throughout this book, we will use the following strategies to help avoid potential pitfalls.

1. *Simulation*: A beautiful aspect of probability is that it is often possible to study problems via *simulation*. Rather than endlessly debating an answer with someone who disagrees with you, you can run a simulation and see empirically who is right. Each chapter in this book ends with a section that gives examples of how to do calculations and simulations in R, a free statistical computing environment.

2. *Biohazards*: Studying common mistakes is important for gaining a stronger understanding of what is and is not valid reasoning in probability. In this book, common mistakes are called *biohazards* and are denoted by ☣ (since making such mistakes can be hazardous to one's health!).

3. *Sanity checks*: After solving a problem one way, we will often try to solve the same problem in a different way or to examine whether our answer makes sense in simple and extreme cases.

1.2 Sample spaces and Pebble World

The mathematical framework for probability is built around *sets*. Imagine that an experiment is performed, resulting in one out of a set of possible outcomes. Before the experiment is performed, it is unknown which outcome will be the result; after, the result "crystallizes" into the actual outcome.

Definition 1.2.1 (Sample space and event)**.** The *sample space* S of an experiment is the set of all possible outcomes of the experiment. An *event* A is a subset of the sample space S, and we say that A *occurred* if the actual outcome is in A.

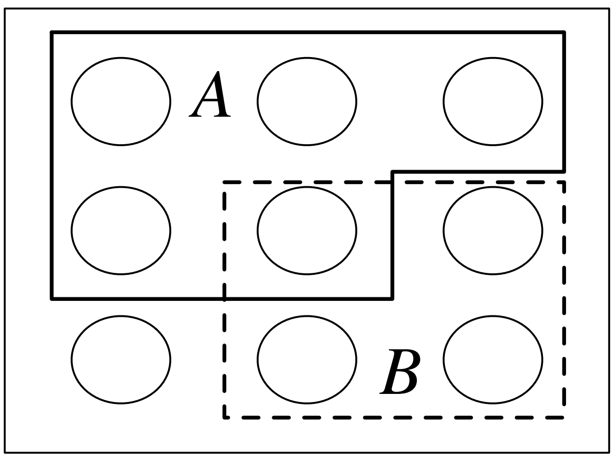

FIGURE 1.1
A sample space as Pebble World, with two events A and B spotlighted.

The sample space of an experiment can be finite, countably infinite, or uncountably infinite (see Section A.1.5 of the math appendix for an explanation of countable and uncountable sets). When the sample space is finite, we can visualize it as *Pebble World*, as shown in Figure 1.1. Each pebble represents an outcome, and an event is a set of pebbles.

Performing the experiment amounts to randomly selecting one pebble. If all the pebbles are of the same mass, all the pebbles are equally likely to be chosen. This special case is the topic of the next two sections. In Section 1.6, we give a general definition of probability that allows the pebbles to differ in mass.

Set theory is very useful in probability, since it provides a rich language for expressing and working with events; Section A.1 of the math appendix provides a review of set theory. Set operations, especially unions, intersections, and complements, make it easy to build new events in terms of already-defined events. These concepts also let us express an event in more than one way; often, one expression for an event is much easier to work with than another expression for the same event.

For example, let S be the sample space of an experiment and let $A, B \subseteq S$ be events. Then the union $A \cup B$ is the event that occurs if and only if *at least one* of A, B occurs, the intersection $A \cap B$ is the event that occurs if and only if *both* A and B occur, and the complement A^c is the event that occurs if and only if A does *not* occur. We also have *De Morgan's laws*:

$$(A \cup B)^c = A^c \cap B^c \text{ and } (A \cap B)^c = A^c \cup B^c,$$

since saying that it is *not* the case that at least one of A and B occur is the same as saying that A does not occur and B does not occur, and saying that it is *not* the case that both occur is the same as saying that at least one does not occur. Analogous results hold for unions and intersections of more than two events.

In the example shown in Figure 1.1, A is a set of 5 pebbles, B is a set of 4 pebbles, $A \cup B$ consists of the 8 pebbles in A or B (including the pebble that is in both), $A \cap B$ consists of the pebble that is in both A and B, and A^c consists of the 4 pebbles that are not in A.

The notion of sample space is very general and abstract, so it is important to have some concrete examples in mind.

Example 1.2.2 (Coin flips)**.** A coin is flipped 10 times. Writing Heads as H and Tails as T, a possible outcome (pebble) is $HHHTHHTTHT$, and the sample space is the set of all possible strings of length 10 of H's and T's. We can (and will) encode H as 1 and T as 0, so that an outcome is a sequence $(s_1, \dots, s_{10})$ with $s_j \in \{0, 1\}$, and the sample space is the set of all such sequences. Now let's look at some events:

1. Let A_1 be the event that the first flip is Heads. As a set,

$$A_1 = \{(1, s_2, \dots, s_{10}) : s_j \in \{0, 1\} \text{ for } 2 \leq j \leq 10\}.$$

This is a subset of the sample space, so it is indeed an event; saying that A_1 occurs is the same thing as saying that the first flip is Heads. Similarly, let A_j be the event that the jth flip is Heads for $j = 2, 3, \dots, 10$.

2. Let B be the event that at least one flip was Heads. As a set,

$$B = \bigcup_{j=1}^{10} A_j.$$

3. Let C be the event that all the flips were Heads. As a set,

$$C = \bigcap_{j=1}^{10} A_j.$$