Introduction to Probability: Excerpt from Chapters 1–2

Joseph K. Blitzstein and Jessica Hwang
Harvard University and Stanford University

Chapman & Hall/CRC Press

Contents

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. For example, Nate Silver's successes in predicting election results, such as in the 2008 and 2012 U.S. presidential elections, were achieved using probability models to make sense of polls and to drive simulations (see Silver [25]).

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." [17]

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.

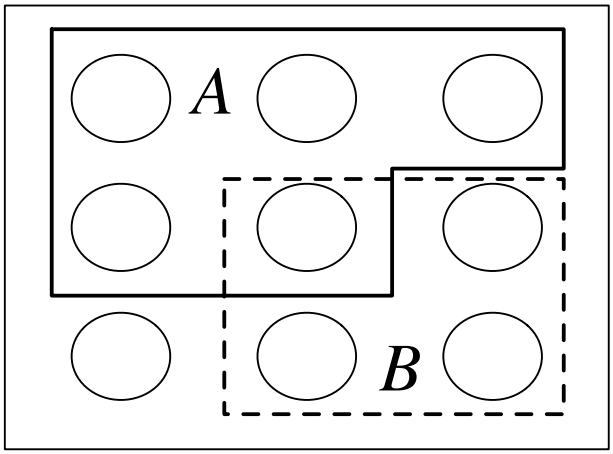


FIGURE 1.1
A sample space as Pebble World, with two events A and B spotlighted.

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 and 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 $(s_1, \ldots, s_{10})$ 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,

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

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 $j = 2, 3, \ldots, 10$.

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,

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

4. Let D be the event that there were at least two consecutive Heads. As a set,

$$D = \bigcup_{j=1}^{9} (A_j \cap A_{j+1}).$$

□

Example 1.2.3 (Pick a card, any card)**.** Pick a card from a standard deck of 52 cards. The sample space S is the set of all 52 cards (so there are 52 pebbles, one for each card). Consider the following four events:

- A: card is an ace.
- B: card has a black suit.
- D: card is a diamond.
- H: card is a heart.

As a set, H consists of 13 cards:

$$\{\text{Ace of Hearts, Two of Hearts, } \ldots, \text{ King of Hearts}\}.$$

We can create various other events in terms of A, B, D, H. For example, $A \cap H$ is the event that the card is the Ace of Hearts, $A \cap B$ is the event {Ace of Spades, Ace of Clubs}, and $A \cup D \cup H$ is the event that the card is red or an ace. Also, note that $(D \cup H)^c = D^c \cap H^c = B$, so B can be expressed in terms of D and H. On the other hand, the event that the card is a spade can't be written in terms of A, B, D, H since none of them are fine-grained enough to be able to distinguish between spades and clubs.

There are *many* other events that could be defined using this sample space. In fact, the counting methods introduced later in this chapter show that there are $2^{52} \approx 4.5 \times 10^{15}$ events in this problem, even though there are only 52 pebbles.

What if the card drawn were a joker? That would indicate that we had the wrong sample space; we are assuming that the outcome of the experiment is guaranteed to be an element of S. □

As the preceding examples demonstrate, events can be described in English or in set notation. Sometimes the English description is easier to interpret while the set notation is easier to manipulate. Let S be a sample space and s_{actual} be the actual outcome of the experiment (the pebble that ends up getting chosen when the experiment is performed). A mini-dictionary for converting between English and

sets is shown below. For example, for events A and B, the English statement "A implies B" says that whenever the event A occurs, the event B also occurs; in terms of sets, this translates into saying that A is a subset of B.

English	**Sets**
Events and occurrences	
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
something must happen	$s_{\text{actual}} \in S$
New events from old events	
A or B (inclusive)	$A \cup B$
A and B	$A \cap B$
not A	A^c
A or B, but not both	$(A \cap B^c) \cup (A^c \cap B)$
at least one of $A_1, \ldots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \cdots \cap A_n$
Relationships between events	
A implies B	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \emptyset$
$A_1, \ldots, A_n$ are a partition of S	$A_1 \cup \cdots \cup A_n = S, A_i \cap A_j = \emptyset$ for $i \neq j$

1.3 Naive definition of probability

Historically, the earliest definition of the probability of an event was to count the number of ways the event could happen and divide by the total number of possible outcomes for the experiment. We call this the *naive definition* since it is restrictive and relies on strong assumptions; nevertheless, it is important to understand, and useful when not misused.

Definition 1.3.1 (Naive definition of probability)**.** Let A be an event for an experiment with a finite sample space S. The *naive probability* of A is

$$P_{\text{naive}}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to } A}{\text{total number of outcomes in } S}.$$

(We use $|A|$ to denote the size of A; see Section A.1.5 of the math appendix.)

In terms of Pebble World, the naive definition just says that the probability of A is the fraction of pebbles that are in A. For example, in Figure 1.1 it says

$$P_{\text{naive}}(A) = \frac{5}{9},\ P_{\text{naive}}(B) = \frac{4}{9},\ P_{\text{naive}}(A \cup B) = \frac{8}{9},\ P_{\text{naive}}(A \cap B) = \frac{1}{9}.$$

For the complements of the events just considered,

$$P_{\text{naive}}(A^c) = \frac{4}{9},\ P_{\text{naive}}(B^c) = \frac{5}{9},\ P_{\text{naive}}((A \cup B)^c) = \frac{1}{9},\ P_{\text{naive}}((A \cap B)^c) = \frac{8}{9}.$$

In general,

$$P_{\text{naive}}(A^c) = \frac{|A^c|}{|S|} = \frac{|S| - |A|}{|S|} = 1 - \frac{|A|}{|S|} = 1 - P_{\text{naive}}(A).$$

In Section 1.6, we will see that this result about complements *always* holds for probability, even when we go beyond the naive definition. A good strategy when trying to find the probability of an event is to start by thinking about whether it will be easier to find the probability of the event or the probability of its complement. De Morgan's laws are especially useful in this context, since it may be easier to work with an intersection than a union, or vice versa.

The naive definition is very restrictive in that it requires S to be finite, with equal mass for each pebble. It has often been misapplied by people who assume equally likely outcomes without justification and make arguments to the effect of "either it will happen or it won't, and we don't know which, so it's 50-50". In addition to sometimes giving absurd probabilities, this type of reasoning isn't even internally consistent. For example, it would say that the probability of life on Mars is 1/2 ("either there is or there isn't life there"), but it would also say that the probability of *intelligent* life on Mars is 1/2, and it is clear intuitively—and by the properties of probability developed in Section 1.6—that the latter should have strictly lower probability than the former. But there are several important types of problems where the naive definition *is* applicable:

- when there is *symmetry* in the problem that makes outcomes equally likely. It is common to assume that a coin has a 50% chance of landing Heads when tossed, due to the physical symmetry of the coin.[1] For a standard, well-shuffled deck of cards, it is reasonable to assume that all orders are equally likely. There aren't certain overeager cards that especially like to be near the top of the deck; any particular location in the deck is equally likely to house any of the 52 cards.
- when the outcomes are equally likely *by design*. For example, consider conducting a survey of n people in a population of N people. A common goal is to obtain a

[1] See Diaconis, Holmes, and Montgomery [8] for a physical argument that the chance of a tossed coin coming up the way it started is about 0.51 (close to but slightly more than 1/2), and Gelman and Nolan [12] for an explanation of why the probability of Heads is close to 1/2 even for a coin that is manufactured to have different weights on the two sides (for standard coin-tossing; allowing the coin to spin is a different matter).

simple random sample, which means that the n people are chosen randomly with all subsets of size n being equally likely. If successful, this ensures that the naive definition is applicable, but in practice this may be hard to accomplish because of various complications, such as not having a complete, accurate list of contact information for everyone in the population.

- when the naive definition serves as a useful *null model*. In this setting, we *assume* that the naive definition applies just to see what predictions it would yield, and then we can compare observed data with predicted values to assess whether the hypothesis of equally likely outcomes is tenable.

1.4 How to count

Calculating the naive probability of an event A involves counting the number of pebbles in A and the number of pebbles in the sample space S. Often the sets we need to count are extremely large. This section introduces some fundamental methods for counting; further methods can be found in books on *combinatorics*, the branch of mathematics that studies counting.

1.4.1 Multiplication rule

In some problems, we can directly count the number of possibilities using a basic but versatile principle called the *multiplication rule*. We'll see that the multiplication rule leads naturally to counting rules for *sampling with replacement* and *sampling without replacement*, two scenarios that often arise in probability and statistics.

Theorem 1.4.1 (Multiplication rule)**.** Consider a compound experiment consisting of two sub-experiments, Experiment A and Experiment B. Suppose that Experiment A has a possible outcomes, and for each of those outcomes Experiment B has b possible outcomes. Then the compound experiment has ab possible outcomes.

To see why the multiplication rule is true, imagine a tree diagram as in Figure 1.2. Let the tree branch a ways according to the possibilities for Experiment A, and for each of those branches create b further branches for Experiment B. Overall, there are $\underbrace{b + b + \cdots + b}_{a} = ab$ possibilities.

☣ **1.4.2.** It is often easier to think about the experiments as being in chronological order, but there is no requirement in the multiplication rule that Experiment A has to be performed before Experiment B.

Example 1.4.3 (Ice cream cones)**.** Suppose you are buying an ice cream cone. You can choose whether to have a cake cone or a waffle cone, and whether to

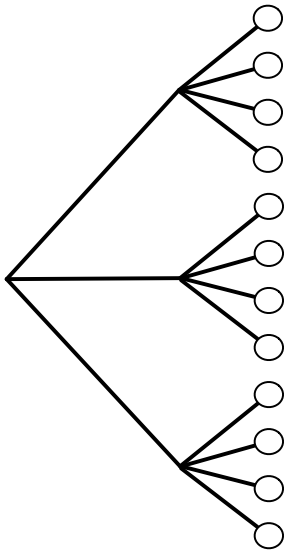


FIGURE 1.2
Tree diagram illustrating the multiplication rule. If Experiment A has 3 possible outcomes and Experiment B has 4 possible outcomes, then overall there are $3 \cdot 4 = 12$ possible outcomes.

have chocolate, vanilla, or strawberry as your flavor. This decision process can be visualized with a tree diagram, as in Figure 1.3.

By the multiplication rule, there are $2 \cdot 3 = 6$ possibilities. This is a very simple example, but is worth thinking through in detail as a foundation for thinking about and visualizing more complicated examples. Soon we will encounter examples where drawing the tree in a legible size would take up more space than exists in the known universe, yet where conceptually we can still think in terms of the ice cream example. Some things to note:

1. It doesn't matter whether you choose the type of cone first ("I'd like a waffle cone with chocolate ice cream") or the flavor first ("I'd like chocolate ice cream on a waffle cone"). Either way, there are $2 \cdot 3 = 3 \cdot 2 = 6$ possibilities.

2. It doesn't matter whether the same flavors are available on a cake cone as on a waffle cone. What matters is that there are exactly 3 flavor choices for each cone choice. If for some strange reason it were forbidden to have chocolate ice cream on a waffle cone, with no substitute flavor available (aside from vanilla and strawberry), there would be $3 + 2 = 5$ possibilities and the multiplication rule wouldn't apply. In larger examples, such complications could make counting the number of possibilities vastly more difficult.

Now suppose you buy *two* ice cream cones on a certain day, one in the afternoon and the other in the evening. Write, for example, (cakeC, waffleV) to mean a cake cone with chocolate in the afternoon, followed by a waffle cone with vanilla in the

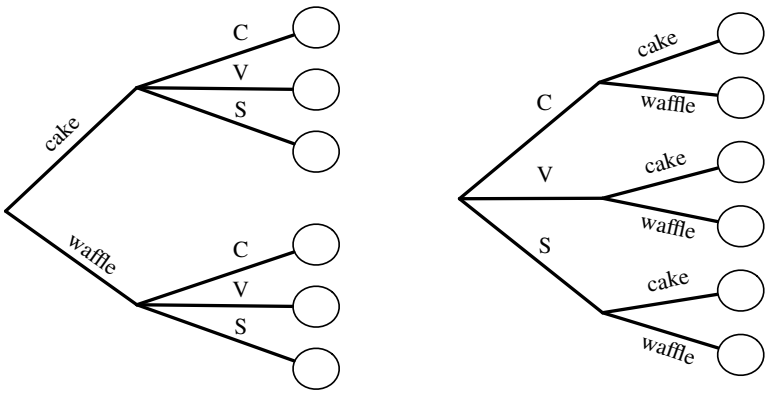


FIGURE 1.3
Tree diagram for choosing an ice cream cone. Regardless of whether the type of cone or the flavor is chosen first, there are $2 \cdot 3 = 3 \cdot 2 = 6$ possibilities.

evening. By the multiplication rule, there are $6^2 = 36$ possibilities in your delicious compound experiment.

But what if you're only interested in what kinds of ice cream cones you had that day, not the order in which you had them, so you don't want to distinguish, for example, between (cakeC, waffleV) and (waffleV, cakeC)? Are there now $36/2 = 18$ possibilities? No, since possibilities like (cakeC, cakeC) were already only listed once each. There are $6 \cdot 5 = 30$ ordered possibilities (x, y) with $x \neq y$, which turn into 15 possibilities if we treat (x, y) as equivalent to (y, x), plus 6 possibilities of the form (x, x), giving a total of 21 possibilities. Note that if the 36 original ordered pairs (x, y) are equally likely, then the 21 possibilities here are *not* equally likely. □

Example 1.4.4 (Subsets)**.** A set with n elements has 2^n subsets, including the empty set $\emptyset$ and the set itself. This follows from the multiplication rule since for each element, we can choose whether to include it or exclude it. For example, the set $\{1, 2, 3\}$ has the 8 subsets $\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$. This result explains why in Example 1.2.3 there are $2^{52} \approx 4.5 \times 10^{15}$ events that can be defined. □

We can use the multiplication rule to arrive at formulas for sampling with and without replacement. Many experiments in probability and statistics can be interpreted in one of these two contexts, so it is appealing that both formulas follow directly from the same basic counting principle.

Theorem 1.4.5 (Sampling with replacement)**.** Consider n objects and making k choices from them, one at a time *with replacement* (i.e., choosing a certain object does not preclude it from being chosen again). Then there are n^k possible outcomes.

For example, imagine a jar with n balls, labeled from 1 to n. We sample balls one at a time with replacement, meaning that each time a ball is chosen, it is returned to the jar. Each sampled ball is a sub-experiment with n possible outcomes, and

there are k sub-experiments. Thus, by the multiplication rule there are n^k ways to obtain a sample of size k.

Theorem 1.4.6 (Sampling without replacement)**.** Consider n objects and making k choices from them, one at a time *without replacement* (i.e., choosing a certain object precludes it from being chosen again). Then there are $n(n-1)\cdots(n-k+1)$ possible outcomes, for $k \leq n$ (and 0 possibilities for $k > n$).

This result also follows directly from the multiplication rule: each sampled ball is again a sub-experiment, and the number of possible outcomes decreases by 1 each time. Note that for sampling k out of n objects without replacement, we need $k \leq n$, whereas in sampling with replacement the objects are inexhaustible.

Example 1.4.7 (Permutations and factorials)**.** A *permutation* of $1, 2, \ldots, n$ is an arrangement of them in some order, e.g., $3, 5, 1, 2, 4$ is a permutation of $1, 2, 3, 4, 5$. By Theorem 1.4.6 with $k = n$, there are $n!$ permutations of $1, 2, \ldots, n$. For example, there are $n!$ ways in which n people can line up for ice cream. (Recall that $n! = n(n-1)(n-2)\cdots 1$ for any positive integer n, and $0! = 1$.) □

Theorems 1.4.5 and 1.4.6 are theorems about *counting*, but when the naive definition applies, we can use them to calculate *probabilities*. This brings us to our next example, a famous problem in probability called the *birthday problem*. The solution incorporates both sampling with replacement and sampling without replacement.

Example 1.4.8 (Birthday problem)**.** There are k people in a room. Assume each person's birthday is equally likely to be any of the 365 days of the year (we exclude February 29), and that people's birthdays are independent (we assume there are no twins in the room). What is the probability that two or more people in the group have the same birthday?

Solution:

There are 365^k ways to assign birthdays to the people in the room, since we can imagine the 365 days of the year being sampled k times, with replacement. By assumption, all of these possibilities are equally likely, so the naive definition of probability applies.

Used directly, the naive definition says we just need to count the number of ways to assign birthdays to k people such that there are two or more people who share a birthday. But this counting problem is hard, since it could be Emma and Steve who share a birthday, or Steve and Naomi, or all three of them, or the three of them could share a birthday while two others in the group share a different birthday, or various other possibilities.

Instead, let's count the complement: the number of ways to assign birthdays to k people such that no two people share a birthday. This amounts to sampling the 365 days of the year *without* replacement, so the number of possibilities is $365 \cdot 364 \cdot 363 \cdots (365-k+1)$ for $k \leq 365$. Therefore the probability of no birthday

matches in a group of k people is

$$P(\text{no birthday match}) = \frac{365 \cdot 364 \cdots (365-k+1)}{365^k},$$

and the probability of at least one birthday match is

$$P(\text{at least 1 birthday match}) = 1 - \frac{365 \cdot 364 \cdots (365-k+1)}{365^k}.$$

Figure 1.4 plots the probability of at least one birthday match as a function of k. The first value of k for which the probability of a match exceeds 0.5 is $k = 23$. Thus, in a group of 23 people, there is a better than 50% chance that two or more of them will have the same birthday. By the time we reach $k = 57$, the probability of a match exceeds 99%.

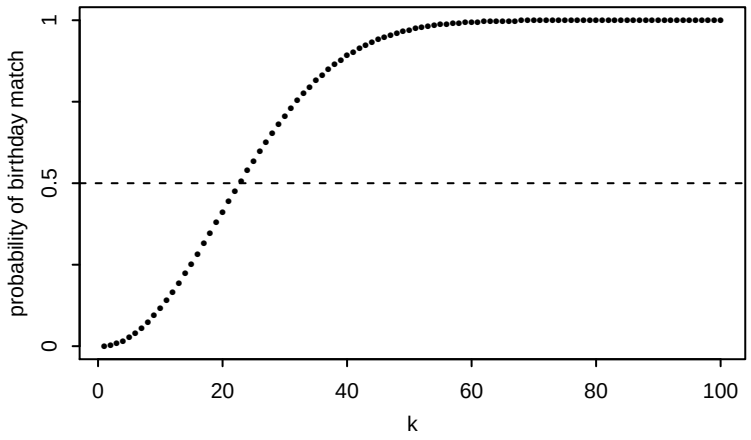


FIGURE 1.4
Probability that in a room of k people, at least two were born on the same day. This probability first exceeds 0.5 when $k = 23$.

Of course, for $k = 366$ we are *guaranteed* to have a match, but it's surprising that even with a much smaller number of people it's overwhelmingly likely that there is a birthday match. For a quick intuition into why it should not be so surprising, note that with 23 people there are $\binom{23}{2} = 253$ *pairs* of people, any of which could be a birthday match.

The birthday problem is much more than a fun party game, and much more than a way to build intuition about coincidences; there are also important applications in statistics and computer science. One of the exercises explores the more general setting in which the probability is not necessarily 1/365 for each day. It turns out that in the non-equal probability case, having at least one match becomes even *more* likely. □

☣ **1.4.9** (Labeling objects)**.** Drawing a sample from a population is a very fundamental concept in statistics. It is important to think of the objects or people in

the population as *named* or *labeled*. For example, if there are n balls in a jar, we can imagine that they have labels from 1 to n, even if the balls look the same to the human eye. In the birthday problem, we can give each person an ID (identification) number, rather than thinking of the people as indistinguishable particles or a faceless mob.

A related example is an instructive blunder made by Leibniz in a seemingly simple problem (see Gorroochurn [15] for discussion of this and a variety of other probability problems from a historical perspective).

Example 1.4.10 (Leibniz's mistake)**.** If we roll two fair dice, which is more likely: a sum of 11 or a sum of 12?

Solution:

Label the dice A and B, and consider each die to be a sub-experiment. By the multiplication rule, there are 36 possible outcomes for ordered pairs of the form (value of A, value of B), and they are equally likely by symmetry. Of these, $(5, 6)$ and $(6, 5)$ are favorable to a sum of 11, while only $(6, 6)$ is favorable to a sum of 12. Therefore a sum of 11 is twice as likely as a sum of 12; the probability is $1/18$ for the former, and $1/36$ for the latter.

However, Leibniz wrongly argued that a sum of 11 and a sum of 12 are equally likely. He claimed that "it is equally likely to throw twelve points, than to throw eleven; because one or the other can be done in only one manner". Here Leibniz was making the mistake of treating the two dice as indistinguishable objects, viewing $(5, 6)$ and $(6, 5)$ as the same outcome.

What are the antidotes to Leibniz's mistake? First, as explained in ☣ 1.4.9, we should *label* the objects in question instead of treating them as indistinguishable. If Leibniz had labeled his dice A and B, or green and orange, or left and right, he would not have made this mistake. Second, before we use counting for probability, we should ask ourselves whether the naive definition applies (see ☣ 1.4.21 for another example showing that caution is needed before applying the naive definition). □

1.4.2 Adjusting for overcounting

In many counting problems, it is not easy to directly count each possibility once and only once. If, however, we are able to count each possibility exactly c times for some c, then we can adjust by dividing by c. For example, if we have exactly double-counted each possibility, we can divide by 2 to get the correct count. We call this *adjusting for overcounting*.

Example 1.4.11 (Committees and teams)**.** Consider a group of four people.

(a) How many ways are there to choose a two-person committee?

(b) How many ways are there to break the people into two teams of two?

Solution:

(a) One way to count the possibilities is by listing them out: labeling the people as 1, 2, 3, 4, the possibilities are $\boxed{12}$, $\boxed{13}$, $\boxed{14}$, $\boxed{23}$, $\boxed{24}$, $\boxed{34}$.

Another approach is to use the multiplication rule with an adjustment for overcounting. By the multiplication rule, there are 4 ways to choose the first person on the committee and 3 ways to choose the second person on the committee, but this counts each possibility twice, since picking 1 and 2 to be on the committee is the same as picking 2 and 1 to be on the committee. Since we have overcounted by a factor of 2, the number of possibilities is $(4 \cdot 3)/2 = 6$.

(b) Here are 3 ways to see that there are 3 ways to form the teams. Labeling the people as $1, 2, 3, 4$, we can directly list out the possibilities: $\boxed{12}\boxed{34}$, $\boxed{13}\boxed{24}$, and $\boxed{14}\boxed{23}$. Listing out all possibilities would quickly become tedious or infeasible with more people though. Another approach is to note that it suffices to specify person 1's teammate (and then the other team is determined). A third way is to use (a) to see that there are 6 ways to choose one team. This overcounts by a factor of 2, since picking 1 and 2 to be a team is equivalent to picking 3 and 4 to be a team. So again the answer is $6/2 = 3$. □

A *binomial coefficient* counts the number of subsets of a certain size for a set, such as the number of ways to choose a committee of size k from a set of n people. Sets and subsets are by definition *unordered*, e.g., $\{3, 1, 4\} = \{4, 1, 3\}$, so we are counting the number of ways to choose k objects out of n, without replacement and without distinguishing between the different orders in which they could be chosen.

Definition 1.4.12 (Binomial coefficient)**.** For any nonnegative integers k and n, the *binomial coefficient* $\binom{n}{k}$, read as "n choose k", is the number of subsets of size k for a set of size n.

For example, $\binom{4}{2} = 6$, as shown in Example 1.4.11. The binomial coefficient $\binom{n}{k}$ is sometimes called a *combination*, but we do not use that terminology here since "combination" is such a useful general-purpose word. Algebraically, binomial coefficients can be computed as follows.

Theorem 1.4.13 (Binomial coefficient formula)**.** For $k \leq n$, we have

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}.$$

For $k > n$, we have $\binom{n}{k} = 0$.

Proof. Let A be a set with $|A| = n$. Any subset of A has size at most n, so $\binom{n}{k} = 0$ for $k > n$. Now let $k \leq n$. By Theorem 1.4.6, there are $n(n-1)\cdots(n-k+1)$ ways to make an *ordered* choice of k elements without replacement. This overcounts each subset of interest by a factor of $k!$ (since we don't care how these elements are ordered), so we can get the correct count by dividing by $k!$. ■

☣ **1.4.14.** The binomial coefficient $\binom{n}{k}$ is often defined in terms of factorials, but keep in mind that $\binom{n}{k}$ is 0 if $k > n$, even though the factorial of a negative number is undefined. Also, the middle expression in Theorem 1.4.13 is often better for computation than the expression with factorials, since factorials grow *extremely* fast. For example,

$$\binom{100}{2} = \frac{100 \cdot 99}{2} = 4950$$

can even be done by hand, whereas computing $\binom{100}{2} = 100!/(98! \cdot 2!)$ by first calculating 100! and 98! would be wasteful and possibly dangerous because of the extremely large numbers involved ($100! \approx 9.33 \times 10^{157}$).

Example 1.4.15 (Club officers)**.** In a club with n people, there are $n(n-1)(n-2)$ ways to choose a president, vice president, and treasurer, and there are $\binom{n}{3} = \frac{n(n-1)(n-2)}{3!}$ ways to choose 3 officers without predetermined titles. □

Example 1.4.16 (Permutations of a word)**.** How many ways are there to permute the letters in the word LALALAAA? To determine a permutation, we just need to choose where the 5 A's go (or, equivalently, just decide where the 3 L's go). So there are

$$\binom{8}{5} = \binom{8}{3} = \frac{8 \cdot 7 \cdot 6}{3!} = 56 \text{ permutations.}$$

How many ways are there to permute the letters in the word STATISTICS? Here are two approaches. We could choose where to put the S's, then where to put the T's (from the remaining positions), then where to put the I's, then where to put the A (and then the C is determined). Alternatively, we can start with 10! and then adjust for overcounting, dividing by 3!3!2! to account for the fact that the S's can be permuted among themselves in any way, and likewise for the T's and I's. This gives

$$\binom{10}{3}\binom{7}{3}\binom{4}{2}\binom{2}{1} = \frac{10!}{3!3!2!} = 50400 \text{ possibilities.}$$

□

Example 1.4.17 (Binomial theorem)**.** The *binomial theorem* states that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

To prove the binomial theorem, expand out the product

$$\underbrace{(x+y)(x+y)\ldots(x+y)}_{n \text{ factors}}.$$

Just as $(a+b)(c+d) = ac+ad+bc+bd$ is the sum of terms where we pick the a or the b from the first factor (but not both) and the c or the d from the second factor (but not both), the terms of $(x+y)^n$ are obtained by picking either the x or the y (but not both) from each factor. There are $\binom{n}{k}$ ways to choose exactly k of the x's, and each such choice yields the term $x^k y^{n-k}$. The binomial theorem follows. □

We can use binomial coefficients to calculate probabilities in many problems for which the naive definition applies.

Example 1.4.18 (Full house in poker)**.** A 5-card hand is dealt from a standard, well-shuffled 52-card deck. The hand is called a *full house* in poker if it consists of three cards of some rank and two cards of another rank, e.g., three 7's and two 10's (in any order). What is the probability of a full house?

Solution:

All of the $\binom{52}{5}$ possible hands are equally likely by symmetry, so the naive definition is applicable. To find the number of full house hands, use the multiplication rule (and imagine the tree). There are 13 choices for what rank we have three of; for concreteness, assume we have three 7's and focus on that branch of the tree. There are $\binom{4}{3}$ ways to choose which 7's we have. Then there are 12 choices for what rank we have two of, say 10's for concreteness, and $\binom{4}{2}$ ways to choose two 10's. Thus,

$$P(\text{full house}) = \frac{13\binom{4}{3}12\binom{4}{2}}{\binom{52}{5}} = \frac{3744}{2598960} \approx 0.00144.$$

The decimal approximation is more useful when playing poker, but the answer in terms of binomial coefficients is exact and *self-annotating* (seeing "$\binom{52}{5}$" is a much bigger hint of its origin than seeing "2598960"). □

Example 1.4.19 (Newton-Pepys problem)**.** Isaac Newton was consulted about the following problem by Samuel Pepys, who wanted the information for gambling purposes. Which of the following events has the highest probability?

A: At least one 6 appears when 6 fair dice are rolled.

B: At least two 6's appear when 12 fair dice are rolled.

C: At least three 6's appear when 18 fair dice are rolled.

Solution:

The three experiments have 6^6, 6^{12}, and 6^{18} possible outcomes, respectively, and by symmetry the naive definition applies in all three experiments.

A: Instead of counting the number of ways to obtain at least one 6, it is easier to count the number of ways to get no 6's. Getting no 6's is equivalent to sampling the numbers 1 through 5 with replacement 6 times, so 5^6 outcomes are favorable to A^c (and $6^6 - 5^6$ are favorable to A). Thus

$$P(A) = 1 - \frac{5^6}{6^6} \approx 0.67.$$

B: Again we count the outcomes in B^c first. There are 5^{12} ways to get no 6's in 12 die rolls. There are $\binom{12}{1}5^{11}$ ways to get exactly one 6: we first choose which die

lands 6, then sample the numbers 1 through 5 with replacement for the other 11 dice. Adding these, we get the number of ways to fail to obtain at least two 6's. Then

$$P(B) = 1 - \frac{5^{12} + \binom{12}{1}5^{11}}{6^{12}} \approx 0.62.$$

C: We count the outcomes in C^c, i.e., the number of ways to get zero, one, or two 6's in 18 die rolls. There are 5^{18} ways to get no 6's, $\binom{18}{1}5^{17}$ ways to get exactly one 6, and $\binom{18}{2}5^{16}$ ways to get exactly two 6's (choose which two dice will land 6, then decide how the other 16 dice will land).

$$P(C) = 1 - \frac{5^{18} + \binom{18}{1}5^{17} + \binom{18}{2}5^{16}}{6^{18}} \approx 0.60.$$

Therefore A has the highest probability.

Newton arrived at the correct answer using similar calculations. Newton also provided Pepys with an intuitive argument for why A was the most likely of the three; however, his intuition was invalid. As explained in Stigler [27], using loaded dice could result in a different ordering of A, B, C, but Newton's intuitive argument did not depend on the dice being fair. □

In this book, we care about counting not for its own sake, but because it sometimes helps us to find probabilities. Here is an example of a neat but treacherous counting problem; the solution is elegant, but it is rare that the result can be used with the naive definition of probability.

Example 1.4.20 (Bose-Einstein)**.** How many ways are there to choose k times from a set of n objects with replacement, if order doesn't matter (we only care about how many times each object was chosen, not the order in which they were chosen)?

Solution:

When order does matter, the answer is n^k by the multiplication rule, but this problem is much harder. We will solve it by solving an *isomorphic* problem (the same problem in a different guise).

Let us find the number of ways to put k indistinguishable particles into n distinguishable boxes. That is, swapping the particles in any way is not considered a separate possibility: all that matters are the *counts* for how many particles are in each box. Any configuration can be encoded as a sequence of |'s and ●'s in a natural way, as illustrated in Figure 1.5.

To be valid, a sequence must start and end with a |, with exactly $n-1$ |'s and k ●'s in between; conversely, any such sequence is a valid encoding for some configuration of particles in boxes. Thus there are $n + k - 1$ slots between the two outer walls, and we need only choose where to put the k ●'s, so the number of possibilities is $\binom{n+k-1}{k}$. This is known as the *Bose-Einstein* value, since the physicists Satyendra Nath Bose and Albert Einstein studied related problems about indistinguishable

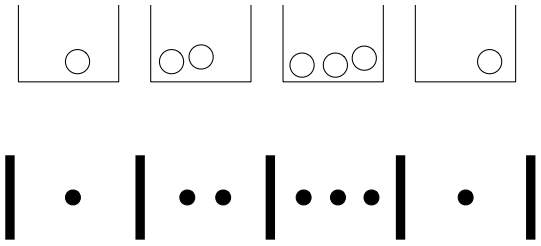


FIGURE 1.5
Bose-Einstein encoding: putting $k = 7$ indistinguishable particles into $n = 4$ distinguishable boxes can be expressed as a sequence of |'s and ●'s, where | denotes a wall and ● denotes a particle.

particles in the 1920s, using their ideas to successfully predict the existence of a strange state of matter known as a Bose-Einstein condensate.

To relate this back to the original question, we can let each box correspond to one of the n objects and use the particles as "check marks" to tally how many times each object is selected. For example, if a certain box contains exactly 3 particles, that means the object corresponding to that box was chosen exactly 3 times. The particles being indistinguishable corresponds to the fact that we don't care about the order in which the objects are chosen. Thus, the answer to the original question is also $\binom{n+k-1}{k}$.

Another isomorphic problem is to count the number of solutions $(x_1, \ldots, x_n)$ to the equation $x_1 + x_2 + \cdots + x_n = k$, where the x_i are nonnegative integers. This is equivalent since we can think of x_i as the number of particles in the ith box.

☣ **1.4.21.** The Bose-Einstein result should *not* be used in the naive definition of probability except in very special circumstances. For example, consider a survey where a sample of size k is collected by choosing people from a population of size n one at a time, with replacement and with equal probabilities. Then the n^k *ordered* samples are equally likely, making the naive definition applicable, but the $\binom{n+k-1}{k}$ unordered samples (where all that matters is how many times each person was sampled) are *not* equally likely.

As another example, with $n = 365$ days in a year and k people, how many possible *unordered* birthday lists are there? For example, for $k = 3$, we want to count lists like (May 1, March 31, April 11), where all permutations are considered equivalent. We can't do a simple adjustment for overcounting such as $n^k/3!$ since, e.g., there are 6 permutations of (May 1, March 31, April 11) but only 3 permutations of (March 31, March 31, April 11). By Bose-Einstein, the number of lists is $\binom{n+k-1}{k}$. But the ordered birthday lists are equally likely, not the unordered lists, so the Bose-Einstein value should not be used in calculating birthday probabilities.

□

1.5 Story proofs

A *story proof* is a proof by interpretation. For counting problems, this often means counting the same thing in two different ways, rather than doing tedious algebra. A story proof often avoids messy calculations and goes further than an algebraic proof toward *explaining* why the result is true. The word "story" has several meanings, some more mathematical than others, but a story proof (in the sense in which we're using the term) is a fully valid mathematical proof. Here are some examples of story proofs, which also serve as further examples of counting.

Example 1.5.1 (Choosing the complement)**.** For any nonnegative integers n and k with $k \leq n$, we have

$$\binom{n}{k} = \binom{n}{n-k}.$$

This is easy to check algebraically (by writing the binomial coefficients in terms of factorials), but a story proof makes the result easier to understand intuitively.

Story proof: Consider choosing a committee of size k in a group of n people. We know that there are $\binom{n}{k}$ possibilities. But another way to choose the committee is to specify which $n-k$ people are *not* on the committee; specifying who is on the committee determines who is *not* on the committee, and vice versa. So the two sides are equal, as they are two ways of counting the same thing. □

Example 1.5.2 (The team captain)**.** For any positive integers n and k with $k \leq n$,

$$n\binom{n-1}{k-1} = k\binom{n}{k}.$$

This is again easy to check algebraically (using the fact that $m! = m(m-1)!$ for any positive integer m), but a story proof is more insightful.

Story proof: Consider a group of n people, from which a team of k will be chosen, one of whom will be the team captain. To specify a possibility, we could first choose the team captain and then choose the remaining $k-1$ team members; this gives the left-hand side. Equivalently, we could first choose the k team members and then choose one of them to be captain; this gives the right-hand side. □

Example 1.5.3 (Vandermonde's identity)**.** A famous relationship between binomial coefficients, called *Vandermonde's identity*, says that

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}.$$

This identity will come up several times in this book. Trying to prove it with a brute force expansion of all the binomial coefficients would be a nightmare. But a story proves the result elegantly and makes it clear *why* the identity holds.

Story proof: Consider a group of m men and n women, from which a committee of size k will be chosen. There are $\binom{m+n}{k}$ possibilities. If there are j men in the committee, then there must be $k-j$ women in the committee. The right-hand side of Vandermonde's identity sums up the cases for j. □

Example 1.5.4 (Partnerships)**.** Let's use a story proof to show that

$$\frac{(2n)!}{2^n \cdot n!} = (2n-1)(2n-3)\cdots 3 \cdot 1.$$

Story proof: We will show that both sides count the number of ways to break $2n$ people into n partnerships. Take $2n$ people, and give them ID numbers from 1 to $2n$. We can form partnerships by lining up the people in some order and then saying the first two are a pair, the next two are a pair, etc. This overcounts by a factor of $n! \cdot 2^n$ since the order of pairs doesn't matter, nor does the order within each pair. Alternatively, count the number of possibilities by noting that there are $2n-1$ choices for the partner of person 1, then $2n-3$ choices for person 2 (or person 3, if person 2 was already paired to person 1), and so on. □

1.6 Non-naive definition of probability

We have now seen several methods for counting outcomes in a sample space, allowing us to calculate probabilities if the naive definition applies. But the naive definition can only take us so far, since it requires equally likely outcomes and can't handle an infinite sample space. To generalize the notion of probability, we'll use the best part about math, which is that you get to *make up your own definitions*. What this means is that we write down a short wish list of how we want probability to behave (in math, the items on the wish list are called *axioms*), and then we define a probability function to be something that satisfies the properties we want!

Here is the general definition of probability that we'll use for the rest of this book. It requires just two axioms, but from these axioms it is possible to prove a vast array of results about probability.

Definition 1.6.1 (General definition of probability)**.** A *probability space* consists of a sample space S and a *probability function* P which takes an event $A \subseteq S$ as input and returns $P(A)$, a real number between 0 and 1, as output. The function P must satisfy the following axioms:

1. $P(\emptyset) = 0$, $P(S) = 1$.

2. If $A_1, A_2, \ldots$ are disjoint events, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j).$$

(Saying that these events are *disjoint* means that they are *mutually exclusive*: $A_i \cap A_j = \emptyset$ for $i \neq j$.)

In Pebble World, the definition says that probability behaves like mass: the mass of an empty pile of pebbles is 0, the total mass of all the pebbles is 1, and if we have non-overlapping piles of pebbles, we can get their combined mass by adding the masses of the individual piles. Unlike in the naive case, we can now have pebbles of differing masses, and we can also have a countably infinite number of pebbles as long as their total mass is 1.

We can even have uncountable sample spaces, such as having S be an area in the plane. In this case, instead of pebbles, we can visualize mud spread out over a region, where the total mass of the mud is 1.

Any function P (mapping events to numbers in the interval $[0, 1]$) that satisfies the two axioms is considered a valid probability function. However, the axioms don't tell us how probability should be *interpreted*; different schools of thought exist.

The *frequentist* view of probability is that it represents a long-run frequency over a large number of repetitions of an experiment: if we say a coin has probability $1/2$ of Heads, that means the coin would land Heads 50% of the time if we tossed it over and over and over.

The *Bayesian* view of probability is that it represents a degree of belief about the event in question, so we can assign probabilities to hypotheses like "candidate A will win the election" or "the defendant is guilty" even if it isn't possible to repeat the same election or the same crime over and over again.

The Bayesian and frequentist perspectives are complementary, and both will be helpful for developing intuition in later chapters. Regardless of how we choose to interpret probability, we can use the two axioms to derive other properties of probability, and these results will hold for *any* valid probability function.

Theorem 1.6.2 (Properties of probability)**.** Probability has the following properties, for any events A and B.

1. $P(A^c) = 1 - P(A)$.
2. If $A \subseteq B$, then $P(A) \leq P(B)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

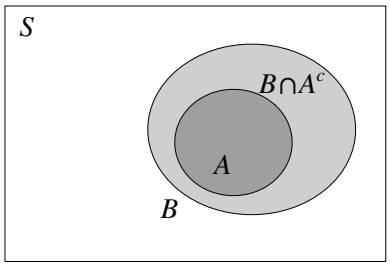
Proof.

1. Since A and A^c are disjoint and their union is S, the second axiom gives

$$P(S) = P(A \cup A^c) = P(A) + P(A^c),$$

But $P(S) = 1$ by the first axiom. So $P(A) + P(A^c) = 1$.

2. If $A \subseteq B$, then we can write B as the union of A and $B \cap A^c$, where $B \cap A^c$ is the part of B not also in A. This is illustrated in the figure below.

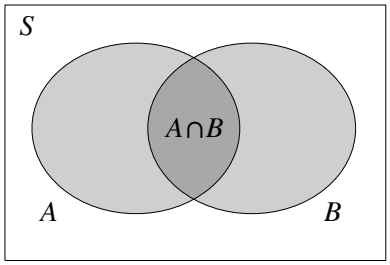


Since A and $B \cap A^c$ are disjoint, we can apply the second axiom:

$$P(B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c).$$

Probability is nonnegative, so $P(B \cap A^c) \geq 0$, proving that $P(B) \geq P(A)$.

3. The intuition for this result can be seen using a Venn diagram like the one below.



The shaded region represents $A \cup B$, but the probability of this region is not $P(A) + P(B)$, because that would count the football-shaped region $A \cap B$ twice. To correct for this, we subtract $P(A \cap B)$. This is a useful intuition, but not a proof.

For a proof using the axioms of probability, we can write $A \cup B$ as the union of the disjoint events A and $B \cap A^c$. Then by the second axiom,

$$P(A \cup B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c).$$

So it suffices to show that $P(B \cap A^c) = P(B) - P(A \cap B)$. Since $A \cap B$ and $B \cap A^c$ are disjoint and their union is B, another application of the second axiom gives us

$$P(A \cap B) + P(B \cap A^c) = P(B).$$

So $P(B \cap A^c) = P(B) - P(A \cap B)$, as desired. ■

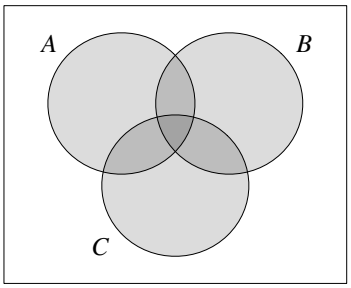
The third property is a special case of *inclusion-exclusion*, a formula for finding the probability of a union of events when the events are not necessarily disjoint. We showed above that for two events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

For three events, inclusion-exclusion says

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

For intuition, consider a triple Venn diagram like the one below.



To get the total area of the shaded region $A \cup B \cup C$, we start by adding the areas of the three circles, $P(A) + P(B) + P(C)$. The three football-shaped regions have each been counted twice, so we then subtract $P(A \cap B) + P(A \cap C) + P(B \cap C)$. Finally, the region in the center has been added three times and subtracted three times, so in order to count it exactly once, we must add it back again. This ensures that each region of the diagram has been counted once and exactly once.

Now we can write inclusion-exclusion for n events.

Theorem 1.6.3 (Inclusion-exclusion)**.** For any events $A_1, \ldots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

This formula can be proven by induction using just the axioms, but instead we'll present a shorter proof in Chapter 4 after introducing some additional tools. The rationale behind the alternating addition and subtraction in the general formula is analogous to the special cases we've already considered.

The next example, *de Montmort's matching problem*, is a famous application of inclusion-exclusion. Pierre Rémond de Montmort was a French mathematician who studied probability in the context of gambling and wrote a treatise [21] devoted to the analysis of various card games. He posed the following problem in 1708, based on a card game called Treize.

Example 1.6.4 (de Montmort's matching problem)**.** Consider a well-shuffled deck of n cards, labeled 1 through n. You flip over the cards one by one, saying the numbers 1 through n as you do so. You win the game if, at some point, the number you say aloud is the same as the number on the card being flipped over (for example, if the 7th card in the deck has the label 7). What is the probability of winning?

Solution:

Let A_i be the event that the ith card in the deck has the number i written on it. We are interested in the probability of the union $A_1 \cup \cdots \cup A_n$: as long as at least one of the cards has a number matching its position in the deck, you will win the game. (An ordering for which you lose is called a *derangement*, though hopefully no one has ever become deranged due to losing at this game.)

To find the probability of the union, we'll use inclusion-exclusion. First,

$$P(A_i) = \frac{1}{n}$$

for all i. One way to see this is with the naive definition of probability, using the full sample space: there are $n!$ possible orderings of the deck, all equally likely, and $(n-1)!$ of these are favorable to A_i (fix the card numbered i to be in the ith position in the deck, and then the remaining $n-1$ cards can be in any order). Another way to see this is by symmetry: the card numbered i is equally likely to be in any of the n positions in the deck, so it has probability $1/n$ of being in the ith spot. Second,

$$P(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)},$$

since we require the cards numbered i and j to be in the ith and jth spots in the deck and allow the remaining $n-2$ cards to be in any order, so $(n-2)!$ out of $n!$ possibilities are favorable to $A_i \cap A_j$. Similarly,

$$P(A_i \cap A_j \cap A_k) = \frac{1}{n(n-1)(n-2)},$$

and the pattern continues for intersections of 4 events, etc.

In the inclusion-exclusion formula, there are n terms involving one event, $\binom{n}{2}$ terms involving two events, $\binom{n}{3}$ terms involving three events, and so forth. By the symmetry of the problem, all n terms of the form $P(A_i)$ are equal, all $\binom{n}{2}$ terms of the form $P(A_i \cap A_j)$ are equal, and the whole expression simplifies considerably:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= \frac{n}{n} - \frac{\binom{n}{2}}{n(n-1)} + \frac{\binom{n}{3}}{n(n-1)(n-2)} - \cdots + (-1)^{n+1} \cdot \frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1} \cdot \frac{1}{n!}. \end{aligned}$$

Comparing this to the Taylor series for $1/e$ (see Section A.8 of the math appendix),

$$e^{-1} = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots,$$

we see that for large n, the probability of winning the game is extremely close to $1 - 1/e$, or about 0.63. Interestingly, as n grows, the probability of winning approaches $1 - 1/e$ instead of going to 0 or 1. With a lot of cards in the deck, the number of possible locations for matching cards increases while the probability of any particular match decreases, and these two forces offset each other and balance to give a probability of about $1 - 1/e$. □

Inclusion-exclusion is a very general formula for the probability of a union of events, but it helps us the most when there is symmetry among the events A_j; otherwise the sum can be extremely tedious. In general, when symmetry is lacking, we should try to use other tools before turning to inclusion-exclusion as a last resort.

1.7 Recap

Probability allows us to quantify uncertainty and randomness in a principled way. Probabilities arise when we perform an experiment: the set of all possible outcomes of the experiment is called the sample space, and a subset of the sample space is called an event. It is useful to be able to go back and forth between describing events in English and writing them down mathematically as sets (often using unions, intersections, and complements).

Pebble World can help us visualize sample spaces and events when the sample space is finite. In Pebble World, each outcome is a pebble, and an event is a set of pebbles. If all the pebbles have the same mass (i.e., are equally likely), we can apply the naive definition of probability, which lets us calculate probabilities by counting.

To this end, we discussed several tools for counting. When counting the number of possibilities, we often use the multiplication rule. If we can't directly use the multiplication rule, we can sometimes count each possibility exactly c times for some c, and then divide by c to get the actual number of possibilities.

An important pitfall to avoid is misapplying the naive definition of probability, implicitly or explicitly assuming equally likely outcomes without justification. One technique to help avoid this is to *give objects labels*, for precision and so that we are not tempted to treat them as indistinguishable.

Moving beyond the naive definition, we define probability to be a function that takes an event and assigns to it a real number between 0 and 1. We require a valid probability function to satisfy two axioms:

1. $P(\emptyset) = 0$, $P(S) = 1$.

2. If $A_1, A_2, \ldots$ are disjoint events, then $P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$.

Many useful properties can be derived just from these axioms. For example,

$$P(A^c) = 1 - P(A)$$

for any event A, and we have the inclusion-exclusion formula

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n)$$

for any events $A_1, \ldots, A_n$. For $n = 2$, this is the much nicer-looking result

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

Figure 1.6 illustrates how a probability function maps events to numbers between 0 and 1. We'll add many new concepts to this diagram as we continue our journey through the field of probability.

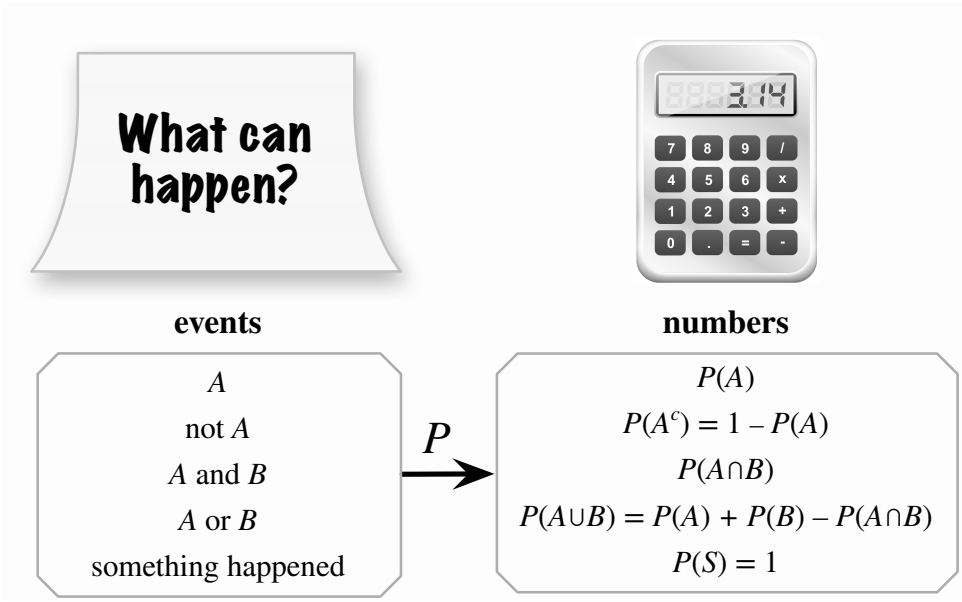


FIGURE 1.6
It is important to distinguish between *events* and *probabilities*. The former are sets, while the latter are numbers. Before the experiment is done, we generally don't know whether or not a particular event will occur (happen). So we assign it a probability of happening, using a probability function P. We can use set operations to define new events in terms of old events, and the properties of probabilities to relate the probabilities of the new events to those of the old events.

2

Conditional probability

We have introduced probability as a language for expressing our degrees of belief or uncertainties about events. Whenever we observe new evidence (i.e., obtain *data*), we acquire information that may affect our uncertainties. A new observation that is consistent with an existing belief could make us more sure of that belief, while a surprising observation could throw that belief into question. *Conditional probability* is the concept that addresses this fundamental question: how should we update our beliefs in light of the evidence we observe?

2.1 The importance of thinking conditionally

Conditional probability is essential for scientific, medical, and legal reasoning, since it shows how to incorporate evidence into our understanding of the world in a logical, coherent manner. In fact, a useful perspective is that *all probabilities are conditional*; whether or not it's written explicitly, there is always background knowledge (or assumptions) built into every probability.

Suppose, for example, that one morning we are interested in the event R that it will rain that day. Let $P(R)$ be our assessment of the probability of rain before looking outside. If we then look outside and see ominous clouds in the sky, then presumably our probability of rain should increase; we denote this new probability by $P(R|C)$ (read as "probability of R given C"), where C is the event of there being ominous clouds. When we go from $P(R)$ to $P(R|C)$, we say that we are "conditioning on C". As the day progresses, we may obtain more and more information about the weather conditions, and we can continually update our probabilities. If we observe that events $B_1, \ldots, B_n$ occurred, then we write our new conditional probability of rain given this evidence as $P(R|B_1, \ldots, B_n)$. If eventually we observe that it does start raining, our conditional probability becomes 1.

Furthermore, we will see that conditioning is a very powerful problem-solving strategy, often making it possible to solve a complicated problem by decomposing it into manageable pieces with case-by-case reasoning. Just as in computer science a common strategy is to break a large problem up into bite-size pieces (or even byte-size pieces), in probability a common strategy is to reduce a complicated probability problem to a bunch of simpler conditional probability problems. In particular, we

will discuss a strategy known as *first-step analysis*, which often allows us to obtain recursive solutions to problems where the experiment has multiple stages.

Due to the central importance of conditioning, both as the means by which we update beliefs to reflect evidence and as a problem-solving strategy, we say that

Conditioning is the soul of statistics.

2.2 Definition and intuition

Definition 2.2.1 (Conditional probability)**.** If A and B are events with $P(B) > 0$, then the *conditional probability* of A given B, denoted by $P(A|B)$, is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Here A is the event whose uncertainty we want to update, and B is the evidence we observe (or want to treat as given). We call $P(A)$ the *prior* probability of A and $P(A|B)$ the *posterior* probability of A ("prior" means before updating based on the evidence, and "posterior" means after updating based on the evidence).

It is important to interpret the event appearing after the vertical conditioning bar as the evidence that we have observed or that is being conditioned on: $P(A|B)$ is the probability of A given the evidence B, *not* the probability of some entity called $A|B$. As discussed in ☣ 2.4.1, there is no such event as $A|B$.

For any event A, $P(A|A) = P(A \cap A)/P(A) = 1$. Upon observing that A has occurred, our updated probability for A is 1. If this weren't the case, we would demand a new definition of conditional probability!

Example 2.2.2 (Two cards)**.** A standard deck of cards is shuffled well. Two cards are drawn randomly, one at a time without replacement. Let A be the event that the first card is a heart, and B be the event that the second card is red. Find $P(A|B)$ and $P(B|A)$.

Solution:

By the naive definition of probability and the multiplication rule,

$$P(A \cap B) = \frac{13 \cdot 25}{52 \cdot 51} = \frac{25}{204},$$

since a favorable outcome is determined by choosing any of the 13 hearts and then any of the remaining 25 red cards. Also, $P(A) = 1/4$ since the 4 suits are equally likely, and

$$P(B) = \frac{26 \cdot 51}{52 \cdot 51} = \frac{1}{2}$$

since there are 26 favorable possibilities for the *second* card, and for each of those, the first card can be any other card (recall from Chapter 1 that chronological order is not needed in the multiplication rule). A neater way to see that $P(B) = 1/2$ is by *symmetry*: from a vantage point before having done the experiment, the second card is equally likely to be any card in the deck. We now have all the pieces needed to apply the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{25/204}{1/2} = \frac{25}{102},$$
$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{25/204}{1/4} = \frac{25}{51}.$$

This is a simple example, but already there are several things worth noting.

1. It's extremely important to be careful about which events to put on which side of the conditioning bar. In particular, $P(A|B) \neq P(B|A)$. The next section explores how $P(A|B)$ and $P(B|A)$ are related in general. Confusing these two quantities is called the *prosecutor's fallacy*. If instead we had defined B to be the event that the second card is a heart, the two conditional probabilities would have been equal.

2. Both $P(A|B)$ and $P(B|A)$ make sense (intuitively and mathematically); the chronological order in which cards were chosen does not dictate which conditional probabilities we can look at. When we calculate conditional probabilities, we are considering what *information* observing one event provides about another event, not whether one event *causes* another.

3. We can also see that $P(B|A) = 25/51$ by a direct interpretation of what conditional probability means: if the first card drawn is a heart, then the remaining cards consist of 25 red cards and 26 black cards (all of which are equally likely to be drawn next), so the conditional probability of getting a red card is $25/(25 + 26) = 25/51$. It is harder to find $P(A|B)$ in this way: if we learn that the second card is red, we might think "that's nice to know, but what we really want to know is whether it's a heart!" The conditional probability results from later sections in this chapter give us methods for getting around this issue. □

To shed more light on what conditional probability means, here are two intuitive interpretations.

Intuition 2.2.3 (Pebble World)**.** Consider a finite sample space, with the outcomes visualized as pebbles with total mass 1. Since A is an event, it is a set of pebbles, and likewise for B. Figure 2.1(a) shows an example.

Now suppose that we learn that B occurred. In Figure 2.1(b), upon obtaining this information, we get rid of all the pebbles in B^c because they are incompatible with the knowledge that B has occurred. Then $P(A \cap B)$ is the total mass of the pebbles remaining in A. Finally, in Figure 2.1(c), we *renormalize*, that is, divide all the masses by a constant so that the new total mass of the remaining pebbles

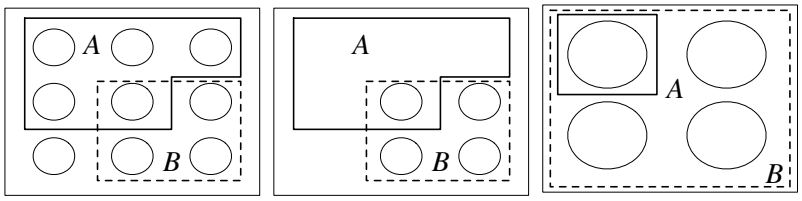


FIGURE 2.1
Pebble World intuition for $P(A|B)$. From left to right: (a) Events A and B are subsets of the sample space. (b) Because we know B occurred, get rid of the outcomes in B^c. (c) In the restricted sample space, renormalize so the total mass is still 1.

is 1. This is achieved by dividing by $P(B)$, the total mass of the pebbles in B. The updated mass of the outcomes corresponding to event A is the conditional probability $P(A|B) = P(A \cap B)/P(B)$.

In this way, our probabilities have been updated in accordance with the observed evidence. Outcomes that contradict the evidence are discarded, and their mass is redistributed among the remaining outcomes, preserving the relative masses of the remaining outcomes. For example, if pebble 2 weighs twice as much as pebble 1 initially, and both are contained in B, then after conditioning on B it is still true that pebble 2 weighs twice as much as pebble 1. But if pebble 2 is not contained in B, then after conditioning on B its mass is updated to 0. □

Intuition 2.2.4 (Frequentist interpretation)**.** Recall that the frequentist interpretation of probability is based on relative frequency over a large number of repeated trials. Imagine repeating our experiment many times, generating a long list of observed outcomes. The conditional probability of A given B can then be thought of in a natural way: it is the fraction of times that A occurs, restricting attention to the trials where B occurs. In Figure 2.2, our experiment has outcomes which can be written as a string of 0's and 1's; B is the event that the first digit is 1 and A is the event that the second digit is 1. Conditioning on B, we circle all the repetitions where B occurred, and then we look at the fraction of circled repetitions in which event A also occurred.

In symbols, let n_A, n_B, n_{AB} be the number of occurrences of $A, B, A \cap B$ respectively in a large number n of repetitions of the experiment. The frequentist interpretation is that

$$P(A) \approx \frac{n_A}{n}, P(B) \approx \frac{n_B}{n}, P(A \cap B) \approx \frac{n_{AB}}{n}.$$

Then $P(A|B)$ is interpreted as n_{AB}/n_B, which equals $(n_{AB}/n)/(n_B/n)$. This interpretation again translates to $P(A|B) = P(A \cap B)/P(B)$. □

For practice with applying the definition of conditional probability, let's do a few more examples. The next three examples all start with the same basic scenario of a family with two children, but subtleties arise depending on the exact information we condition on.

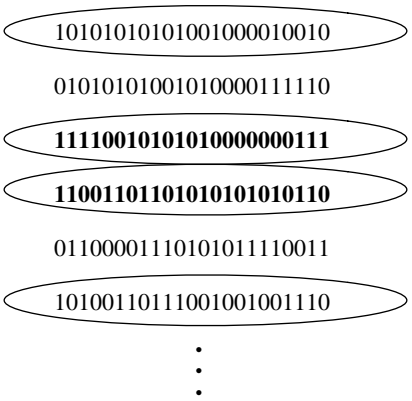


FIGURE 2.2
Frequentist intuition for $P(A|B)$. The repetitions where B occurred are circled; among these, the repetitions where A occurred are highlighted in bold. $P(A|B)$ is the long-run relative frequency of the repetitions where A occurs, within the subset of repetitions where B occurs.

Example 2.2.5 (Elder is a girl vs. at least one girl)**.** A family has two children, and it is known that at least one is a girl. What is the probability that both are girls, given this information? What if it is known that the *elder* child is a girl?

Solution:

Assume each child is equally likely to be a boy or a girl, independently.[1] Then

$$P(\text{both girls}|\text{at least one girl}) = \frac{P(\text{both girls, at least one girl})}{P(\text{at least one girl})} = \frac{1/4}{3/4} = 1/3,$$

$$P(\text{both girls}|\text{elder is a girl}) = \frac{P(\text{both girls, elder is a girl})}{P(\text{elder is a girl})} = \frac{1/4}{1/2} = 1/2.$$

It may seem counterintuitive at first that the two results are different, since there is no reason for us to care whether the elder child is a girl as opposed to the younger child. Indeed, by symmetry,

$$P(\text{both girls}|\text{younger is a girl}) = P(\text{both girls}|\text{elder is a girl}) = 1/2.$$

However, there is no such symmetry between the conditional probabilities $P(\text{both girls}|\text{elder is a girl})$ and $P(\text{both girls}|\text{at least one girl})$. Saying that the elder child is a girl designates a *specific* child, and then the other child (the younger

[1]Independence is introduced formally in a later section, but for now it can just be thought of intuitively: knowing whether the elder child is a girl gives no information about whether the younger child is as girl, and vice versa. For simplicity, this problem assumes that each child is a boy or a girl, with equal probabilities, though in reality slightly more boys than girls are born (in most countries), and a small percentage of babies are intersex. See Matthews [20] for a report on the ratio of boys born to girls born in the U.S.

child) has a 50% chance of being a girl. "At least one" does *not* refer to a specific child. Conditioning on a specific child being a girl knocks away 2 of the 4 "pebbles" in the sample space $\{GG, GB, BG, BB\}$, while conditioning on at least one child being a girl only knocks away BB. □

Example 2.2.6 (Random child is a girl)**.** A family has two children. You randomly run into one of the two, and see that she is a girl. What is the probability that both are girls, given this information? Assume that you are equally likely to run into either child, and that which one you run into has nothing to do with gender.

Solution:

Intuitively, the answer should be 1/2: imagine that the child we encountered is in front of us and the other is at home. Both being girls just says that the child who is at home is a girl, which seems to have nothing to do with the fact that the child in front of us is a girl. But let us check this more carefully, using the definition of conditional probability. This is also good practice with writing events in set notation.

Let G_1, G_2, G_3 be the events that the elder, younger, and random child is a girl, respectively. We have $P(G_1) = P(G_2) = P(G_3) = 1/2$ by symmetry (we are implicitly assuming that any specific child, in the absence of any other information, is equally likely to be a boy or a girl). By the naive definition of probability, $P(G_1 \cap G_2) = 1/4$. Thus,

$$P(G_1 \cap G_2 | G_3) = P(G_1 \cap G_2 \cap G_3)/P(G_3) = (1/4)/(1/2) = 1/2,$$

since $G_1 \cap G_2 \cap G_3 = G_1 \cap G_2$ (if both children are girls, it guarantees that the random child is a girl). So the probability that both are girls is 1/2, consistent with our intuition.

Keep in mind though that in order to arrive at 1/2, an assumption was needed about how the random child was selected. In statistical language, we say that we collected a *random sample*; here the sample consists of one of the two children. One of the most important principles in statistics is that it is essential to think carefully about how the sample was collected, not just stare at the raw data without understanding where they came from. To take a simple extreme case, suppose that a repressive law forbids a boy from leaving the house if he has a sister. Then "the random child is a girl" is equivalent to "at least one of the children is a girl", so the problem reduces to the first part of Example 2.2.5. □

Example 2.2.7 (A girl born in winter)**.** A family has two children. Find the probability that both children are girls, given that at least one of the two is a girl who was born in winter. Assume that the four seasons are equally likely and that gender is independent of season (this means that knowing the gender gives no information about the probabilities of the seasons, and vice versa).

Solution:

By the definition of conditional probability,

$$P(\text{both girls}|\text{at least one winter girl}) = \frac{P(\text{both girls, at least one winter girl})}{P(\text{at least one winter girl})}.$$

Since the probability that a specific child is a winter-born girl is 1/8, the denominator equals

$$P(\text{at least one winter girl}) = 1 - (7/8)^2.$$

To compute the numerator, use the fact that "both girls, at least one winter girl" is the same event as "both girls, at least one winter child"; then use the assumption that gender and season are independent:

$$\begin{aligned} P(\text{both girls, at least one winter girl}) &= P(\text{both girls, at least one winter child}) \\ &= (1/4)P(\text{at least one winter child}) \\ &= (1/4)(1 - P(\text{both are non-winter})) \\ &= (1/4)(1 - (3/4)^2). \end{aligned}$$

Altogether, we obtain

$$P(\text{both girls}|\text{at least one winter girl}) = \frac{(1/4)(1-(3/4)^2)}{1-(7/8)^2} = \frac{7/64}{15/64} = 7/15.$$

At first this result seems absurd! In Example 2.2.5 we saw that the conditional probability of both children being girls, given that at least one is a girl, is 1/3; why should it be any different when we learn that at least one is a winter-born girl? The point is that information about the birth season brings "at least one is a girl" closer to "a specific one is a girl". Conditioning on more and more specific information brings the probability closer and closer to 1/2.

For example, conditioning on "at least one is a girl who was born on a March 31 at 8:20 pm" comes very close to specifying a child, and learning information about a specific child does not give us information about the other child. The seemingly irrelevant information such as season of birth interpolates between the two parts of Example 2.2.5. □

2.3 Bayes' rule and the law of total probability

The definition of conditional probability is simple—just a ratio of two probabilities—but it has far-reaching consequences. The first consequence is obtained easily by moving the denominator in the definition to the other side of the equation.

Theorem 2.3.1. For any events A and B with positive probabilities,

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A).$$

This follows from taking the definition of $P(A|B)$ and multiplying both sides by $P(B)$, and then taking the definition of $P(B|A)$ and multiplying both sides by $P(A)$. At first sight this theorem may not seem very useful: it *is* the definition of conditional probability, just written slightly differently, and anyway it seems circular to use $P(A|B)$ to help find $P(A \cap B)$ when $P(A|B)$ was defined in terms of $P(A \cap B)$. But we will see that the theorem is in fact very useful, since it often turns out to be possible to find conditional probabilities without going back to the definition, and in such cases Theorem 2.3.1 can help us more easily find $P(A \cap B)$.

Applying Theorem 2.3.1 repeatedly, we can generalize to the intersection of n events.

Theorem 2.3.2. For any events $A_1, \ldots, A_n$ with positive probabilities,

$$P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \ldots, A_{n-1}).$$

The commas denote intersections.

In fact, this is $n!$ theorems in one, since we can permute $A_1, \ldots, A_n$ however we want without affecting the left-hand side. Often the right-hand side will be much easier to compute for some orderings than for others. For example,

$$P(A_1, A_2, A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) = P(A_2)P(A_3|A_2)P(A_1|A_2, A_3),$$

and there are 4 other expansions of this form too. It often takes practice and thought to be able to know which ordering to use.

We are now ready to introduce the two main theorems of this chapter—Bayes' rule and the law of total probability—which will allow us to compute conditional probabilities in a wide range of problems. Bayes' rule is an extremely famous, extremely useful result that relates $P(A|B)$ to $P(B|A)$.

Theorem 2.3.3 (Bayes' rule)**.**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

This follows immediately from Theorem 2.3.1, which in turn followed immediately from the definition of conditional probability. Yet Bayes' rule has important implications and applications in probability and statistics, since it is so often necessary to find conditional probabilities, and often $P(B|A)$ is much easier to find directly than $P(A|B)$ (or vice versa).

Another way to write Bayes' rule is in terms of *odds* rather than probability.

Definition 2.3.4 (Odds)**.** The *odds* of an event A are

$$\text{odds}(A) = P(A)/P(A^c).$$

For example, if $P(A) = 2/3$, we say the odds in favor of A are 2 to 1. (This is sometimes written as $2:1$, and is sometimes stated as 1 to 2 odds against A; care is needed since some sources do not explicitly state whether they are referring to odds in favor or odds against an event.) Of course we can also convert from odds back to probability:

$$P(A) = \text{odds}(A)/(1 + \text{odds}(A)).$$

By taking the Bayes' rule expression for $P(A|B)$ and dividing it by the Bayes' rule expression for $P(A^c|B)$, we arrive at the odds form of Bayes' rule.

Theorem 2.3.5 (Odds form of Bayes' rule)**.** For any events A and B with positive probabilities, the odds of A after conditioning on B are

$$\frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)}\frac{P(A)}{P(A^c)}.$$

In words, this says that the *posterior odds* $P(A|B)/P(A^c|B)$ are equal to the *prior odds* $P(A)/P(A^c)$ times the factor $P(B|A)/P(B|A^c)$, which is known in statistics as the *likelihood ratio*. Sometimes it is convenient to work with this form of Bayes' rule to get the posterior odds, and then if desired we can convert from odds back to probability.

The *law of total probability (LOTP)* relates conditional probability to unconditional probability. It is essential for fulfilling the promise that conditional probability can be used to decompose complicated probability problems into simpler pieces, and it is often used in tandem with Bayes' rule.

Theorem 2.3.6 (Law of total probability (LOTP))**.** Let $A_1, \ldots, A_n$ be a partition of the sample space S (i.e., the A_i are disjoint events and their union is S), with $P(A_i) > 0$ for all i. Then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Proof. Since the A_i form a partition of S, we can decompose B as

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n).$$

This is illustrated in Figure 2.3, where we have chopped B into the smaller pieces $B \cap A_1$ through $B \cap A_n$. By the second axiom of probability, because these pieces are disjoint, we can add their probabilities to get $P(B)$:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n).$$

Now we can apply Theorem 2.3.1 to each of the $P(B \cap A_i)$:

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n).$$

■

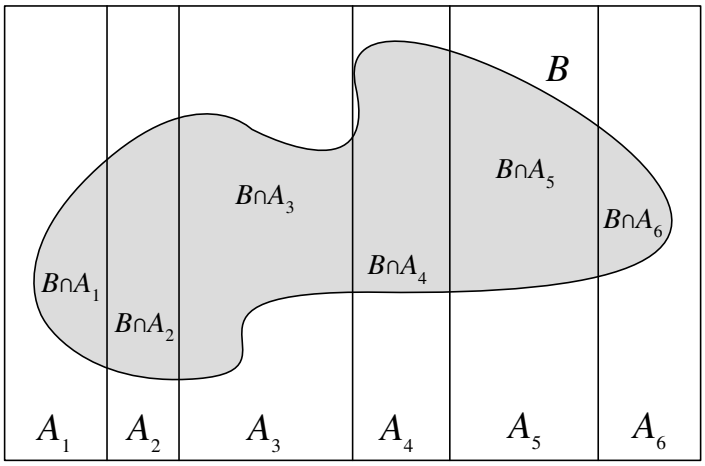


FIGURE 2.3
The A_i partition the sample space; $P(B)$ is equal to $\sum_i P(B \cap A_i)$.

The law of total probability tells us that to get the unconditional probability of B, we can divide the sample space into disjoint slices A_i, find the conditional probability of B within each of the slices, then take a weighted sum of the conditional probabilities, where the weights are the probabilities $P(A_i)$. The choice of how to divide up the sample space is crucial: a well-chosen partition will reduce a complicated problem into simpler pieces, whereas a poorly chosen partition will only exacerbate our problems, requiring us to calculate n difficult probabilities instead of just one!

The next few examples show how we can use Bayes' rule together with the law of total probability to update our beliefs based on observed evidence.

Example 2.3.7 (Random coin)**.** You have one fair coin, and one biased coin which lands Heads with probability 3/4. You pick one of the coins at random and flip it three times. It lands Heads all three times. Given this information, what is the probability that the coin you picked is the fair one?

Solution:

Let A be the event that the chosen coin lands Heads three times and let F be the event that we picked the fair coin. We are interested in $P(F|A)$, but it is easier to find $P(A|F)$ and $P(A|F^c)$ since it helps to know which coin we have; this suggests

using Bayes' rule and the law of total probability. Doing so, we have

$$\begin{aligned} P(F|A) &= \frac{P(A|F)P(F)}{P(A)} \\ &= \frac{P(A|F)P(F)}{P(A|F)P(F) + P(A|F^c)P(F^c)} \\ &= \frac{(1/2)^3 \cdot 1/2}{(1/2)^3 \cdot 1/2 + (3/4)^3 \cdot 1/2} \\ &\approx 0.23. \end{aligned}$$

Before flipping the coin, we thought we were equally likely to have picked the fair coin as the biased coin: $P(F) = P(F^c) = 1/2$. Upon observing three Heads, however, it becomes more likely that we've chosen the biased coin than the fair coin, so $P(F|A)$ is only about 0.23. □

☣ **2.3.8** (Prior vs. posterior)**.** It would *not* be correct in the calculation in the above example to say after the first step, "$P(A) = 1$ because we know A happened." It is true that $P(A|A) = 1$, but $P(A)$ is the *prior* probability of A and $P(F)$ is the *prior* probability of F—both are the probabilities before we observe any data in the experiment. These must not be confused with *posterior* probabilities conditional on the evidence A.

Example 2.3.9 (Testing for a rare disease)**.** A patient named Fred is tested for a disease called conditionitis, a medical condition that afflicts 1% of the population. The test result is positive, i.e., the test claims that Fred has the disease. Let D be the event that Fred has the disease and T be the event that he tests positive.

Suppose that the test is "95% accurate"; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that $P(T|D) = 0.95$ and $P(T^c|D^c) = 0.95$. The quantity $P(T|D)$ is known as the *sensitivity* or *true positive rate* of the test, and $P(T^c|D^c)$ is known as the *specificity* or *true negative rate*.

Find the conditional probability that Fred has conditionitis, given the evidence provided by the test result.

Solution:

Applying Bayes' rule and the law of total probability, we have

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} \\ &= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} \\ &= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \\ &\approx 0.16. \end{aligned}$$

So there is only a 16% chance that Fred has conditionitis, given that he tested positive, even though the test seems to be quite reliable!

Many people, including doctors, find it surprising that the conditional probability of having the disease given a positive test result is only 16%, even though the test is 95% accurate (see Gigerenzer and Hoffrage [14]). The key to understanding this surprisingly high posterior probability is to realize that there are two factors at play: the evidence from the test, and our *prior information* about the prevalence of the disease. Although the test provides evidence in favor of disease, conditionitis is also a rare condition! The conditional probability $P(D|T)$ reflects a balance between these two factors, appropriately weighing the rarity of the disease against the rarity of a mistaken test result.

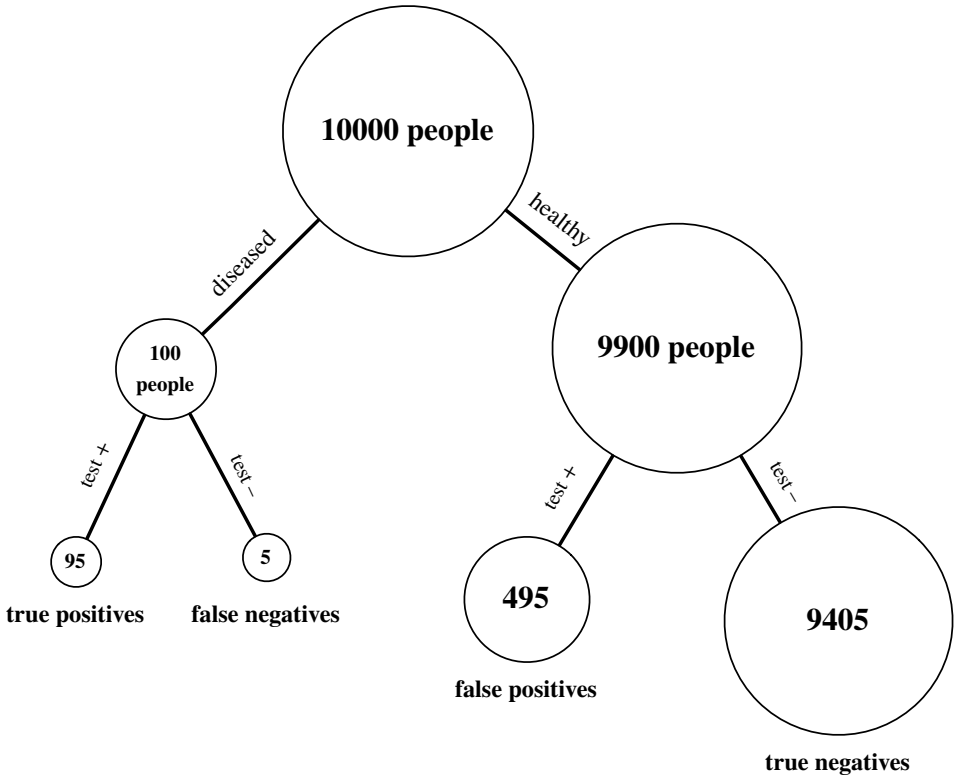


FIGURE 2.4
Testing for a rare disease in a population of 10000 people, where the prevalence of the disease is 1% and the true positive and true negative rates are both equal to 95%. Bubbles are not to scale.

For further intuition, consider a population of 10000 people as illustrated in Figure 2.4, where 100 have conditionitis and 9900 don't; this corresponds to a 1% disease rate. If we tested everybody in the population, we'd expect that out of the 100 diseased individuals, 95 would test positive and 5 would test negative. Out of the 9900 healthy individuals, we'd expect $(0.95)(9900) \approx 9405$ to test negative and 495 to test positive.

Now let's focus in on those individuals who test positive; that is, let's condition on a

positive test result. The 95 true positives (i.e., the individuals who test positive and have the disease) are far outnumbered by the 495 false positives (i.e., the individuals who test positive despite not having the disease). So most people who test positive for the disease are actually disease-free! □

2.4 Conditional probabilities are probabilities

When we condition on an event E, we update our beliefs to be consistent with this knowledge, effectively putting ourselves in a universe where we know that E occurred. Within our new universe, however, the laws of probability operate just as before. Conditional probability satisfies all the properties of probability! Therefore, any of the results we have derived about probability are still valid if we replace all unconditional probabilities with probabilities conditional on E. In particular:

- Conditional probabilities are between 0 and 1.
- $P(S|E) = 1$, $P(\emptyset|E) = 0$.
- If $A_1, A_2, \ldots$ are disjoint, then $P(\cup_{j=1}^{\infty} A_j|E) = \sum_{j=1}^{\infty} P(A_j|E)$.
- $P(A^c|E) = 1 - P(A|E)$.
- Inclusion-exclusion: $P(A \cup B|E) = P(A|E) + P(B|E) - P(A \cap B|E)$.

☣ **2.4.1.** When we write $P(A|E)$, it does *not* mean that $A|E$ is an event and we're taking its probability; $A|E$ is not an event. Rather, $P(\cdot|E)$ is a probability function which assigns probabilities in accordance with the knowledge that E has occurred, and $P(\cdot)$ is a different probability function which assigns probabilities without regard for whether E has occurred or not. When we take an event A and plug it into the $P(\cdot)$ function, we'll get a number, $P(A)$; when we plug it into the $P(\cdot|E)$ function, we'll get another number, $P(A|E)$, which incorporates the information (if any) provided by knowing that E occurred.

To prove mathematically that conditional probabilities are probabilities, fix an event E with $P(E) > 0$, and for any event A, define $\tilde{P}(A) = P(A|E)$. This notation helps emphasize the fact that we are fixing E and treating $P(\cdot|E)$ as our new probability function. We just need to check the two axioms of probability. First,

$$\tilde{P}(\emptyset) = P(\emptyset|E) = \frac{P(\emptyset \cap E)}{P(E)} = 0, \tilde{P}(S) = P(S|E) = \frac{P(S \cap E)}{P(E)} = 1.$$

Second, if $A_1, A_2, \ldots$ are disjoint events, then

$$\tilde{P}(A_1 \cup A_2 \cup \cdots) = \frac{P((A_1 \cap E) \cup (A_2 \cap E) \cup \cdots)}{P(E)} = \frac{\sum_{j=1}^{\infty} P(A_j \cap E)}{P(E)} = \sum_{j=1}^{\infty} \tilde{P}(A_j).$$

So $\tilde{P}$ satisfies the axioms of probability.

Conversely, *all* probabilities can be thought of as conditional probabilities: whenever we make a probability statement, there is always some background information that we are conditioning on, even if we don't state it explicitly. Consider the rain example from the beginning of this chapter. It would be natural to base the initial probability of rain today, $P(R)$, on the fraction of days in the past on which it rained. But which days in the past should we look at? If it's November 1, should we only count past rainy days in autumn, thus conditioning on the season? What about conditioning on the exact month, or the exact day? We could ask the same about location: should we look at days when it rained in our exact location, or is it enough for it to have rained somewhere nearby? In order to determine the seemingly unconditional probability $P(R)$, we actually have to make decisions about what background information to condition on! These choices require careful thought and different people may come up with different "prior" probabilities $P(R)$ (though everyone can agree on how to update based on new evidence).

Since all probabilities are conditional on background information, we can imagine that there is always a vertical conditioning bar, with background knowledge K to the right of the vertical bar. Then the unconditional probability $P(A)$ is just shorthand for $P(A|K)$; the background knowledge is absorbed into the letter P instead of being written explicitly.

To summarize our discussion in a nutshell:

Conditional probabilities are probabilities, and all probabilities are conditional.

We now state conditional forms of Bayes' rule and the law of total probability. These are obtained by taking the ordinary forms of Bayes' rule and LOTP and adding E to the right of the vertical bar everywhere.

Theorem 2.4.2 (Bayes' rule with extra conditioning)**.** Provided that $P(A \cap E) > 0$ and $P(B \cap E) > 0$, we have

$$P(A|B,E) = \frac{P(B|A,E)P(A|E)}{P(B|E)}.$$

Theorem 2.4.3 (LOTP with extra conditioning)**.** Let $A_1, \ldots, A_n$ be a partition of S. Provided that $P(A_i \cap E) > 0$ for all i, we have

$$P(B|E) = \sum_{i=1}^{n} P(B|A_i,E)P(A_i|E).$$

The extra conditioning forms of Bayes' rule and LOTP can be proved similarly to how we verified that $\tilde{P}$ satisfies the axioms of probability, but they also follow directly from the "metatheorem" that *conditional probabilities are probabilities.*

Example 2.4.4 (Random coin, continued)**.** Continuing with the scenario from Example 2.3.7, suppose that we have now seen our chosen coin land Heads three

times. If we toss the coin a fourth time, what is the probability that it will land Heads once more?

Solution:

As before, let A be the event that the chosen coin lands Heads three times, and define a new event H for the chosen coin landing Heads on the fourth toss. We are interested in $P(H|A)$. It would be very helpful to know whether we have the fair coin. LOTP with extra conditioning gives us $P(H|A)$ as a weighted average of $P(H|F, A)$ and $P(H|F^c, A)$, and within these two conditional probabilities we *do* know whether we have the fair coin:

$$\begin{aligned} P(H|A) &= P(H|F, A)P(F|A) + P(H|F^c, A)P(F^c|A) \\ &= \frac{1}{2} \cdot 0.23 + \frac{3}{4} \cdot (1 - 0.23) \\ &\approx 0.69. \end{aligned}$$

The posterior probabilities $P(F|A)$ and $P(F^c|A)$ are from our answer to Example 2.3.7.

An equivalent way to solve this problem is to define a new probability function $\tilde{P}$ such that for any event B, $\tilde{P}(B) = P(B|A)$. This new function assigns probabilities that are updated with the knowledge that A occurred. Then by the ordinary law of total probability,

$$\tilde{P}(H) = \tilde{P}(H|F)\tilde{P}(F) + \tilde{P}(H|F^c)\tilde{P}(F^c),$$

which is exactly the same as our use of LOTP with extra conditioning. This once again illustrates the principle that conditional probabilities are probabilities. □

We often want to condition on more than one piece of information, and we now have several ways of doing that. For example, here are some approaches for finding $P(A|B, C)$:

1. We can think of B, C as the single event $B \cap C$ and use the definition of conditional probability to get

$$P(A|B, C) = \frac{P(A, B, C)}{P(B, C)}.$$

This is a natural approach if it's easiest to think about B and C in tandem. We can then try to evaluate the numerator and denominator. For example, we can use LOTP in both the numerator and the denominator, or we can write the numerator as $P(B, C|A)P(A)$ (which would give us a version of Bayes' rule) and use LOTP to help with the denominator.

2. We can use Bayes' rule with extra conditioning on C to get

$$P(A|B, C) = \frac{P(B|A, C)P(A|C)}{P(B|C)}.$$

This is a natural approach if we want to think of everything in our problem as being conditioned on C.

3. We can use Bayes' rule with extra conditioning on B to get

$$P(A|B,C) = \frac{P(C|A,B)P(A|B)}{P(C|B)}.$$

This is the same as the previous approach, except with the roles of B and C swapped. We mention it separately just to emphasize that it's a bad idea to plug into a formula without thinking about which event should play which role.

It is both challenging and powerful that there are a variety of ways to approach this kind of conditioning problem.

A

Math

A.1 Sets

A set is a Many that allows itself to be thought of as a One.
– Georg Cantor

Amazon should put their cloud in a cloud, so the cloud will have the redundancy of the cloud.
– @dowens

A *set* is a collection of objects. The objects can be anything: numbers, people, cats, courses, even other sets! The language of sets allows us to talk precisely about *events*. If S is a set, then the notation $x \in S$ indicates that x is an element or member of the set S (and $x \notin S$ indicates that x is not in S). We can think of the set as a club, with precisely defined criteria for membership. For example:

1. $\{1, 3, 5, 7, \dots\}$ is the set of all odd numbers;
2. {Worf, Jack, Tobey} is the set of Joe's cats;
3. $[3, 7]$ is the closed interval consisting of all real numbers between 3 and 7;
4. $\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$ is the set of all possible outcomes if a coin is flipped twice (where, for example, HT means the first flip lands Heads and the second lands Tails).

To describe a set (when it's tedious or impossible to list out its elements), we can give a rule that says whether each possible object is or isn't in the set. For example, $\{(x, y) : x \text{ and } y \text{ are real numbers and } x^2 + y^2 \leq 1\}$ is the disk in the plane of radius 1, centered at the origin.

A.1.1 The empty set

Bu Fu to Chi Po: "No, no! You have merely painted what is! Anyone can paint what is; the real secret is to paint what isn't."
Chi Po: "But what is there that isn't?"
– Oscar Mandel [19]

The smallest set, which is both subtle and important, is the *empty set*, which is the set that has no elements whatsoever. It is denoted by $\emptyset$ or by $\{\}$. Make sure not to confuse $\emptyset$ with $\{\emptyset\}$! The former has no elements, while the latter has one element. If we visualize the empty set as an empty paper bag, then we can visualize $\{\emptyset\}$ as a paper bag inside of a paper bag.

A.1.2 Subsets

If A and B are sets, then we say A is a *subset* of B (and write $A \subseteq B$) if every element of A is also an element of B. For example, the set of all integers is a subset of the set of all real numbers. It is always true that $\emptyset$ and A itself are subsets of A; these are the extreme cases for subsets. A general strategy for showing that $A \subseteq B$ is to let x be an arbitrary element of A, and then show that x must also be an element of B. A general strategy for showing that $A = B$ for two sets A and B is to show that each is a subset of the other.

A.1.3 Unions, intersections, and complements

The *union* of two sets A and B, written as $A \cup B$, is the set of all objects that are in A or B (or both). The *intersection* of A and B, written as $A \cap B$, is the set of all objects that are in both A and B. We say that A and B are *disjoint* if $A \cap B = \emptyset$. For n sets $A_1, \ldots, A_n$, the union $A_1 \cup A_2 \cup \cdots \cup A_n$ is the set of all objects that are in *at least one* of the A_j's, while the intersection $A_1 \cap A_2 \cap \cdots \cap A_n$ is the set of all objects that are in *all* of the A_j's.

In many applications, all the sets we're working with are subsets of some set S (in probability, this may be the set of all possible outcomes of some experiment). When S is clear from the context, we define the *complement* of a set A to be the set of all objects in S that are *not* in A; this is denoted by A^c.

Unions, intersections, and complements can be visualized easily using Venn diagrams, such as the one below. The union is the entire shaded region, while the intersection is the football-shaped region of points that are in both A and B. The complement of A is all points in the rectangle that are outside of A.

Note that the area of the region $A \cup B$ is the area of A plus the area of B, minus the area of $A \cap B$ (this is a basic form of what is called the *inclusion-exclusion principle*).

De Morgan's laws give an elegant, useful duality between unions and intersections:

$$(A_1 \cup A_2 \cup \cdots \cup A_n)^c = A_1^c \cap A_2^c \cap \cdots \cap A_n^c$$

$$(A_1 \cap A_2 \cap \cdots \cap A_n)^c = A_1^c \cup A_2^c \cup \cdots \cup A_n^c$$

It is much more important to *understand* De Morgan's laws than to memorize them!

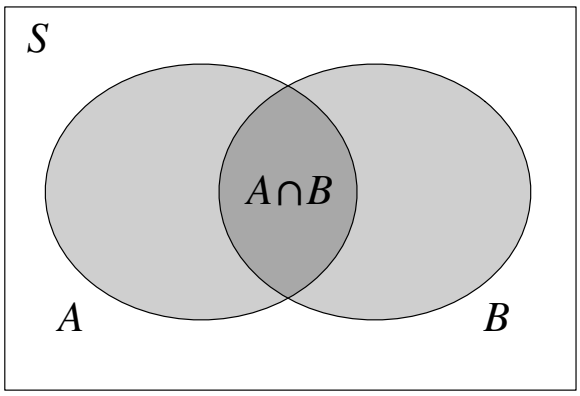


FIGURE A.1
A Venn diagram.

The first law says that not being in at least one of the A_j is the same thing as not being in A_1, nor being in A_2, nor being in A_3, etc. For example, let A_j be the set of people who like the jth Star Wars prequel (for $j \in \{1, 2, 3\}$). Then $(A_1 \cup A_2 \cup A_3)^c$ is the set of people for whom it is *not* the case that they like at least one of the prequels, but that's the same as $A_1^c \cap A_2^c \cap A_3^c$, the set of people who don't like *The Phantom Menace*, don't like *Attack of the Clones*, and don't like *Revenge of the Sith*.

The second law says that not being in all of the A_j is the same thing as being outside at least one of the A_j. For example, let the A_j be defined as in the previous paragraph. If it is not the case that you like all of the Star Wars prequels (making you a member of the set $(A_1 \cap A_2 \cap A_3)^c$), then there must be at least one prequel that you don't like (making you a member of the set $A_1^c \cup A_2^c \cup A_3^c$), and vice versa.

Proving the following facts about sets (not just drawing Venn diagrams, though they are very helpful for building intuition) is good practice:

1. $A \cap B$ and $A \cap B^c$ are disjoint, with $(A \cap B) \cup (A \cap B^c) = A$.
2. $A \cap B = A$ if and only if $A \subseteq B$.
3. $A \subseteq B$ if and only if $B^c \subseteq A^c$.

A.1.4 Partitions

A collection of subsets $A_1, \ldots, A_n$ of a set S is a *partition* of S if $A_1 \cup \cdots \cup A_n = S$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. In words, a partition is a collection of disjoint subsets whose union is the entire set. For example, the set of even numbers $\{0, 2, 4, \ldots\}$ and the set of odd numbers $\{1, 3, 5, \ldots\}$ form a partition of the set of nonnegative integers.

A.1.5 Cardinality

A set may be finite or infinite. If A is a finite set, we write $|A|$ for the number of elements in A, which is called its *size* or *cardinality*. For example, $|\{2, 4, 6, 8, 10\}| = 5$ since there are 5 elements in this set. A very useful fact is that A and B are finite sets, then

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

This is a form of the inclusion-exclusion result from Chapter 1. It says that to count how many elements are in the union of A and B, we can add the separate counts for each, and then adjust for the fact that we have double-counted the elements (if any) that are in both A and B.

Two sets A and B are said to have the *same size* or *same cardinality* if they can be put into one-to-one correspondence, i.e., if each element of A can be paired up with exactly one element of B, with no unpaired elements in either set. We say that A is *smaller than* B if there is *not* a one-to-one correspondence between A and B, but there *is* a one-to-one correspondence between A and a *subset* of B.

For example, suppose that we want to count the number of people in a movie theater with 100 seats. Assume that no one in the theater is standing, and no seat has more than one person in it. The obvious thing to do is to go around counting people one by one (though it's surprisingly easy to miss someone or accidentally count someone twice). But if every seat is occupied, then a much easier method is to note that there must be 100 people, since there are 100 seats and there is a one-to-one correspondence between people and seats. If some seats are empty, then there must be fewer than 100 people there.

This idea of looking at one-to-one correspondences makes sense both for finite and for infinite sets. Consider the perfect squares $1^2, 2^2, 3^2, \ldots$. Galileo pointed out the paradoxical result that on the one hand it seems like there are fewer perfect squares than positive integers (since every perfect square is a positive integer, but lots of positive integers aren't perfect squares), but on the other hand it seems like these two sets have the same size since they can be put into one-to-one correspondence: pair 1^2 with 1, pair 2^2 with 2, pair 3^2 with 3, etc.

The resolution of Galileo's paradox is to realize that intuitions about finite sets don't necessarily carry over to infinite sets. By definition, the set of all perfect squares and the set of all positive integers do have the same size. Another famous example of this is *Hilbert's hotel*. For any hotel in the real world the number of rooms is finite. If every room is occupied, there is no way to accommodate more guests, other than by cramming more people into already occupied rooms.

Now consider an imaginary hotel with an infinite sequence of rooms, numbered $1, 2, 3, \ldots$. Assume that all the rooms are occupied, and that a weary traveler arrives, looking for a room. Can the hotel give the traveler a room, without leaving any of the current guests without a room? Yes, one way is to have the guest in room n

move to room $n+1$, for all $n = 1, 2, 3, \ldots$. This frees up room 1, so the traveler can stay there.

What if *infinitely many* travelers arrive at the same time, such that their cardinality is the same as that of the positive integers (so we can label the travelers as traveler 1, traveler 2, ...)? The hotel could fit them in one by one by repeating the above procedure over and over again, but it would take forever (infinitely many moves) to accommodate everyone, and it would be bad for business to make the current guests keep moving over and over again. Can the room assignments be updated just *once* so that everyone has a room? Yes, one way is to have the guest in room n move to room to $2n$ for all $n = 1, 2, 3, \ldots$, and then have traveler n move into room $2n - 1$. In this way, the current guests occupy all the even-numbered rooms, and the new guests occupy all the odd-numbered rooms.

An infinite set is called *countably infinite* if it has the same cardinality as the set of all positive integers. A set is called *countable* if it is finite or countably infinite, and *uncountable* otherwise. The mathematician Cantor showed that not all infinite sets are the same size. In particular, the set of all real numbers is uncountable, as is any interval in the real line of positive length.

A.1.6 Cartesian product

The *Cartesian product* of two sets A and B is the set

$$A \times B = \{(a, b) : a \in A, b \in B\}.$$

For example, $[0, 1] \times [0, 1]$ is the square $\{(x, y) : x, y \in [0, 1]\}$, and $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ is two-dimensional Euclidean space.

A.2 Functions

Let A and B be sets. A *function* from A to B is a deterministic rule that, given an element of A as input, provides an element of B as an output. That is, a function from A to B is a machine that takes an x in A and "maps" it to some y in B. Different x's can map to the same y, but each x only maps to one y. Here A is called the *domain* and B is called the *target*. The notation $f : A \to B$ says that f is a function mapping A into B. The *range* of f is $\{y \in B : f(x) = y \text{ for some } x \in A\}$.

Of course, we have many familiar examples, such as the function f given by $f(x) = x^2$, for all real x. It is important to distinguish between f (the function) and $f(x)$ (the value of the function when evaluated at x). That is, f is a rule, while $f(x)$ is a number for each number x. The function g given by $g(x) = e^{-x^2/2}$ is exactly the

same as the function g given by $g(t) = e^{-t^2/2}$; what matters is the rule, not the name we use for the input.

A function f from the real line to the real line is *continuous* if $f(x) \to f(a)$ as $x \to a$, for any value of a. It is called *right-continuous* if this is true when approaching from the right, i.e., $f(x) \to f(a)$ as $x \to a$ while ranging over values with $x > a$.

In general though, A needn't consist of numbers, and f needn't be given by an explicit formula. For example, let A be the set of all positive-valued, continuous functions on $[0, 1]$, and f be the rule that takes a function in A as input, and gives the area under its curve (from 0 to 1) as output.

In probability, it is extremely useful to consider functions whose domains are the set of all possible outcomes of some experiment. It may be difficult to write down a formula for the function, but it's still valid as long as it's defined unambiguously.

A.2.1 One-to-one functions

Let f be a function from A to B. Then f is a *one-to-one* function if $f(x) \neq f(y)$ whenever $x \neq y$. That is, any two distinct inputs in A get mapped to two distinct outputs in B; for each y in B, there can be at most one x in A that maps to it.

Let f be a one-to-one function from A to B, and let C be the range of f (so C is the subset of B consisting of the elements of B that do get mapped to by an element of A). Then there is an *inverse function* $f^{-1} : C \to A$, defined by letting $f^{-1}(y)$ be the unique element $x \in A$ such that $f(x) = y$.

For example, let $f(x) = x^2$ for all real x. This is *not* a one-to-one function since, for example, $f(3) = f(-3)$. But now assume instead that the domain of f is chosen to be $[0, \infty)$, so we are defining f as a function from $[0, \infty)$ to $[0, \infty)$. Then f is one-to-one, and its inverse function is given by $f^{-1}(y) = \sqrt{y}$ for all $y \in [0, \infty)$.

A.2.2 Increasing and decreasing functions

Let $f : A \to \mathbb{R}$, where A is a set of real numbers. Then f is an *increasing* function if $x \leq y$ implies $f(x) \leq f(y)$ (for all $x, y \in A$). Note that these definition allows there to be regions where f is flat, e.g., the constant function that is equal to 42 everywhere is an increasing function. We say that f is *strictly increasing* if $x < y$ implies $f(x) < f(y)$. For example, $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^3$ is a strictly increasing function.

☣ **A.2.1.** Some references say "nondecreasing" or "weakly increasing" where we say "increasing", and "increasing" where we say "strictly increasing".

Similarly, f is a *decreasing* function if $x \leq y$ implies $f(x) \geq f(y)$, and is a *strictly decreasing* function if $x < y$ implies $f(x) > f(y)$. For example, $f : (0, \infty) \to (0, \infty)$ with $f(x) = 1/x$ is a strictly decreasing function.

A *monotone* function is a function that is either increasing or decreasing. A *strictly monotone* function is a function that is either strictly increasing or strictly increasing. Note that any strictly monotone function is one-to-one.

A.2.3 Even and odd functions

Let f be a function from $\mathbb{R}$ to $\mathbb{R}$. We say f is an *even function* if $f(x) = f(-x)$ for all x, and we say f is an *odd function* if $-f(x) = f(-x)$ for all x. If neither of these conditions is satisfied, then f is neither even nor odd. Figure A.2 shows the graphs of two even functions and two odd functions.

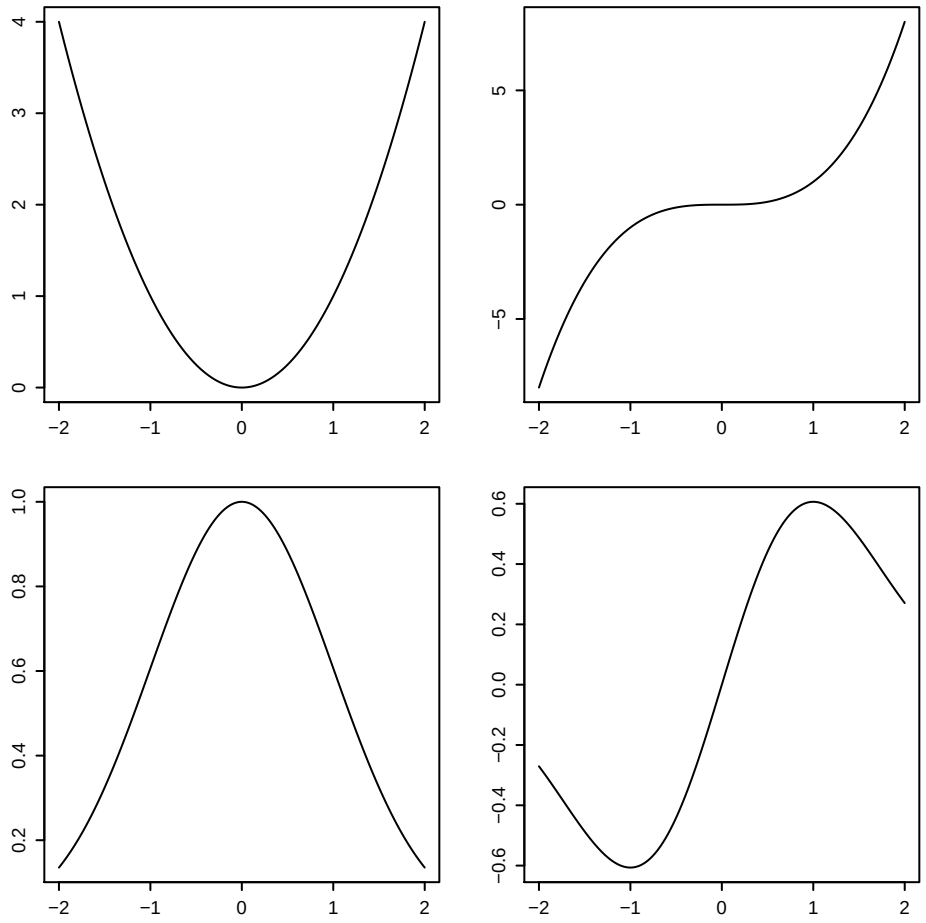


FIGURE A.2
Even and odd functions. The graphs on the left are even functions: $f(x) = x^2$ on the top and $f(x) = e^{-x^2/2}$ on the bottom. The graphs on the right are odd functions: $f(x) = x^3$ on the top and $f(x) = xe^{-x^2/2}$ on the bottom.

Even and odd functions have nice symmetry properties. The graph of an even func-

tion remains the same if you reflect it about the vertical axis, and the graph of an odd function remains the same if you rotate it 180 degrees around the origin.

Even functions have the property that for any a,

$$\int_{-a}^{a} f(x)dx = 2\int_{0}^{a} f(x)dx,$$

assuming the integral exists. This is because the area under the function from $-a$ to 0 is equal to the area under the function from 0 to a. Odd functions have the property that for any a,

$$\int_{-a}^{a} f(x)dx = 0,$$

again assuming the integral exists. This is because the area under the function from $-a$ to 0 cancels the area under the function from 0 to a.

A.2.4 Convex and concave functions

A function g whose domain is an interval I is *convex* if

$$g(px_1 + (1-p)x_2) \leq pg(x_1) + (1-p)g(x_2)$$

for all $x_1, x_2 \in I$ and $p \in (0,1)$. Geometrically, this says that if we draw a line segment connecting two points on the graph of g, then the line segment lies above the graph of g. If the derivative g' exists, then an equivalent definition is that every tangent line to the graph of g lies below the graph. If g'' exists, then an equivalent definition is that $g''(x) \geq 0$ for all $x \in I$. An example is shown in Figure A.3. A simple example is $g(x) = x^2$, whose second derivative is $g''(x) = 2 > 0$.

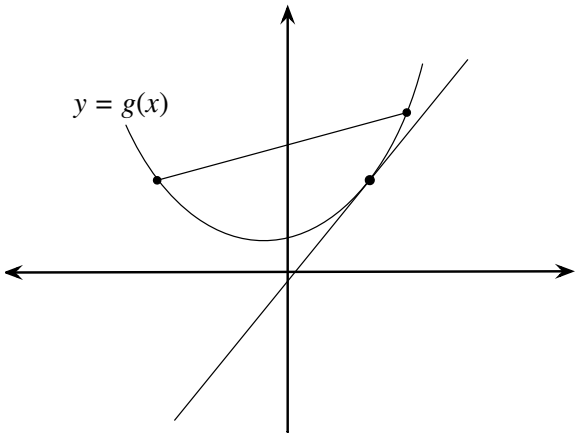


FIGURE A.3
Graph of a convex function g. We have $g''(x) \geq 0$. Any line segment connecting two points on the curve lies above the curve. Any tangent line lies below the curve.

A function g is *concave* if $-g$ is convex. If g'' exists, then g is concave if and only if $g''(x) \leq 0$ for all x in the domain. For example, $g(x) = \log(x)$ defines a concave function on $(0, \infty)$ since $g''(x) = -1/x^2 < 0$ for all $x \in (0, \infty)$.

A.2.5 Exponential and logarithmic functions

Exponential functions are functions of the form $f(x) = a^x$ for some number $a > 0$. If $a > 1$, the function is increasing, and if $0 < a < 1$, the function is decreasing. The most common exponential function we'll work with is $f(x) = e^x$, and a very useful limit result to know is that

$$\left(1 + \frac{x}{n}\right)^n \to e^x$$

as $n \to \infty$, for any real number x. This has an interpretation in terms of a bank paying compound interest on a deposit: as compounding occurs more and more times per year, the growth rate approaches exponential growth.

Here are some properties of exponential functions:

1. $a^x a^y = a^{x+y}$.
2. $a^x b^x = (ab)^x$.
3. $(a^x)^y = a^{xy}$.

The inverse of an exponential function is a logarithmic function: for positive y, $\log_a y$ is defined to be the number x such that $a^x = y$. Throughout this book, when we write $\log y$ without explicitly specifying the base, we are referring to the *natural logarithm* (base e).

Here are some properties of logarithms:

1. $\log_a x + \log_a y = \log_a xy$.
2. $\log_a x^n = n \log_a x$.
3. $\log_a x = \frac{\log x}{\log a}$.

A.2.6 Floor function and ceiling function

The floor function is defined by letting $\lfloor x \rfloor$ be the greatest integer less than or equal to x. That is, it says to round down to an integer. For example, $\lfloor 3.14 \rfloor = 3$, $\lfloor -1.3 \rfloor = -2$, and $\lfloor 5 \rfloor = 5$. (Some books denote the floor function by $[x]$ but this is bad notation since it does not suggest a corresponding notation for the ceiling function, and since square brackets are also used for other purposes.)

The ceiling function is defined by letting $\lceil x \rceil$ be the smallest integer greater than or equal to x. For example, $\lceil 3.14 \rceil = 4$, $\lceil -1.3 \rceil = -1$, and $\lceil 5 \rceil = 5$.

A.2.7 Factorial function and gamma function

The factorial function takes a positive integer n and returns the product of the integers from 1 to n, denoted $n!$ and read "n factorial":

$$n! = 1 \cdot 2 \cdot 3 \ldots n.$$

Also, we define $0! = 1$. This convention makes sense since if we think of $n!$ as the number of ways in which n people can line up, there is 1 way for $n = 0$ (this just means there is no one there, so the line is empty). It is also very helpful since, for example, we can then say that $n!/(n-1)! = n$ for all positive integers n without running into trouble when $n = 1$.

The factorial function grows extremely quickly as n grows. A famous, useful approximation for factorials is *Stirling's formula*,

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

The ratio of the two sides converges to 1 as $n \to \infty$. For example, direct calculation gives $52! \approx 8.066 \times 10^{67}$, while Stirling's formula says $52! \approx 8.053 \times 10^{67}$.

The *gamma function* Γ generalizes the factorial function to positive real numbers; it is defined by

$$\Gamma(a) = \int_0^\infty x^a e^{-x} \frac{dx}{x}, \quad a > 0,$$

and has the property that

$$\Gamma(n) = (n-1)!$$

for all positive integers n. Also, $\Gamma(1/2) = \sqrt{\pi}$.

An important property of the gamma function, which generalizes the fact that $n! = n \cdot (n-1)!$, is that

$$\Gamma(a+1) = a\Gamma(a)$$

for all $a > 0$. See Chapter 8 for more about the gamma function.

A.3 Matrices

> Neo: What is the Matrix?
> Trinity: The answer is out there, Neo, and it's looking for you, and it will find you if you want it to.
> – *The Matrix* (film from 1999)

A matrix is a rectangular array of numbers, such as $\begin{pmatrix} 3 & 1/e \\ 2\pi & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 3 \end{pmatrix}$. We

say that the dimensions of a matrix are m by n (also written $m \times n$) if it has m rows and n columns (so the former example is 2 by 2, while the latter is 2 by 3). The matrix is called *square* if $m = n$. If $m = 1$, we have a *row vector*; if $n = 1$, we have a *column vector*.

A.3.1 Matrix addition and multiplication

To *add* two matrices A and B with the same dimensions, just add the corresponding entries, e.g.,

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix}.$$

When we *multiply* an m by n matrix A by an n by r matrix B, we obtain an m by r matrix AB; note that matrix multiplication is only well-defined when the number of columns of A equals the number of rows of B. The row i, column j entry of AB is $\sum_{k=1}^{n} a_{ik}b_{kj}$, where a_{ij} and b_{ij} are the row i, column j entries of A and B, respectively. For example, here is how to multiply a 2×3 matrix by a 3×1 vector:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 \cdot 7 + 2 \cdot 8 + 3 \cdot 9 \\ 4 \cdot 7 + 5 \cdot 8 + 6 \cdot 9 \end{pmatrix} = \begin{pmatrix} 50 \\ 122 \end{pmatrix}.$$

Note that AB may not equal BA, even if both are defined. To multiply a matrix A by a scalar, just multiply each entry by that scalar.

The *transpose* of a matrix A is the matrix whose row i, column j entry is the row j, column i entry of A. It is denoted by A' and read as "A transpose" or "A prime". The rows of A are the columns of A', and the columns of A are the rows of A'. If A and B are matrices such that the product AB is defined, then $(AB)' = B'A'$.

The *determinant* of a 2 by 2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is defined to be $ad - bc$. Determinants can also be defined for n by n matrices, in a recursive manner not reviewed here.

A.3.2 Eigenvalues and eigenvectors

An *eigenvalue* of an $n \times n$ matrix A is a number λ such that

$$A\mathbf{v} = \lambda\mathbf{v}$$

for some $n \times 1$ column vector $\mathbf{v}$, where the elements of $\mathbf{v}$ are not all zero. The vector $\mathbf{v}$ is called an *eigenvector* of A, or sometimes a *right eigenvector*. (A *left eigenvector* of A would be a row vector $\mathbf{w}$ satisfying $\mathbf{w}A = \lambda\mathbf{w}$ for some λ.) This definition says that when A and $\mathbf{v}$ are multiplied, $\mathbf{v}$ just gets stretched by the constant λ.

Some matrices have no real eigenvalues, but the *Perron-Frobenius theorem* tells us that in a special case that is of particular interest to us in Chapter 11, eigenvalues exist and have nice properties. Let A be a square matrix whose entries are nonnegative and whose rows sum to 1. Further assume that for all i and j, there exists $k \geq 1$ such that the row i, column j entry of A^k is positive. Then the Perron-Frobenius theorem says that 1 is an eigenvalue of A, 1 is in fact the largest eigenvalue of A, and there is a corresponding eigenvector whose entries are all positive.

A.4 Difference equations

A *difference equation* describes a sequence of numbers recursively in terms of earlier terms in the sequence. For example, $a_0, a_1, \ldots$ is a *Fibonacci sequence* if

$$a_i = a_{i-1} + a_{i-2}$$

for all $i \geq 2$. There are infinitely many Fibonacci sequences, but such a sequence is uniquely determined after a_0 and a_1 are specified (these are called the *initial conditions* or *boundary conditions*). For example, $a_0 = 0, a_1 = 1$ yields the Fibonacci sequence $0, 1, 1, 2, 3, 5, 8, 13, \ldots$.

Difference equations often arise in probability, especially when applying LOTP. In this section, we will show how to solve difference equations of the form

$$p_i = p \cdot p_{i+1} + q \cdot p_{i-1},$$

where $p \neq 0$ and $q = 1 - p$. (This equation comes up in the gambler's ruin problem.) The first step is to *guess* a solution of the form $p_i = x^i$. Plugging this into the above, we have

$$x^i = p \cdot x^{i+1} + q \cdot x^{i-1},$$

which reduces to $px^2 - x + q = 0$. This is called the *characteristic equation*, and the solution to the difference equation depends on whether the characteristic equation has one or two distinct roots. If there are two distinct roots r_1 and r_2, then the solution is of the form

$$p_i = ar_1^i + br_2^i,$$

for some constants a and b. If there is only one distinct root r, then the solution is of the form

$$p_i = ar^i + bir^i.$$

In our case, the characteristic equation has roots 1 and q/p, since

$$\frac{1 \pm \sqrt{1 - 4p(1-p)}}{2p} = \frac{1 \pm \sqrt{(2p-1)^2}}{2p} = \frac{1 \pm |2p-1|}{2p}$$

is $(1 + 2p - 1)/(2p) = 1$ or $(2 - 2p)/(2p) = q/p$. The roots are distinct if $p \neq q$ and are both equal to 1 if $p = q$. So we have

$$p_i = \begin{cases} a + b\left(\frac{q}{p}\right)^i, & p \neq q, \\ a + bi, & p = q. \end{cases}$$

This is called the *general solution* of the difference equation, since we have not yet specified the constants a and b. To get a *specific solution*, we need to know two points in the sequence in order to solve for a and b.

A.5 Differential equations

Differential equations are the continuous version of difference equations. A differential equation uses derivatives to describe a function or collection of functions. For example, the differential equation

$$\frac{dy}{dx} = 3y$$

describes a collection of functions that have the following property: the instantaneous rate of change of the function at any point (x, y) is equal to cy. This is an example of a *separable* differential equation because we can separate the x's and y's, putting them on opposite sides of the equation:

$$\frac{dy}{y} = 3dx.$$

Now we can integrate both sides, giving $\log y = 3x + c$, or equivalently,

$$y = Ce^{3x},$$

where C is any constant. This is called the *general solution* of the differential equation, and it tells us that all functions satisfying the differential equation are of the form $y = Ce^{3x}$ for some C. To get a *specific solution*, we need to specify one point on the graph, which allows us to solve for C.

Separable differential equations are a special case. In general, it may not be possible to rearrange the x's and y's to be on opposite sides, so there are additional methods for solving non-separable differential equations, which we will not review here.

A.6 Partial derivatives

If you can do ordinary derivatives, you can do partial derivatives: just hold all the other input variables constant except for the one you're differentiating with respect to. For example, let $f(x, y) = y \sin(x^2+y^3)$. Then the partial derivative with respect to x is

$$\frac{\partial f(x,y)}{\partial x} = 2xy\cos(x^2 + y^3),$$

and the partial derivative with respect to y is

$$\frac{\partial f(x,y)}{\partial y} = \sin(x^2 + y^3) + 3y^3\cos(x^2 + y^3).$$

The *Jacobian* of a function which maps $(x_1, \ldots, x_n)$ to $(y_1, \ldots, y_n)$ is the n by n matrix of all possible partial derivatives, given by

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}.$$

A.7 Multiple integrals

If you can do single integrals, you can do multiple integrals: just do more than one integral, holding variables other than the current variable of integration constant. For example,

$$\begin{aligned} \int_0^1 \int_0^y (x-y)^2 dxdy &= \int_0^1 \int_0^y (x^2 - 2xy + y^2) dxdy \\ &= \int_0^1 (x^3/3 - x^2y + xy^2)\Big|_0^y dy \\ &= \int_0^1 (y^3/3 - y^3 + y^3) dy \\ &= \frac{1}{12}. \end{aligned}$$

A.7.1 Change of order of integration

We can also integrate in the other order, $dydx$ rather than $dxdy$, as long as we are careful about the limits of integration. Since we're integrating over all (x, y) with x

and y between 0 and 1 such that $x \leq y$, to integrate the other way we write

$$\begin{aligned}\int_0^1 \int_x^1 (x-y)^2 dydx &= \int_0^1 \int_x^1 (x^2 - 2xy + y^2) dydx \\ &= \int_0^1 (x^2 y - xy^2 + y^3/3)\Big|_x^1 dx \\ &= \int_0^1 (x^2 - x + 1/3 - x^3 + x^3 - x^3/3) dx \\ &= \left(x^3/3 - x^2/2 + x/3 - \frac{x^4}{12}\right)\Big|_0^1 \\ &= \frac{1}{12}.\end{aligned}$$

A.7.2 Change of variables

In making a change of variables with multiple integrals, a Jacobian is needed. Let's state the two-dimensional version, for concreteness. Suppose we make a change of variables (transformation) from (x, y) to (u, v), say with $x = g(u, v), y = h(u, v)$. Then

$$\int\int f(x,y)dxdy = \int\int f(g(u,v), h(u,v)) \left|\frac{\partial(x,y)}{\partial(u,v)}\right| dudv,$$

over the appropriate limits of integration, where $\left|\frac{\partial(x,y)}{\partial(u,v)}\right|$ is the absolute value of the determinant of the Jacobian. We assume that the partial derivatives exist and are continuous, and that the determinant is nonzero.

For example, let's find the area of a circle of radius 1. To find the area of a region, we just need to integrate 1 over that region (so any difficulty comes from the limits of integration; the function we're integrating is just the constant 1). So the area is

$$\iint\limits_{x^2+y^2\leq 1} 1 dxdy = \int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} 1 dxdy = 2\int_{-1}^1 \sqrt{1-y^2} dy.$$

Note that the limits for the inner variable (x) of the double integral can depend on the outer variable (y), while the outer limits are constants. The last integral can be done with a trigonometric substitution, but instead let's simplify the problem by transforming to polar coordinates: let

$$x = r\cos\theta, y = r\sin\theta,$$

where r is the distance from (x, y) to the origin and $\theta \in [0, 2\pi)$ is the angle. The Jacobian matrix of this transformation is

$$\frac{d(x,y)}{d(r,\theta)} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta, \end{pmatrix}$$

so the absolute value of the determinant is $r(\cos^2\theta + \sin^2\theta) = r$. That is, $dxdy$ becomes $rdrd\theta$. So the area of the circle is

$$\int_0^{2\pi}\int_0^1 rdrd\theta = \int_0^{2\pi}\frac{1}{2}d\theta = \pi.$$

For a circle of radius r, it follows immediately that the area is πr^2 since we can imagine converting our units of measurement to the unit for which the radius is 1.

This may seem like a lot of work just to get such a familiar result, but it served as illustration and with similar methods, we can get the volume of a ball in any number of dimensions! It turns out that the volume of a ball of radius 1 in n dimensions is $\frac{\pi^{n/2}}{\Gamma(n/2+1)}$, where Γ is the gamma function (see Section A.2.7).

A.8 Sums

> "So you've got to the end of our race-course?" said the Tortoise. "Even though it does consist of an infinite series of distances? I thought some wiseacre or another had proved that the thing couldn't be done?"
> "It can be done," said Achilles; "It has been done! Solvitur ambulando. You see, the distances were constantly diminishing."
> – Lewis Carroll [3]

There are several kinds of sums that come up frequently in probability.

A.8.1 Geometric series

A series of the form $\sum_{n=0}^{\infty} x^n$ is called a *geometric series*. For $|x| < 1$, we have

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

The series diverges if $|x| \geq 1$. A series of the same form except with finitely many terms is called a *finite geometric series*. For $x \neq 1$ the sum is

$$\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}.$$

A.8.2 Taylor series for e^x

The Taylor series for e^x is

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x, \text{ for all } x.$$

A.8.3 Harmonic series and other sums with a fixed exponent

It is also useful to know that $\sum_{n=1}^{\infty} 1/n^c$ converges for $c > 1$ and diverges for $c \leq 1$. For $c = 1$, this is called the *harmonic series*. The sum of the first n terms of the harmonic series can be approximated using

$$\sum_{k=1}^{n} \frac{1}{k} \approx \log(n) + \gamma$$

for n large, where $\gamma \approx 0.577$.

The sum of the first n positive integers is

$$\sum_{k=1}^{n} k = n(n+1)/2.$$

For squares of integers, we have

$$\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6.$$

For cubes of integers, amazingly, the sum is the square of the sum of the first n positive integers! That is,

$$\sum_{k=1}^{n} k^3 = (n(n+1)/2)^2.$$

A.8.4 Binomial theorem

The *binomial theorem* states that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k},$$

where $\binom{n}{k}$ is a *binomial coefficient*, defined as the number of ways to choose k objects out of n, with order not mattering. An explicit formula for $\binom{n}{k}$ in terms of factorials is

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

A proof of the binomial theorem is given in Example 1.4.17.

A.9 Pattern recognition

Much of math and statistics is really about *pattern recognition*: seeing the essential structure of a problem, recognizing when one problem is essentially the same as another problem (just in a different guise), noticing symmetry, and so on. We will see many examples of this kind of thinking in this book. For example, suppose we have the series $\sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \lambda^k / k!$, with λ a positive constant. The $e^{-\lambda}$ can be taken out from the sum, and then the structure of the series exactly matches up with the structure of the Taylor series for e^x. Therefore

$$\sum_{k=0}^{\infty} \frac{e^{tk} e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)},$$

valid for all real t.

Similarly, suppose we want the Taylor series for $1/(1-x^3)$ about $x = 0$. It would be tedious to start taking derivatives of this function. Instead, note that this function is reminiscent of the result of summing a geometric series. Therefore

$$\frac{1}{1-x^3} = \sum_{n=0}^{\infty} x^{3n},$$

valid for $|x^3| < 1$ (which is equivalent to $|x| < 1$). What matters is the structure, not what names we use for variables!

A.10 Common sense and checking answers

It is very easy to make mistakes in probability, so checking answers is especially important. Some useful strategies for checking answers are (a) seeing whether the answer makes sense intuitively (though as we have often seen in this book, probability has many results that seem counterintuitive at first), (b) making sure the answer isn't a category error or a case of a biohazard, (c) trying out simple examples, (d) trying out extreme examples, and (e) looking for alternative methods to solve the problem (including methods that might only give a bound or approximation, such as using an inequality from Chapter 10 or running a simulation).

For example, here are some wrong arguments where the claim defies common sense. These illustrate the importance of doing sanity checks.

1. $$\text{“}\int_{-1}^{1} \frac{1}{x^2} dx = (-x^{-1})\Big|_{-1}^{1} = -2.\text{”}$$

This makes no sense intuitively, since $1/x^2$ is a *positive quantity*; it would be a miracle if its integral were negative!

2. "Let us find $\int \frac{1}{x} dx$ using integration by parts. Let $u = 1/x, dv = dx$. Then

$$\int \frac{1}{x} dx = uv - \int v du = 1 + \int \frac{x}{x^2} dx = 1 + \int \frac{1}{x} dx,$$

which implies $0 = 1$ since we can cancel out $\int \frac{1}{x} dx$ from both sides."

3. What is wrong with the following "proof" that all horses are the same color? (This example is due to George Pólya.) "Let n be the number of horses, and use induction on n to "prove" that in every group of n horses, all the horses have the same color. For the base case $n = 1$, there is only one horse, which clearly must be its own color. Now assume the claim is true for $n = k$, and show that it is true for $n = k + 1$. Consider a group of $k + 1$ horses. Excluding the oldest horse, we have k horses, which by the inductive hypothesis must all be the same color. Excluding the youngest horse, we have k horses, which again by the inductive hypothesis must have the same color. Thus, all the horses have the same color."

4. "There are 19 integers between 12 and 31 (inclusive) since $31 - 12 = 19$. More generally, there are $m - n$ numbers in the list $n, n+1, \ldots, m$ if n and m are integers with $m \geq n$." Such off-by-one errors are very common in math and programming, but they can easily be avoided by checking simple and extreme cases, e.g., there are 10 integers from 1 to 10 (inclusive) even though $10 - 1 = 9$.

Bibliography

[1] Donald J. Albers and Gerald L. Alexanderson. *More Mathematical People: Contemporary Conversations*. Academic Press, 1990.

[2] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

[3] Lewis Carroll. What the Tortoise Said to Achilles. *Mind*, 4(14):278–290, 1895.

[4] Jian Chen and Jeffrey S. Rosenthal. Decrypting classical cipher text using Markov chain Monte Carlo. *Statistics and Computing*, 22(2):397–413, 2012.

[5] William G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 1968.

[6] Persi Diaconis. Statistical problems in ESP research. *Science*, 201(4351):131–136, 1978.

[7] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

[8] Persi Diaconis, Susan Holmes, and Richard Montgomery. Dynamical bias in the coin toss. *SIAM Review*, 49(2):211–235, 2007.

[9] Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435, 1976.

[10] Bradley Efron and Ronald Thisted. Did Shakespeare write a newly-discovered poem? *Biometrika*, 74:445–455, 1987.

[11] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 2013.

[12] Andrew Gelman and Deborah Nolan. You can load a die, but you can't bias a coin. *The American Statistician*, 56(4):308–311, 2002.

[13] Andrew Gelman, Boris Shor, Joseph Bafumi, and David K. Park. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do (Expanded Edition)*. Princeton University Press, 2009.

[14] Gerd Gigerenzer and Ulrich Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684, 1995.

[15] Prakash Gorroochurn. *Classic Problems of Probability*. John Wiley & Sons, 2012.

[16] Richard Hamming. You and your research. *IEEE Potentials*, pages 37–40, October 1993.

[17] David P. Harrington. The randomized clinical trial. *Journal of the American Statistical Association*, 95(449):312–315, 2000.

[18] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[19] Oscar Mandel. *Chi Po and the Sorcerer: A Chinese Tale for Children and Philosophers*. Charles E. Tuttle Company, 1964.

[20] T.J. Mathews and Brady E. Hamilton. Trend analysis of the sex ratio at birth in the United States. *National Vital Statistics Reports*, 53(20):1–17, 2005.

[21] Pierre Rémond de Montmort. *Essay d'Analyse sur les Jeux de Hazard*. Quilau, Paris, 1708.

[22] John Allen Paulos. *Innumeracy: Mathematical Illiteracy and Its Consequences*. Macmillan, 1988.

[23] Horst Rinne. *The Weibull Distribution: A Handbook*. CRC Press, 2008.

[24] James G. Sanderson. Testing ecological patterns. *American Scientist*, 88:332–339, 2000.

[25] Nate Silver. *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*. Penguin, 2012.

[26] Tom W. Smith, Peter Marsden, Michael Hout, and Jibum Kim. General social surveys, 1972–2012. *Sponsored by National Science Foundation. NORC, Chicago: National Opinion Research Center*, 2013.

[27] Stephen M. Stigler. Isaac Newton as a probabilist. *Statistical Science*, 21(3):400–403, 2006.

[28] Tom Stoppard. *Rosencrantz & Guildenstern Are Dead*. Samuel French, Inc., 1967.

[29] R.J. Stroeker. On the sum of consecutive cubes being a perfect square. *Compositio Mathematica*, 97:295–307, 1995.

[30] Amos Tversky and Daniel Kahneman. Causal schemas in judgments under uncertainty. In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

[31] Herbert S. Wilf. *generatingfunctionology*. A K Peters/CRC Press, 3rd edition, 2005.