# Chapter 1

# Probability Models

**CHAPTER OUTLINE**

This chapter introduces the basic concept of the entire course, namely probability. We discuss why probability was introduced as a scientific concept and how it has been formalized mathematically in terms of a probability model. Following this we develop some of the basic mathematical results associated with the probability model.

## 1.1 | Probability: A Measure of Uncertainty

Often in life we are confronted by our own ignorance. Whether we are pondering tonight's traffic jam, tomorrow's weather, next week's stock prices, an upcoming election, or where we left our hat, often we do not know an outcome with certainty. Instead, we are forced to guess, to estimate, to hedge our bets.

Probability is the science of uncertainty. It provides precise mathematical rules for understanding and analyzing our own ignorance. It does not tell us tomorrow's weather or next week's stock prices; rather, it gives us a framework for working with our limited knowledge and for making sensible decisions based on what we do and do not know.

To say there is a 40% chance of rain tomorrow is not to know tomorrow's weather. Rather, it is to know what we do not know about tomorrow's weather.

In this text, we will develop a more precise understanding of what it means to say there is a 40% chance of rain tomorrow. We will learn how to work with ideas of randomness, probability, expected value, prediction, estimation, etc., in ways that are sensible and mathematically clear.

1

There are also other sources of randomness besides uncertainty. For example, computers often use *pseudorandom numbers* to make games fun, simulations accurate, and searches efficient. Also, according to the modern theory of quantum mechanics, the makeup of atomic matter is in some sense *truly* random. All such sources of randomness can be studied using the techniques of this text.

Another way of thinking about probability is in terms of *relative frequency*. For example, to say a coin has a 50% chance of coming up heads can be interpreted as saying that, if we flipped the coin many, many times, then approximately half of the time it would come up heads. This interpretation has some limitations. In many cases (such as tomorrow's weather or next week's stock prices), it is impossible to repeat the experiment many, many times. Furthermore, what precisely does "approximately" mean in this case? However, despite these limitations, the relative frequency interpretation is a useful way to think of probabilities and to develop intuition about them.

Uncertainty has been with us forever, of course, but the mathematical theory of probability originated in the seventeenth century. In 1654, the Paris gambler Le Chevalier de Méré asked Blaise Pascal about certain probabilities that arose in gambling (such as, if a game of chance is interrupted in the middle, what is the probability that each player would have won had the game continued?). Pascal was intrigued and corresponded with the great mathematician and lawyer Pierre de Fermat about these questions. Pascal later wrote the book *Traité du Triangle Arithmetique*, discussing binomial coefficients (Pascal's triangle) and the binomial probability distribution.

At the beginning of the twentieth century, Russians such as Andrei Andreyevich Markov, Andrey Nikolayevich Kolmogorov, and Pafnuty L. Chebyshev (and American Norbert Wiener) developed a more formal mathematical theory of probability. In the 1950s, Americans William Feller and Joe Doob wrote important books about the mathematics of probability theory. They popularized the subject in the western world, both as an important area of pure mathematics and as having important applications in physics, chemistry, and later in computer science, economics, and finance.

### 1.1.1 | Why Do We Need Probability Theory?

Probability theory comes up very often in our daily lives. We offer a few examples here.

Suppose you are considering buying a "Lotto 6/49" lottery ticket. In this lottery, you are to pick six distinct integers between 1 and 49. Another six distinct integers between 1 and 49 are then selected at random by the lottery company. If the two sets of six integers are identical, then you win the jackpot.

After mastering Section 1.4, you will know how to calculate that the probability of the two sets matching is equal to one chance in 13,983,816. That is, it is about 14 million times more likely that you will not win the jackpot than that you will. (These are not very good odds!)

Suppose the lottery tickets cost $1 each. After mastering expected values in Chapter 3, you will know that you should not even *consider* buying a lottery ticket unless the jackpot is more than $14 million (which it usually is not). Furthermore, if the jackpot is ever more than $14 million, then likely many other people will buy lottery tickets

that week, leading to a larger probability that you will have to *share* the jackpot with other winners even if you do win — so it is probably not in your favor to buy a lottery ticket even then.

Suppose instead that a "friend" offers you a bet. He has three cards, one red on both sides, one black on both sides, and one red on one side and black on the other. He mixes the three cards in a hat, picks one at random, and places it flat on the table with only one side showing. Suppose that one side is red. He then offers to bet his $4 against your $3 that the other side of the card is also red.

At first you might think it sounds like the probability that the other side is also red is 50%; thus a good bet. However, after mastering conditional probability (Section 1.5), you will know that, conditional on one side being red, the conditional probability that the other side is also red is equal to 2/3. So, by the theory of expected values (Chapter 3), you will know that you should not accept your "friend's" bet.

Finally, suppose your "friend" suggests that you flip a coin one thousand times. Your "friend" says that if the coin comes up heads at least six hundred times, then he will pay you $100; otherwise, you have to pay him just $1.

At first you might think that, while 500 heads is the most likely, there is still a *reasonable* chance that 600 heads will appear — at least good enough to justify accepting your friend's $100 to $1 bet. However, after mastering the laws of large numbers (Chapter 4), you will know that as the number of coin flips gets large, it becomes more and more likely that the number of heads is very close to half of the total number of coin flips. In fact, in this case, there is less than one chance in ten billion of getting more than 600 heads! Therefore, you should not accept this bet, either.

As these examples show, a good understanding of probability theory will allow you to correctly assess probabilities in everyday situations, which will in turn allow you to make wiser decisions. It might even save you money!

Probability theory also plays a key role in many important applications of science and technology. For example, the design of a nuclear reactor must be such that the escape of radioactivity into the environment is an extremely rare event. Of course, we would like to say that it is categorically impossible for this to ever happen, but reactors are complicated systems, built up from many interconnected subsystems, each of which we know will fail to function properly at some time. Furthermore, we can never definitely say that a natural event like an earthquake cannot occur that would damage the reactor sufficiently to allow an emission. The best we can do is try to quantify our uncertainty concerning the failures of reactor components or the occurrence of natural events that would lead to such an event. This is where probability enters the picture. Using probability as a tool to deal with the uncertainties, the reactor can be designed to ensure that an unacceptable emission has an extremely small probability — say, once in a billion years — of occurring.

The gambling and nuclear reactor examples deal essentially with the concept of *risk* — the risk of losing money, the risk of being exposed to an injurious level of radioactivity, etc. In fact, we are exposed to risk all the time. When we ride in a car, or take an airplane flight, or even walk down the street, we are exposed to risk. We know that the risk of injury in such circumstances is never zero, yet we still engage in these activities. This is because we intuitively realize that the probability of an accident occurring is extremely low.

So we are using probability every day in our lives to assess risk. As the problems we face, individually or collectively, become more complicated, we need to refine and develop our rough, intuitive ideas about probability to form a clear and precise approach. This is why probability theory has been developed as a subject. In fact, the insurance industry has been developed to help us cope with risk. Probability is the tool used to determine what you pay to reduce your risk or to compensate you or your family in case of a personal injury.

### Summary of Section 1.1

- Probability theory provides us with a precise understanding of uncertainty.
- This understanding can help us make predictions, make better decisions, assess risk, and even make money.

### DISCUSSION TOPICS

**1.1.1** Do you think that tomorrow's weather and next week's stock prices are "really" random, or is this just a convenient way to discuss and analyze them?

**1.1.2** Do you think it is possible for probabilities to depend on who is observing them, or at what time?

**1.1.3** Do you find it surprising that probability theory was not discussed as a mathematical subject until the seventeenth century? Why or why not?

**1.1.4** In what ways is probability important for such subjects as physics, computer science, and finance? Explain.

**1.1.5** What are examples from your own life where thinking about probabilities did save — or could have saved — you money or helped you to make a better decision? (List as many as you can.)

**1.1.6** Probabilities are often depicted in popular movies and television programs. List as many examples as you can. Do you think the probabilities were portrayed there in a "reasonable" way?

## 1.2 | Probability Models

A formal definition of probability begins with a *sample space*, often written $S$. This sample space is any set that lists all possible *outcomes* (or, *responses*) of some unknown experiment or situation. For example, perhaps

$$S = \{\text{rain, snow, clear}\}$$

when predicting tomorrow's weather. Or perhaps $S$ is the set of all positive real numbers, when predicting next week's stock price. The point is, $S$ can be any set at all, even an infinite set. We usually write $s$ for an element of $S$, so that $s \in S$. Note that $S$ describes only those things that we are interested in; if we are studying weather, then rain and snow are in $S$, but tomorrow's stock prices are not.

A probability model also requires a collection of *events*, which are subsets of $S$ to which probabilities can be assigned. For the above weather example, the subsets {rain}, {snow}, {rain, snow}, {rain, clear}, {rain, snow, clear}, and even the empty set $\emptyset = \{\ \}$, are all examples of subsets of $S$ that could be events. Note that here the comma means "or"; thus, {rain, snow} is the event that it will rain *or* snow. We will generally assume that *all* subsets of $S$ are events. (In fact, in complicated situations there are some technical restrictions on what subsets can or cannot be events, according to the mathematical subject of measure theory. But we will not concern ourselves with such technicalities here.)

Finally, and most importantly, a probability model requires a *probability measure*, usually written $P$. This probability measure must assign, to each event $A$, a probability $P(A)$. We require the following properties:

1. $P(A)$ is always a nonnegative real number, between 0 and 1 inclusive.

2. $P(\emptyset) = 0$, i.e., if $A$ is the empty set $\emptyset$, then $P(A) = 0$.

3. $P(S) = 1$, i.e., if $A$ is the entire sample space $S$, then $P(A) = 1$.

4. $P$ is *(countably) additive*, meaning that if $A_1, A_2, \ldots$ is a finite or countable sequence of disjoint events, then

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots . \qquad (1.2.1)$$

The first of these properties says that we shall measure all probabilities on a scale from 0 to 1, where 0 means impossible and 1 (or 100%) means certain. The second property says the probability that *nothing* happens is 0; in other words, it is impossible that *no* outcome will occur. The third property says the probability that *something* happens is 1; in other words, it is certain that *some* outcome must occur.

The fourth property is the most subtle. It says that we can calculate probabilities of complicated events by adding up the probabilities of smaller events, provided those smaller events are *disjoint* and together contain the entire complicated event. Note that events are *disjoint* if they contain no outcomes in common. For example, {rain} and {snow, clear} are disjoint, whereas {rain} and {rain, clear} are not disjoint. (We are assuming for simplicity that it cannot both rain *and* snow tomorrow.) Thus, we should have $P(\{\text{rain}\}) + P(\{\text{snow, clear}\}) = P(\{\text{rain, snow, clear}\})$, but do *not* expect to have $P(\{\text{rain}\}) + P(\{\text{rain, clear}\}) = P(\{\text{rain, rain, clear}\})$ (the latter being the same as $P(\{\text{rain, clear}\})$).

We now formalize the definition of a probability model.

**Definition 1.2.1** A *probability model* consists of a nonempty set called the sample space $S$; a collection of events that are subsets of $S$; and a probability measure $P$ assigning a probability between 0 and 1 to each event, with $P(\emptyset) = 0$ and $P(S) = 1$ and with $P$ additive as in (1.2.1).

**EXAMPLE 1.2.1**

Consider again the weather example, with $S = \{$rain, snow, clear$\}$. Suppose that the probability of rain is 40%, the probability of snow is 15%, and the probability of a clear day is 45%. We can express this as $P(\{$rain$\}) = 0.40$, $P(\{$snow$\}) = 0.15$, and $P(\{$clear$\}) = 0.45$.

For this example, of course $P(\emptyset) = 0$, i.e., it is impossible that *nothing* will happen tomorrow. Also $P(\{$rain, snow, clear$\}) = 1$, because we are assuming that exactly *one* of rain, snow, or clear must occur tomorrow. (To be more realistic, we might say that we are predicting the weather at exactly 11:00 A.M. tomorrow.) Now, what is the probability that it will rain *or* snow tomorrow? Well, by the additivity property, we see that

$$P(\{\text{rain, snow}\}) = P(\{\text{rain}\}) + P(\{\text{snow}\}) = 0.40 + 0.15 = 0.55.$$

We thus conclude that, as expected, there is a 55% chance of rain *or* snow tomorrow. ∎

**EXAMPLE 1.2.2**

Suppose your candidate has a 60% chance of winning an election in progress. Then $S = \{$win, lose$\}$, with $P(\text{win}) = 0.6$ and $P(\text{lose}) = 0.4$. Note that $P(\text{win}) + P(\text{lose}) = 1$. ∎

**EXAMPLE 1.2.3**

Suppose we flip a fair coin, which can come up either heads $(H)$ or tails $(T)$ with equal probability. Then $S = \{H, T\}$, with $P(H) = P(T) = 0.5$. Of course, $P(H) + P(T) = 1$. ∎

**EXAMPLE 1.2.4**

Suppose we flip three fair coins in a row and keep track of the sequence of heads and tails that result. Then

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Furthermore, each of these eight outcomes is equally likely. Thus, $P(HHH) = 1/8$, $P(TTT) = 1/8$, etc. Also, the probability that the first coin is heads *and* the second coin is tails, but the third coin can be anything, is equal to the sum of the probabilities of the events $HTH$ and $HTT$, i.e., $P(HTH) + P(HTT) = 1/8 + 1/8 = 1/4$. ∎

**EXAMPLE 1.2.5**

Suppose we flip three fair coins in a row but care only about the number of heads that result. Then $S = \{0, 1, 2, 3\}$. However, the probabilities of these four outcomes are *not* all equally likely; we will see later that in fact $P(0) = P(3) = 1/8$, while $P(1) = P(2) = 3/8$. ∎

We note that it is possible to define probability models on more complicated (e.g., uncountably infinite) sample spaces as well.

**EXAMPLE 1.2.6**

Suppose that $S = [0, 1]$ is the unit interval. We can define a probability measure $P$ on $S$ by saying that

$$P([a, b]) = b - a, \qquad \text{whenever } 0 \leq a \leq b \leq 1. \tag{1.2.2}$$

In words, for any[1] subinterval $[a, b]$ of $[0, 1]$, the probability of the interval is simply the *length* of that interval. This example is called the *uniform distribution on* $[0, 1]$. The uniform distribution is just the first of many distributions on uncountable state spaces. Many further examples will be given in Chapter 2. ∎

## 1.2.1 | Venn Diagrams and Subsets

*Venn diagrams* provide a very useful graphical method for depicting the sample space $S$ and subsets of it. For example, in Figure 1.2.1 we have a Venn diagram showing the subset $A \subset S$ and the *complement*

$$A^c = \{s : s \notin A\}$$

of $A$. The rectangle denotes the entire sample space $S$. The circle (and its interior) denotes the subset $A$; the region outside the circle, but inside $S$, denotes $A^c$.
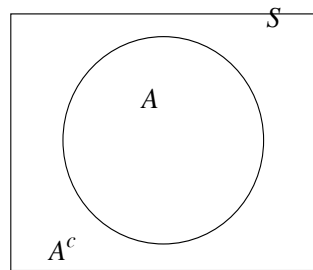


Figure 1.2.1: Venn diagram of the subsets $A$ and $A^c$ of the sample space $S$.

Two subsets $A \subset S$ and $B \subset S$ are depicted as two circles, as in Figure 1.2.2 on the next page. The *intersection*

$$A \cap B = \{s : s \in A \text{ and } s \in B\}$$

of the subsets $A$ and $B$ is the set of elements common to both sets and is depicted by the region where the two circles overlap. The set

$$A \cap B^c = \{s : s \in A \text{ and } s \in B\}$$

is called the *complement of B in A* and is depicted as the region inside the $A$ circle, but not inside the $B$ circle. This is the set of elements in $A$ but not in $B$. Similarly, we have the complement of $A$ in $B$, namely, $A^c \cap B$. Observe that the sets $A \cap B$, $A \cap B^c$, and $A^c \cap B$ are mutually disjoint.

---

[1]For the uniform distribution on $[0, 1]$, it turns out that not all subsets of $[0, 1]$ can properly be regarded as *events* for this model. However, this is merely a technical property, and any subset that we can explicitly write down will always be an event. See more advanced probability books, e.g., page 3 of *A First Look at Rigorous Probability Theory,* Second Edition, by J. S. Rosenthal (World Scientific Publishing, Singapore, 2006).

The *union*
$$A \cup B = \{s : s \in A \text{ or } s \in B\}$$
of the sets $A$ and $B$ is the set of elements that are in either $A$ or $B$. In Figure 1.2.2, it is depicted by the region covered by both circles. Notice that $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$.

There is one further region in Figure 1.2.2. This is the complement of $A \cup B$, namely, the set of elements that are in neither $A$ nor $B$. So we immediately have
$$(A \cup B)^c = A^c \cap B^c.$$
Similarly, we can show that
$$(A \cap B)^c = A^c \cup B^c,$$
namely, the subset of elements that are not in both $A$ and $B$ is given by the set of elements not in $A$ or not in $B$.
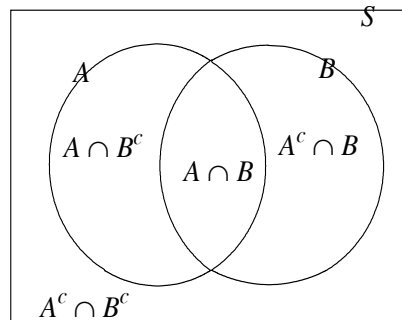


Figure 1.2.2: Venn diagram depicting the subsets $A$, $B$, $A \cap B$, $A \cap B^c$, $A^c \cap B$, $A^c \cap B^c$, and $A \cup B$.

Finally, we note that if $A$ and $B$ are disjoint subsets, then it makes sense to depict these as drawn in Figure 1.2.3, i.e., as two nonoverlapping circles because they have no elements in common.
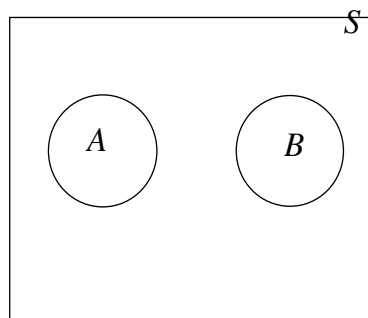


Figure 1.2.3: Venn diagram of the disjoint subsets $A$ and $B$.

## Summary of Section 1.2

- A probability model consists of a sample space $S$ and a probability measure $P$ assigning probabilities to each event.
- Different sorts of sets can arise as sample spaces.
- Venn diagrams provide a convenient method for representing sets and the relationships among them.

## EXERCISES

**1.2.1** Suppose $S = \{1, 2, 3\}$, with $P(\{1\}) = 1/2$, $P(\{2\}) = 1/3$, and $P(\{3\}) = 1/6$.
(a) What is $P(\{1, 2\})$?
(b) What is $P(\{1, 2, 3\})$?
(c) List all events $A$ such that $P(A) = 1/2$.

**1.2.2** Suppose $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$, with $P(\{s\}) = 1/8$ for $1 \leq s \leq 8$.
(a) What is $P(\{1, 2\})$?
(b) What is $P(\{1, 2, 3\})$?
(c) How many events $A$ are there such that $P(A) = 1/2$?

**1.2.3** Suppose $S = \{1, 2, 3\}$, with $P(\{1\}) = 1/2$ and $P(\{1, 2\}) = 2/3$. What must $P(\{2\})$ be?

**1.2.4** Suppose $S = \{1, 2, 3\}$, and we try to define $P$ by $P(\{1, 2, 3\}) = 1$, $P(\{1, 2\}) = 0.7$, $P(\{1, 3\}) = 0.5$, $P(\{2, 3\}) = 0.7$, $P(\{1\}) = 0.2$, $P(\{2\}) = 0.5$, $P(\{3\}) = 0.3$. Is $P$ a valid probability measure? Why or why not?

**1.2.5** Consider the uniform distribution on $[0, 1]$. Let $s \in [0, 1]$ be any outcome. What is $P(\{s\})$? Do you find this result surprising?

**1.2.6** Label the subregions in the Venn diagram in Figure 1.2.4 using the sets $A$, $B$, and $C$ and their complements (just as we did in Figure 1.2.2).
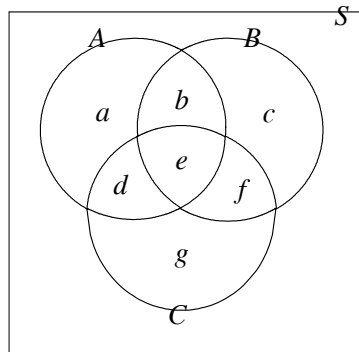


Figure 1.2.4: Venn diagram of subsets $A$, $B$, and $C$.

**1.2.7** On a Venn diagram, depict the set of elements that are in subsets $A$ or $B$ but *not* in both. Also write this as a subset involving unions and intersections of $A$, $B$, and their complements.

**1.2.8** Suppose $S = \{1, 2, 3\}$, and $P(\{1, 2\}) = 1/3$, and $P(\{2, 3\}) = 2/3$. Compute $P(\{1\})$, $P(\{2\})$, and $P(\{3\})$.

**1.2.9** Suppose $S = \{1, 2, 3, 4\}$, and $P(\{1\}) = 1/12$, and $P(\{1, 2\}) = 1/6$, and $P(\{1, 2, 3\}) = 1/3$. Compute $P(\{1\})$, $P(\{2\})$, $P(\{3\})$, and $P(\{4\})$.

**1.2.10** Suppose $S = \{1, 2, 3\}$, and $P(\{1\}) = P(\{3\}) = 2\,P(\{2\})$. Compute $P(\{1\})$, $P(\{2\})$, and $P(\{3\})$.

**1.2.11** Suppose $S = \{1, 2, 3\}$, and $P(\{1\}) = P(\{2\}) + 1/6$, and $P(\{3\}) = 2\,P(\{2\})$. Compute $P(\{1\})$, $P(\{2\})$, and $P(\{3\})$.

**1.2.12** Suppose $S = \{1, 2, 3, 4\}$, and $P(\{1\}) - 1/8 = P(\{2\}) = 3\,P(\{3\}) = 4\,P(\{4\})$. Compute $P(\{1\})$, $P(\{2\})$, $P(\{3\})$, and $P(\{4\})$.

### PROBLEMS

**1.2.13** Consider again the uniform distribution on [0, 1]. Is it true that

$$P([0, 1]) = \sum_{s \in [0,1]} P(\{s\})?$$

How does this relate to the additivity property of probability measures?

**1.2.14** Suppose $S$ is a finite or countable set. Is it possible that $P(\{s\}) = 0$ for every single $s \in S$? Why or why not?

**1.2.15** Suppose $S$ is an uncountable set. Is it possible that $P(\{s\}) = 0$ for every single $s \in S$? Why or why not?

### DISCUSSION TOPICS

**1.2.16** Does the additivity property make sense intuitively? Why or why not?

**1.2.17** Is it important that we always have $P(S) = 1$? How would probability theory change if this were not the case?

## 1.3 | Properties of Probability Models

The additivity property of probability measures automatically implies certain basic properties. These are true for *any* probability model at all.

If $A$ is any event, we write $A^c$ (read "$A$ complement") for the event that $A$ does *not* occur. In the weather example, if $A = \{rain\}$, then $A^c = \{snow, clear\}$. In the coin examples, if $A$ is the event that the first coin is heads, then $A^c$ is the event that the first coin is tails.

Now, $A$ and $A^c$ are always disjoint. Furthermore, their union is always the entire sample space: $A \cup A^c = S$. Hence, by the additivity property, we must have $P(A) + P(A^c) = P(S)$. But we always have $P(S) = 1$. Thus, $P(A) + P(A^c) = 1$, or

$$P(A^c) = 1 - P(A). \tag{1.3.1}$$

In words, the probability that any event does *not* occur is equal to one minus the probability that it *does* occur. This is a very helpful fact that we shall use often.

   Now suppose that $A_1, A_2, \ldots$ are events that form a *partition* of the sample space $S$. This means that $A_1, A_2, \ldots$ are disjoint and, furthermore, that their union is equal to $S$, i.e., $A_1 \cup A_2 \cup \cdots = S$. We have the following basic theorem that allows us to decompose the calculation of the probability of $B$ into the sum of the probabilities of the sets $A_i \cap B$. Often these are easier to compute.

---

**Theorem 1.3.1** (*Law of total probability, unconditioned version*) Let $A_1, A_2, \ldots$ be events that form a partition of the sample space $S$. Let $B$ be any event. Then

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

---

**PROOF**   The events $(A_1 \cap B)$, $(A_2 \cap B)$, ... are disjoint, and their union is $B$. Hence, the result follows immediately from the additivity property (1.2.1). ∎

A somewhat more useful version of the law of total probability, and applications of its use, are provided in Section 1.5.

   Suppose now that $A$ and $B$ are two events such that $A$ *contains* $B$ (in symbols, $A \supseteq B$). In words, all outcomes in $B$ are also in $A$. Intuitively, $A$ is a "larger" event than $B$, so we would expect its probability to be larger. We have the following result.

---

**Theorem 1.3.2** Let $A$ and $B$ be two events with $A \supseteq B$. Then

$$P(A) = P(B) + P(A \cap B^c). \qquad (1.3.2)$$

---

**PROOF**   We can write $A = B \cup (A \cap B^c)$, where $B$ and $A \cap B^c$ are disjoint. Hence, $P(A) = P(B) + P(A \cap B^c)$ by additivity. ∎

Because we always have $P(A \cap B^c) \geq 0$, we conclude the following.

---

**Corollary 1.3.1** (*Monotonicity*) Let $A$ and $B$ be two events, with $A \supseteq B$. Then

$$P(A) \geq P(B).$$

---

On the other hand, rearranging (1.3.2), we obtain the following.

---

**Corollary 1.3.2** Let $A$ and $B$ be two events, with $A \supseteq B$. Then

$$P(A \cap B^c) = P(A) - P(B). \qquad (1.3.3)$$

---

   More generally, even if we do not have $A \supseteq B$, we have the following property.

---

**Theorem 1.3.3** (*Principle of inclusion–exclusion, two-event version*) Let $A$ and $B$ be two events. Then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \qquad (1.3.4)$$

---

**PROOF**   We can write $A \cup B = (A \cap B^c) \cup (B \cap A^c) \cup (A \cap B)$, where $A \cap B^c$, $B \cap A^c$, and $A \cap B$ are disjoint. By additivity, we have

$$P(A \cup B) = P(A \cap B^c) + P(B \cap A^c) + P(A \cap B). \qquad (1.3.5)$$

On the other hand, using Corollary 1.3.2 (with $B$ replaced by $A \cap B$), we have

$$P(A \cap B^c) = P(A \cap (A \cap B)^c) = P(A) - P(A \cap B) \qquad (1.3.6)$$

and similarly,

$$P(B \cap A^c) = P(B) - P(A \cap B). \qquad (1.3.7)$$

Substituting (1.3.6) and (1.3.7) into (1.3.5), the result follows. ∎

A more general version of the principle of inclusion–exclusion is developed in Challenge 1.3.10.

Sometimes we do not need to evaluate the probability content of a union; we need only know it is bounded above by the sum of the probabilities of the individual events. This is called subadditivity.

---

**Theorem 1.3.4** (*Subadditivity*) Let $A_1, A_2, \ldots$ be a finite or countably infinite sequence of events, not necessarily disjoint. Then

$$P(A_1 \cup A_2 \cup \cdots) \leq P(A_1) + P(A_2) + \cdots .$$

---

**PROOF**   See Section 1.7 for the proof of this result. ∎

We note that some properties in the definition of a probability model actually follow from other properties. For example, once we know the probability $P$ is additive and that $P(S) = 1$, it follows that we *must* have $P(\emptyset) = 0$. Indeed, because $S$ and $\emptyset$ are disjoint, $P(S \cup \emptyset) = P(S) + P(\emptyset)$. But of course, $P(S \cup \emptyset) = P(S) = 1$, so we must have $P(\emptyset) = 0$.

Similarly, once we know $P$ is additive on countably infinite sequences of disjoint events, it follows that $P$ must be additive on finite sequences of disjoint events, too. Indeed, given a finite disjoint sequence $A_1, \ldots, A_n$, we can just set $A_i = \emptyset$ for all $i > n$, to get a countably infinite disjoint sequence with the same union and the same sum of probabilities.

## Summary of Section 1.3

- The probability of the complement of an event equals one minus the probability of the event.
- Probabilities always satisfy the basic properties of total probability, subadditivity, and monotonicity.
- The principle of inclusion–exclusion allows for the computation of $P(A \cup B)$ in terms of simpler events.

## EXERCISES

**1.3.1** Suppose $S = \{1, 2, \ldots, 100\}$. Suppose further that $P(\{1\}) = 0.1$.
(a) What is the probability $P(\{2, 3, 4, \ldots, 100\})$?
(b) What is the smallest possible value of $P(\{1, 2, 3\})$?

**1.3.2** Suppose that Al watches the six o'clock news 2/3 of the time, watches the eleven o'clock news 1/2 of the time, and watches both the six o'clock and eleven o'clock news 1/3 of the time. For a randomly selected day, what is the probability that Al watches only the six o'clock news? For a randomly selected day, what is the probability that Al watches neither news?

**1.3.3** Suppose that an employee arrives late 10% of the time, leaves early 20% of the time, and both arrives late *and* leaves early 5% of the time. What is the probability that on a given day that employee will either arrive late *or* leave early (or both)?

**1.3.4** Suppose your right knee is sore 15% of the time, and your left knee is sore 10% of the time. What is the largest possible percentage of time that at least one of your knees is sore? What is the smallest possible percentage of time that at least one of your knees is sore?

**1.3.5** Suppose a fair coin is flipped five times in a row.
(a) What is the probability of getting all five heads?
(b) What is the probability of getting at least one tail?

**1.3.6** Suppose a card is chosen uniformly at random from a standard 52-card deck.
(a) What is the probability that the card is a jack?
(b) What is the probability that the card is a club?
(c) What is the probability that the card is both a jack and a club?
(d) What is the probability that the card is either a jack or a club (or both)?

**1.3.7** Suppose your team has a 40% chance of winning or tying today's game and has a 30% chance of winning today's game. What is the probability that today's game will be a tie?

**1.3.8** Suppose 55% of students are female, of which 4/5 (44%) have long hair, and 45% are male, of which 1/3 (15% of all students) have long hair. What is the probability that a student chosen at random will either be female or have long hair (or both)?

## PROBLEMS

**1.3.9** Suppose we choose a positive integer at random, according to some unknown probability distribution. Suppose we know that $P(\{1, 2, 3, 4, 5\}) = 0.3$, that $P(\{4, 5, 6\}) = 0.4$, and that $P(\{1\}) = 0.1$. What are the largest and smallest possible values of $P(\{2\})$?

## CHALLENGES

**1.3.10** Generalize the principle of inclusion–exclusion, as follows.
(a) Suppose there are three events $A$, $B$, and $C$. Prove that

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

(b) Suppose there are $n$ events $A_1, A_2, \ldots, A_n$. Prove that

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{n} P(A_i \cap A_j) + \sum_{\substack{i,j,k=1 \\ i<j<k}}^{n} P(A_i \cap A_j \cap A_k)$$

$$- \cdots \pm P(A_1 \cap \cdots \cap A_n).$$

(Hint: Use induction.)

### DISCUSSION TOPICS

**1.3.11** Of the various theorems presented in this section, which ones do you think are the most important? Which ones do you think are the least important? Explain the reasons for your choices.

## 1.4 | Uniform Probability on Finite Spaces

If the sample space $S$ is finite, then one possible probability measure on $S$ is the *uniform* probability measure, which assigns probability $1/|S|$ to each outcome. Here $|S|$ is the number of elements in the sample space $S$. By additivity, it then follows that for any event $A$ we have

$$P(A) = \frac{|A|}{|S|}. \tag{1.4.1}$$

**EXAMPLE 1.4.1**
Suppose we roll a six-sided die. The possible outcomes are $S = \{1, 2, 3, 4, 5, 6\}$, so that $|S| = 6$. If the die is fair, then we believe each outcome is equally likely. We thus set $P(\{i\}) = 1/6$ for each $i \in S$ so that $P(\{3\}) = 1/6$, $P(\{4\}) = 1/6$, etc. It follows from (1.4.1) that, for example, $P(\{3, 4\}) = 2/6 = 1/3$, $P(\{1, 5, 6\}) = 3/6 = 1/2$, etc. This is a good model of rolling a fair six-sided die once. ∎

**EXAMPLE 1.4.2**
For a second example, suppose we flip a fair coin once. Then $S = \{\text{heads, tails}\}$, so that $|S| = 2$, and $P(\{\text{heads}\}) = P(\{\text{tails}\}) = 1/2$. ∎

**EXAMPLE 1.4.3**
Suppose now that we flip *three different* fair coins. The outcome can be written as a sequence of three letters, with each letter being $H$ (for heads) or $T$ (for tails). Thus,

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Here $|S| = 8$, and each of the events is equally likely. Hence, $P(\{HHH\}) = 1/8$, $P(\{HHH, TTT\}) = 2/8 = 1/4$, etc. Note also that, by additivity, we have, for example, that $P(\text{exactly two heads}) = P(\{HHT, HTH, THH\}) = 1/8 + 1/8 + 1/8 = 3/8$, etc. ∎

**EXAMPLE 1.4.4**

For a final example, suppose we roll a fair six-sided die *and* flip a fair coin. Then we can write

$$S = \{1H, 2H, 3H, 4H, 5H, 6H, 1T, 2T, 3T, 4T, 5T, 6T\}.$$

Hence, $|S| = 12$ in this case, and $P(s) = 1/12$ for each $s \in S$. ∎

## 1.4.1 Combinatorial Principles

Because of (1.4.1), problems involving uniform distributions on finite sample spaces often come down to being able to compute the sizes $|A|$ and $|S|$ of the sets involved. That is, we need to be good at *counting* the number of elements in various sets. The science of counting is called *combinatorics*, and some aspects of it are very sophisticated. In the remainder of this section, we consider a few simple combinatorial rules and their application in probability theory when the uniform distribution is appropriate.

**EXAMPLE 1.4.5** *Counting Sequences: The Multiplication Principle*

Suppose we flip three fair coins and roll two fair six-sided dice. What is the probability that all three coins come up heads and that both dice come up 6? Each coin has two possible outcomes (heads and tails), and each die has six possible outcomes $\{1, 2, 3, 4, 5, 6\}$. The total number of possible outcomes of the three coins and two dice is thus given by *multiplying* three 2's and two 6's, i.e., $2 \times 2 \times 2 \times 6 \times 6 = 288$. This is sometimes referred to as the *multiplication principle*. There are thus 288 possible outcomes of our experiment (e.g., $HHH66$, $HTH24$, $TTH15$, etc.). Of these outcomes, only one (namely, $HHH66$) counts as a success. Thus, the probability that all three coins come up heads and both dice come up 6 is equal to 1/288.

Notice that we can obtain this result in an alternative way. The chance that any one of the coins comes up heads is 1/2, and the chance that any one die comes up 6 is 1/6. Furthermore, these events are all *independent* (see the next section). Under independence, the probability that they *all* occur is given by the product of their individual probabilities, namely,

$$(1/2)(1/2)(1/2)(1/6)(1/6) = 1/288.$$

More generally, suppose we have $k$ finite sets $S_1, \ldots, S_k$ and we want to count the number of sequences of length $k$ where the $i$th element comes from $S_i$, i.e., count the number of elements in

$$S = \{(s_1, \ldots, s_k) : s_i \in S_i\} = S_1 \times \cdots \times S_k.$$

The multiplication principle says that the number of such sequences is obtained by multiplying together the number of elements in each set $S_i$, i.e.,

$$|S| = |S_1| \cdots |S_k|. \; ∎$$

**EXAMPLE 1.4.6**

Suppose we roll two fair six-sided dice. What is the probability that the sum of the numbers showing is equal to 10? By the above multiplication principle, the total number of possible outcomes is equal to $6 \times 6 = 36$. Of these outcomes, there are three that sum to 10, namely, $(4, 6)$, $(5, 5)$, and $(6, 4)$. Thus, the probability that the sum is 10 is equal to 3/36, or 1/12. ∎

**EXAMPLE 1.4.7** *Counting Permutations*

Suppose four friends go to a restaurant, and each checks his or her coat. At the end of the meal, the four coats are *randomly* returned to the four people. What is the probability that each of the four people gets his or her own coat? Here the total number of different ways the coats can be returned is equal to $4 \times 3 \times 2 \times 1$, or 4! (i.e., four factorial). This is because the first coat can be returned to any of the four friends, the second coat to any of the three remaining friends, and so on. Only one of these assignments is correct. Hence, the probability that each of the four people gets his or her own coat is equal to 1/4!, or 1/24.

Here we are counting permutations, or sequences of elements from a set where no element appears more than once. We can use the multiplication principle to count permutations more generally. For example, suppose $|S| = n$ and we want to count the number of permutations of length $k \leq n$ obtained from $S$, i.e., we want to count the number of elements of the set

$$\left\{ (s_1, \ldots, s_k) : s_i \in S, s_i \neq s_j \text{ when } i \neq j \right\}.$$

Then we have $n$ choices for the first element $s_1$, $n - 1$ choices for the second element, and finally $n - (k - 1) = n - k + 1$ choices for the last element. So there are $n(n - 1) \cdots (n - k + 1)$ permutations of length $k$ from a set of $n$ elements. This can also be written as $n!/(n - k)!$. Notice that when $k = n$, there are

$$n! = n(n - 1) \cdots 2 \cdot 1$$

permutations of length $n$. ∎

**EXAMPLE 1.4.8** *Counting Subsets*

Suppose 10 fair coins are flipped. What is the probability that exactly seven of them are heads? Here each possible sequence of 10 heads or tails (e.g., $HHHTTTHTTT$, $THTTTTHHHT$, etc.) is equally likely, and by the multiplication principle the total number of possible outcomes is equal to 2 multiplied by itself 10 times, or $2^{10} = 1024$. Hence, the probability of any particular sequence occurring is 1/1024. But of these sequences, how many have exactly seven heads?

To answer this, notice that we may specify such a sequence by giving the positions of the seven heads, which involves choosing a subset of size 7 from the set of possible indices $\{1, \ldots, 10\}$. There are $10!/3! = 10 \cdot 9 \cdots 5 \cdot 4$ different permutations of length 7 from $\{1, \ldots, 10\}$, and each such permutation specifies a sequence of seven heads and three tails. But we can permute the indices specifying where the heads go in 7! different ways without changing the sequence of heads and tails. So the total number of outcomes with exactly seven heads is equal to $10!/3!7! = 120$. The probability that exactly seven of the ten coins are heads is therefore equal to 120/1024, or just under 12%.

In general, if we have a set $S$ of $n$ elements, then the number of different subsets of size $k$ that we can construct by choosing elements from $S$ is

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

which is called the *binomial coefficient*. This follows by the same argument, namely, there are $n!/(n-k)!$ permutations of length $k$ obtained from the set; each such permutation, and the $k!$ permutations obtained by permuting it, specify a unique subset of $S$.
∎

It follows, for example, that the probability of obtaining exactly $k$ heads when flipping a total of $n$ fair coins is given by

$$\binom{n}{k} 2^{-n} = \frac{n!}{k!\,(n-k)!} 2^{-n}.$$

This is because there are $\binom{n}{k}$ different patterns of $k$ heads and $n-k$ tails, and a total of $2^n$ different sequences of $n$ heads and tails.

More generally, if each coin has probability $\theta$ of being heads (and probability $1-\theta$ of being tails), where $0 \le \theta \le 1$, then the probability of obtaining exactly $k$ heads when flipping a total of $n$ such coins is given by

$$\binom{n}{k} \theta^k (1-\theta)^{n-k} = \frac{n!}{k!\,(n-k)!} \theta^k (1-\theta)^{n-k}, \tag{1.4.2}$$

because each of the $\binom{n}{k}$ different patterns of $k$ heads and $n-k$ tails has probability $\theta^k (1-\theta)^{n-k}$ of occurring (this follows from the discussion of independence in Section 1.5.2). If $\theta = 1/2$, then this reduces to the previous formula.

**EXAMPLE 1.4.9** *Counting Sequences of Subsets and Partitions*
Suppose we have a set $S$ of $n$ elements and we want to count the number of elements of

$$\left\{ (S_1, S_2, \ldots, S_l) : S_i \subset S, |S_i| = k_i, S_i \cap S_j = \emptyset \text{ when } i \ne j \right\},$$

namely, we want to count the number of sequences of $l$ subsets of a set where no two subsets have any elements in common and the $i$th subset has $k_i$ elements. By the multiplication principle, this equals

$$\binom{n}{k_1}\binom{n-k_1}{k_2} \cdots \binom{n - k_1 - \cdots - k_{l-1}}{k_l}$$
$$= \frac{n!}{k_1! \cdots k_{l-1}! k_l! \,(n - k_1 - \cdots - k_l)!}, \tag{1.4.3}$$

because we can choose the elements of $S_1$ in $\binom{n}{k_1}$ ways, choose the elements of $S_2$ in $\binom{n-k_1}{k_2}$ ways, etc.

When we have that $S = S_1 \cup S_2 \cup \cdots \cup S_l$, in addition to the individual sets being mutually disjoint, then we are counting the number of *ordered partitions* of a set of $n$

elements with $k_1$ elements in the first set, $k_2$ elements in the second set, etc. In this case, (1.4.3) equals

$$\binom{n}{k_1 \; k_2 \; \ldots \; k_l} = \frac{n!}{k_1! k_2! \cdots k_l!}, \tag{1.4.4}$$

which is called the *multinomial coefficient*. ∎

For example, how many different bridge hands are there? By this we mean how many different ways can a deck of 52 cards be divided up into four hands of 13 cards each, with the hands labelled North, East, South, and West, respectively. By (1.4.4), this equals

$$\binom{52}{13 \; 13 \; 13 \; 13} = \frac{52!}{13! \, 13! \, 13! \, 13!} \approx 5.364474 \times 10^{28},$$

which is a very large number.

## Summary of Section 1.4

- The uniform probability distribution on a finite sample space $S$ satisfies $P(A) = |A| \, / \, |S|$.
- Computing $P(A)$ in this case requires computing the *sizes* of the sets $A$ and $S$. This may require combinatorial principles such as the multiplication principle, factorials, and binomial/multinomial coefficients.

## EXERCISES

**1.4.1** Suppose we roll eight fair six-sided dice.
(a) What is the probability that all eight dice show a 6?
(b) What is the probability that all eight dice show the same number?
(c) What is the probability that the sum of the eight dice is equal to 9?

**1.4.2** Suppose we roll 10 fair six-sided dice. What is the probability that there are exactly two 2's showing?

**1.4.3** Suppose we flip 100 fair independent coins. What is the probability that at least three of them are heads? (Hint: You may wish to use (1.3.1).)

**1.4.4** Suppose we are dealt five cards from an ordinary 52-card deck. What is the probability that
(a) we get all four aces, plus the king of spades?
(b) all five cards are spades?
(c) we get no pairs (i.e., all five cards are different values)?
(d) we get a full house (i.e., three cards of a kind, plus a different pair)?

**1.4.5** Suppose we deal four 13-card bridge hands from an ordinary 52-card deck. What is the probability that
(a) all 13 spades end up in the same hand?
(b) all four aces end up in the same hand?

**1.4.6** Suppose we pick two cards at random from an ordinary 52-card deck. What is the probability that the sum of the values of the two cards (where we count jacks, queens, and kings as 10, and count aces as 1) is at least 4?

**1.4.7** Suppose we keep dealing cards from an ordinary 52-card deck until the first jack appears. What is the probability that at least 10 cards go by *before* the first jack?

**1.4.8** In a well-shuffled ordinary 52-card deck, what is the probability that the ace of spades and the ace of clubs are adjacent to each other?

**1.4.9** Suppose we repeatedly roll two fair six-sided dice, considering the *sum* of the two values showing each time. What is the probability that the first time the sum is exactly 7 is on the third roll?

**1.4.10** Suppose we roll three fair six-sided dice. What is the probability that two of them show the same value, but the third one does not?

**1.4.11** Consider two urns, labelled urn #1 and urn #2. Suppose urn #1 has 5 red and 7 blue balls. Suppose urn #2 has 6 red and 12 blue balls. Suppose we pick three balls uniformly at random from each of the two urns. What is the probability that all six chosen balls are the same color?

**1.4.12** Suppose we roll a fair six-sided die and flip three fair coins. What is the probability that the total number of heads is equal to the number showing on the die?

**1.4.13** Suppose we flip two pennies, three nickels, and four dimes. What is the probability that the total value of all coins showing heads is equal to $0.31?

## PROBLEMS

**1.4.14** Show that a probability measure defined by (1.4.1) is always additive in the sense of (1.2.1).

**1.4.15** Suppose we roll eight fair six-sided dice. What is the probability that the sum of the eight dice is equal to 9? What is the probability that the sum of the eight dice is equal to 10? What is the probability that the sum of the eight dice is equal to 11?

**1.4.16** Suppose we roll one fair six-sided die, and flip six coins. What is the probability that the number of heads is equal to the number showing on the die?

**1.4.17** Suppose we roll 10 fair six-sided dice. What is the probability that there are exactly two 2's showing *and* exactly three 3's showing?

**1.4.18** Suppose we deal four 13-card bridge hands from an ordinary 52-card deck. What is the probability that the North and East hands each have exactly the same number of spades?

**1.4.19** Suppose we pick a card at random from an ordinary 52-card deck and also flip 10 fair coins. What is the probability that the number of heads equals the value of the card (where we count jacks, queens, and kings as 10, and count aces as 1)?

## CHALLENGES

**1.4.20** Suppose we roll two fair six-sided dice and flip 12 coins. What is the probability that the number of heads is equal to the sum of the numbers showing on the two dice?

**1.4.21** *(The birthday problem)* Suppose there are $C$ people, each of whose birthdays (month and day only) are equally likely to fall on any of the 365 days of a normal (i.e., non-leap) year.
(a) Suppose $C = 2$. What is the probability that the two people have the same exact

birthday?

(b) Suppose $C \geq 2$. What is the probability that all $C$ people have the same exact birthday?

(c) Suppose $C \geq 2$. What is the probability that *some pair* of the $C$ people have the same exact birthday? (Hint: You may wish to use (1.3.1).)

(d) What is the smallest value of $C$ such that the probability in part (c) is more than 0.5? Do you find this result surprising?

# 1.5 | Conditional Probability and Independence

Consider again the three-coin example as in Example 1.4.3, where we flip three different fair coins, and

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

with $P(s) = 1/8$ for each $s \in S$. What is the probability that the first coin comes up heads? Well, of course, this should be $1/2$. We can see this more formally by saying that $P(\text{first coin heads}) = P(\{HHH, HHT, HTH, HTT\}) = 4/8 = 1/2$, as it should.

But suppose now that an informant tells us that exactly two of the three coins came up heads. *Now* what is the probability that the first coin was heads?

The point is that this informant has changed our available information, i.e., changed our level of ignorance. It follows that our corresponding probabilities should also change. Indeed, if we know that exactly two of the coins were heads, then we know that the outcome was one of $HHT$, $HTH$, and $THH$. Because those three outcomes should (in this case) still all be equally likely, and because only the first two correspond to the first coin being heads, we conclude the following: If we know that exactly two of the three coins are heads, *then* the probability that the first coin is heads is $2/3$.

More precisely, we have computed a *conditional probability*. That is, we have determined that, *conditional* on knowing that exactly two coins came up heads, the *conditional probability* of the first coin being heads is $2/3$. We write this in mathematical notation as

$$P(\text{first coin heads} \mid \text{two coins heads}) = 2/3.$$

Here the vertical bar | stands for "conditional on," or "given that."

## 1.5.1 | Conditional Probability

In general, given two events $A$ and $B$ with $P(B) > 0$, the *conditional probability* of $A$ given $B$, written $P(A \mid B)$, stands for the fraction of the time that $A$ occurs once we *know* that $B$ occurs. It is computed as the ratio of the probability that $A$ and $B$ *both* occur, divided by the probability that $B$ occurs, as follows.

**Definition 1.5.1** Given two events $A$ and $B$, with $P(B) > 0$, the conditional probability of $A$ given $B$ is equal to

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \qquad (1.5.1)$$

The motivation for (1.5.1) is as follows. The event $B$ will occur a fraction $P(B)$ of the time. Also, *both* $A$ and $B$ will occur a fraction $P(A \cap B)$ of the time. The ratio $P(A \cap B)/P(B)$ thus gives the *proportion* of the times when $B$ occurs, that $A$ *also* occurs. That is, if we *ignore* all the times that $B$ does not occur and consider only those times that $B$ does occur, then the ratio $P(A \cap B)/P(B)$ equals the fraction of the time that $A$ will also occur. This is precisely what is meant by the conditional probability of $A$ given $B$.

In the example just computed, $A$ is the event that the first coin is heads, while $B$ is the event that exactly two coins were heads. Hence, in mathematical terms, $A = \{HHH, HHT, HTH, HTT\}$ and $B = \{HHT, HTH, THH\}$. It follows that $A \cap B = \{HHT, HTH\}$. Therefore,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{HHT, HTH\})}{P(\{HHT, HTH, THH\})} = \frac{2/8}{3/8} = 2/3,$$

as already computed.

On the other hand, we similarly compute that

$$P(\text{first coin tails} \mid \text{two coins heads}) = 1/3.$$

We thus see that conditioning on some event (such as "two coins heads") can make probabilities either increase (as for the event "first coin heads") or decrease (as for the event "first coin tails").

The definition of $P(B \mid A)$ immediately leads to the *multiplication formula*

$$P(A \cap B) = P(A)P(B \mid A). \qquad (1.5.2)$$

This allows us to compute the joint probability of $A$ and $B$ when we are given the probability of $A$ and the conditional probability of $B$ given $A$.

Conditional probability allows us to express Theorem 1.3.1, the law of total probability, in a different and sometimes more helpful way.

**Theorem 1.5.1** (*Law of total probability, conditioned version*) Let $A_1, A_2, \ldots$ be events that form a partition of the sample space $S$, each of positive probability. Let $B$ be any event. Then $P(B) = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots$.

**PROOF**    The multiplication formula (1.5.2) gives that $P(A_i \cap B) = P(A_i)P(A_i \mid B)$. The result then follows immediately from Theorem 1.3.1. ∎

**EXAMPLE 1.5.1**
Suppose a class contains 60% girls and 40% boys. Suppose that 30% of the girls have long hair, and 20% of the boys have long hair. A student is chosen uniformly at random from the class. What is the probability that the chosen student will have long hair?

To answer this, we let $A_1$ be the set of girls and $A_2$ be the set of boys. Then $\{A_1, A_2\}$ is a partition of the class. We further let $B$ be the set of all students with long hair.

We are interested in $P(B)$. We compute this by Theorem 1.5.1 as

$$P(B) = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) = (0.6)(0.3) + (0.4)(0.2) = 0.26,$$

so there is a 26% chance that the randomly chosen student has long hair. ∎

Suppose now that $A$ and $B$ are two events, each of positive probability. In some applications, we are given the values of $P(A)$, $P(B)$, and $P(B \mid A)$ and want to compute $P(A \mid B)$. The following result establishes a simple relationship among these quantities.

---

**Theorem 1.5.2** (*Bayes' theorem*) Let $A$ and $B$ be two events, each of positive probability. Then
$$P(A \mid B) = \frac{P(A)}{P(B)} P(B \mid A).$$

---

**PROOF**   We compute that

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

This gives the result. ∎

Standard applications of the multiplication formula, the law of total probabilities, and Bayes' theorem occur with *two-stage systems*. The response for such systems can be thought of as occurring in two steps or stages. Typically, we are given the probabilities for the first stage and the conditional probabilities for the second stage. The multiplication formula is then used to calculate joint probabilities for what happens at both stages; the law of total probability is used to compute the probabilities for what happens at the second stage; and Bayes' theorem is used to calculate the conditional probabilities for the first stage, given what has occurred at the second stage. We illustrate this by an example.

**EXAMPLE 1.5.2**
Suppose urn #1 has 3 red and 2 blue balls, and urn #2 has 4 red and 7 blue balls. Suppose one of the two urns is selected with probability $1/2$ each, and then one of the balls within that urn is picked uniformly at random.

What is the probability that urn #2 is selected at the first stage (event $A$) and a blue ball is selected at the second stage (event $B$)? The multiplication formula provides the correct way to compute this probability as

$$P(A \cap B) = P(A)P(B \mid A) = \frac{1}{2}\frac{7}{11} = \frac{7}{22}.$$

Suppose instead we want to compute the probability that a blue ball is obtained. Using the law of total probability (Theorem 1.5.1), we have that

$$P(B) = P(A)P(B \mid A) + P(A^c)P(B \mid A^c) = \frac{1}{2}\frac{2}{5} + \frac{1}{2}\frac{7}{11}.$$

Now suppose we are given the information that the ball picked is blue. Then, using Bayes' theorem, the conditional probability that we had selected urn #2 is given by

$$
\begin{aligned}
P(A \mid B) &= \frac{P(A)}{P(B)} P(B \mid A) = \left( \frac{1/2}{(1/2)(2/5) + (1/2)(7/11)} \right) \frac{7}{11} \\
&= 35/57 = 0.614.
\end{aligned}
$$

Note that, without the information that a blue ball occurred at the second stage, we have that

$$P(\text{urn \#2 selected}) = 1/2.$$

We see that knowing the ball was blue significantly *increases* the probability that urn #2 was selected. ∎

We can represent a two-stage system using a tree, as in Figure 1.5.1. It can be helpful to draw such a figure when carrying out probability computations for such systems. There are two possible outcomes at the first stage and three possible outcomes at the second stage.
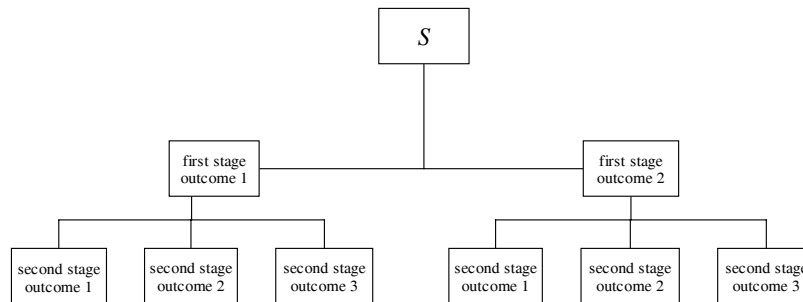


Figure 1.5.1: A tree depicting a two-stage system with two possible outcomes at the first stage and three possible outcomes at the second stage.

## 1.5.2 | Independence of Events

Consider now Example 1.4.4, where we roll one fair die and flip one fair coin, so that

$$S = \{1H, 2H, 3H, 4H, 5H, 6H, 1T, 2T, 3T, 4T, 5T, 6T\}$$

and $P(\{s\}) = 1/12$ for each $s \in S$. Here the probability that the die comes up 5 is equal to $P(\{5H, 5T\}) = 2/12 = 1/6$, as it should be.

But now, what is the probability that the die comes up 5, *conditional* on knowing that the coin came up tails? Well, we can compute that probability as

$$
\begin{aligned}
P(\text{die} = 5 \mid \text{coin} = \text{tails}) \quad &= \quad \frac{P(\text{die} = 5 \text{ and coin} = \text{tails})}{P(\text{coin} = \text{tails})} \\
&= \quad \frac{P(\{5T\})}{P(\{1T, 2T, 3T, 4T, 5T, 6T\})} \\
&= \quad \frac{1/12}{6/12} = 1/6.
\end{aligned}
$$

This is the same as the unconditional probability, $P(\text{die} = 5)$. It seems that knowing that the coin was tails had no effect whatsoever on the probability that the coin came up 5. This property is called *independence*. We say that the coin and the die are *independent* in this example, to indicate that the occurrence of one does not have any influence on the probability of the other occurring.

More formally, we make the following definition.

---

**Definition 1.5.2**   Two events $A$ and $B$ are *independent* if

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

---

Now, because $P(A \mid B) = P(A \cap B)/P(B)$, we see that $A$ and $B$ are independent if and only if $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$, provided that $P(A) > 0$ and $P(B) > 0$. Definition 1.5.2 has the advantage that it remains valid even if $P(B) = 0$ or $P(A) = 0$, respectively. Intuitively, events $A$ and $B$ are independent if neither one has any impact on the probability of the other.

### EXAMPLE 1.5.3
In Example 1.4.4, if $A$ is the event that the die was 5, and $B$ is the event that the coin was tails, then $P(A) = P(\{5H, 5T\}) = 2/12 = 1/6$, and

$$P(B) = P(\{1T, 2T, 3T, 4T, 5T, 6T\}) = 6/12 = 1/2.$$

Also, $P(A \cap B) = P(\{5T\}) = 1/12$, which is indeed equal to $(1/6)(1/2)$. Hence, $A$ and $B$ are independent in this case. ∎

For multiple events, the definition of independence is somewhat more involved.

---

**Definition 1.5.3**   A collection of events $A_1, A_2, A_3, \ldots$ are *independent* if

$$P(A_{i_1} \cap \cdots \cap A_{i_j}) = P(A_{i_1}) \cdots P(A_{i_j})$$

for *any* finite subcollection $A_{i_1}, \ldots, A_{i_j}$ of distinct events.

---

**EXAMPLE 1.5.4**
According to Definition 1.5.3, three events $A$, $B$, and $C$ are independent if *all* of the
following equations hold:

$$\begin{aligned} P(A \cap B) &= P(A)P(B), \\ P(A \cap C) &= P(A)P(C), \\ P(B \cap C) &= P(B)P(C), \end{aligned} \qquad (1.5.3)$$

and

$$P(A \cap B \cap C) = P(A)P(B)P(C). \qquad (1.5.4)$$

It is not sufficient to check just *some* of these conditions to verify independence. For
example, suppose that $S = \{1, 2, 3, 4\}$, with $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) =$
$1/4$. Let $A = \{1, 2\}$, $B = \{1, 3\}$, and $C = \{1, 4\}$. Then each of the three equations
(1.5.3) holds, but equation (1.5.4) does not hold. Here, the events $A$, $B$, and $C$ are
called *pairwise independent*, but they are not independent. ∎

## Summary of Section 1.5

- Conditional probability measures the probability that $A$ occurs given that $B$ occurs; it is given by $P(A \mid B) = P(A \cap B) / P(B)$.
- Conditional probability satisfies its own law of total probability.
- Events are independent if they have no effect on each other's probabilities. Formally, this means that $P(A \cap B) = P(A)P(B)$.
- If $A$ and $B$ are independent, and $P(A) > 0$ and $P(B) > 0$, then $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.

## EXERCISES

**1.5.1** Suppose that we roll four fair six-sided dice.
(a) What is the conditional probability that the first die shows 2, conditional on the
event that exactly three dice show 2?
(b) What is the conditional probability that the first die shows 2, conditional on the
event that *at least* three dice show 2?

**1.5.2** Suppose we flip two fair coins and roll one fair six-sided die.
(a) What is the probability that the number of heads equals the number showing on the
die?
(b) What is the conditional probability that the number of heads equals the number
showing on the die, conditional on knowing that the die showed 1?
(c) Is the answer for part (b) larger or smaller than the answer for part (a)? Explain
intuitively why this is so.

**1.5.3** Suppose we flip three fair coins.
(a) What is the probability that all three coins are heads?
(b) What is the conditional probability that all three coins are heads, conditional on
knowing that the number of heads is odd?

(c) What is the conditional probability that all three coins are heads, given that the number of heads is even?

**1.5.4** Suppose we deal five cards from an ordinary 52-card deck. What is the conditional probability that all five cards are spades, given that at least four of them are spades?

**1.5.5** Suppose we deal five cards from an ordinary 52-card deck. What is the conditional probability that the hand contains all four aces, given that the hand contains at least four aces?

**1.5.6** Suppose we deal five cards from an ordinary 52-card deck. What is the conditional probability that the hand contains no pairs, given that it contains no spades?

**1.5.7** Suppose a baseball pitcher throws fastballs 80% of the time and curveballs 20% of the time. Suppose a batter hits a home run on 8% of all fastball pitches, and on 5% of all curveball pitches. What is the probability that this batter will hit a home run on this pitcher's next pitch?

**1.5.8** Suppose the probability of snow is 20%, and the probability of a traffic accident is 10%. Suppose further that the *conditional* probability of an accident, given that it snows, is 40%. What is the conditional probability that it snows, given that there is an accident?

**1.5.9** Suppose we roll two fair six-sided dice, one red and one blue. Let $A$ be the event that the two dice show the same value. Let $B$ be the event that the sum of the two dice is equal to 12. Let $C$ be the event that the red die shows 4. Let $D$ be the event that the blue die shows 4.
(a) Are $A$ and $B$ independent?
(b) Are $A$ and $C$ independent?
(c) Are $A$ and $D$ independent?
(d) Are $C$ and $D$ independent?
(e) Are $A$, $C$, and $D$ all independent?

**1.5.10** Consider two urns, labelled urn #1 and urn #2. Suppose, as in Exercise 1.4.11, that urn #1 has 5 red and 7 blue balls, that urn #2 has 6 red and 12 blue balls, and that we pick three balls uniformly at random from each of the two urns. Conditional on the fact that all six chosen balls are the same color, what is the conditional probability that this color is red?

**1.5.11** Suppose we roll a fair six-sided die and then flip a number of fair coins equal to the number showing on the die. (For example, if the die shows 4, then we flip 4 coins.)
(a) What is the probability that the number of heads equals 3?
(b) Conditional on knowing that the number of heads equals 3, what is the conditional probability that the die showed the number 5?

**1.5.12** Suppose we roll a fair six-sided die and then pick a number of cards from a well-shuffled deck equal to the number showing on the die. (For example, if the die shows 4, then we pick 4 cards.)
(a) What is the probability that the number of jacks in our hand equals 2?
(b) Conditional on knowing that the number of jacks in our hand equals 2, what is the conditional probability that the die showed the number 3?

## PROBLEMS

**1.5.13** Consider three cards, as follows: One is red on both sides, one is black on both sides, and one is red on one side and black on the other. Suppose the cards are placed in a hat, and one is chosen at random. Suppose further that this card is placed flat on the table, so we can see one side only.
(a) What is the probability that this one side is red?
(b) Conditional on this one side being red, what is the probability that the card showing is the one that is red on both sides? (Hint: The answer is somewhat surprising.)
(c) Suppose you wanted to verify the answer in part (b), using an actual, physical experiment. Explain how you could do this.

**1.5.14** Prove that $A$ and $B$ are independent if and only if $A^C$ and $B$ are independent.

**1.5.15** Let $A$ and $B$ be events of positive probability. Prove that $P(A \mid B) > P(A)$ if and only if $P(B \mid A) > P(B)$.

## CHALLENGES

**1.5.16** Suppose we roll three fair six-sided dice. Compute the conditional probability that the first die shows 4, given that the sum of the three numbers showing is 12.

**1.5.17** *(The game of craps)* The game of craps is played by rolling two fair, six-sided dice. On the first roll, if the sum of the two numbers showing equals 2, 3, or 12, then the player immediately loses. If the sum equals 7 or 11, then the player immediately wins. If the sum equals any other value, then this value becomes the player's "point." The player then repeatedly rolls the two dice, until such time as he or she either rolls the point value again (in which case he or she wins) or rolls a 7 (in which case he or she loses).
(a) Suppose the player's point is equal to 4. Conditional on this, what is the conditional probability that he or she will win (i.e., will roll another 4 before rolling a 7)? (Hint: The final roll will be either a 4 or 7; what is the conditional probability that it is a 4?)
(b) For $2 \leq i \leq 12$, let $p_i$ be the conditional probability that the player will win, conditional on having rolled $i$ on the first roll. Compute $p_i$ for all $i$ with $2 \leq i \leq 12$. (Hint: You've already done this for $i = 4$ in part (b). Also, the cases $i = 2, 3, 7, 11, 12$ are trivial. The other cases are similar to the $i = 4$ case.)
(c) Compute the overall probability that a player will win at craps. (Hint: Use part (b) and Theorem 1.5.1.)

**1.5.18** *(The Monty Hall problem)* Suppose there are three doors, labeled A, B, and C. A new car is behind one of the three doors, but you don't know which. You select one of the doors, say, door A. The host then opens one of doors B or C, as follows: If the car is behind B, then they open C; if the car is behind C, then they open B; if the car is behind A, then they open either B or C with probability 1/2 each. (In any case, the door opened by the host will *not* have the car behind it.) The host then gives you the *option* of either sticking with your original door choice (i.e., A), or switching to the remaining unopened door (i.e., whichever of B or C the host did not open). You then win (i.e., get to keep the car) if and only if the car is behind your final door selection. (Source: *Parade Magazine*, "Ask Marilyn" column, September 9, 1990.) Suppose for definiteness that the host opens door B.

(a) If you stick with your original choice (i.e., door A), conditional on the host having opened door B, then what is your probability of winning? (Hint: First condition on the true location of the car. Then use Theorem 1.5.2.)

(b) If you switch to the remaining door (i.e., door C), conditional on the host having opened door B, then what is your probability of winning?

(c) Do you find the result of parts (a) and (b) surprising? How could you design a physical experiment to verify the result?

(d) Suppose we change the rules so that, if you originally chose A and the car was indeed behind A, then the host always opens door B. How would the answers to parts (a) and (b) change in this case?

(e) Suppose we change the rules so that, if you originally chose A, then the host always opens door B *no matter where the car is*. We then *condition* on the fact that door B happened *not* to have a car behind it. How would the answers to parts (a) and (b) change in this case?

## DISCUSSION TOPICS

**1.5.19** Suppose two people each flip a fair coin simultaneously. Will the results of the two flips usually be independent? Under what sorts of circumstances might they not be independent? (List as many such circumstances as you can.)

**1.5.20** Suppose you are able to repeat an experiment many times, and you wish to check whether or not two events are independent. How might you go about this?

**1.5.21** The Monty Hall problem (Challenge 1.5.18) was originally presented by Marilyn von Savant, writing in the "Ask Marilyn" column of *Parade Magazine*. She gave the correct answer. However, many people (including some well-known mathematicians, plus many laypeople) wrote in to complain that her answer was incorrect. The controversy dragged on for months, with many letters and very strong language written by both sides (in the end, von Savant was vindicated). Part of the confusion lay in the assumptions being made, e.g., some people misinterpreted her question as that of the modified version of part (e) of Challenge 1.5.18. However, a lot of the confusion was simply due to mathematical errors and misunderstandings. (Source: *Parade Magazine*, "Ask Marilyn" column, September 9, 1990; December 2, 1990; February 17, 1991; July 7, 1991.)

(a) Does it surprise you that so many people, including well-known mathematicians, made errors in solving this problem? Why or why not?

(b) Does it surprise you that so many people, including many laypeople, cared so strongly about the answer to this problem? Why or why not?

# 1.6 | Continuity of $P$

Suppose $A_1, A_2, \ldots$ is a sequence of events that are getting "closer" (in some sense) to another event, $A$. Then we might expect that the probabilities $P(A_1), P(A_2), \ldots$ are getting close to $P(A)$, i.e., that $\lim_{n \to \infty} P(A_n) = P(A)$. But can we be sure about this?

Properties like this, which say that $P(A_n)$ is close to $P(A)$ whenever $A_n$ is "close" to $A$, are called *continuity properties*. The above question can thus be translated, roughly, as asking whether or not probability measures $P$ are "continuous." It turns out that $P$ is indeed continuous in some sense.

Specifically, let us write $\{A_n\} \nearrow A$ and say that the sequence $\{A_n\}$ *increases* to $A$, if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, and also $\bigcup_{n=1}^{\infty} A_n = A$. That is, the sequence of events is an *increasing* sequence, and furthermore its union is equal to $A$. For example, if $A_n = (1/n, n]$, then $A_1 \subseteq A_2 \subseteq \cdots$ and $\bigcup_{n=1}^{\infty} A_n = (0, \infty)$. Hence, $\{(1/n, n]\} \nearrow (0, \infty)$. Figure 1.6.1 depicts an increasing sequence of subsets.
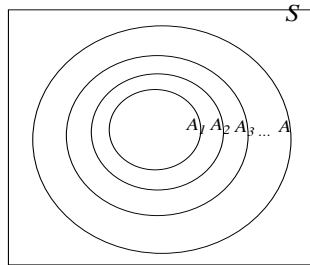


Figure 1.6.1: An increasing sequence of subsets $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$.

Similarly, let us write $\{A_n\} \searrow A$ and say that the sequence $\{A_n\}$ *decreases* to $A$, if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, and also $\bigcap_{n=1}^{\infty} A_n = A$. That is, the sequence of events is a *decreasing* sequence, and furthermore its intersection is equal to $A$. For example, if $A_n = (-1/n, 1/n]$, then $A_1 \supseteq A_2 \supseteq \cdots$ and $\bigcap_{n=1}^{\infty} A_n = \{0\}$. Hence, $\{(-1/n, 1/n]\} \searrow \{0\}$. Figure 1.6.2 depicts a decreasing sequence of subsets.
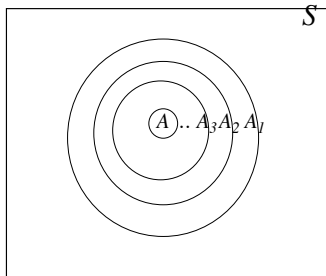


Figure 1.6.2: A decreasing sequence of subsets $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$.

We will consider such sequences of sets at several points in the text. For this we need the following result.

**Theorem 1.6.1** Let $A, A_1, A_2, \dots$ be events, and suppose that either $\{A_n\} \nearrow A$ or $\{A_n\} \searrow A$. Then

$$\lim_{n \to \infty} P(A_n) = P(A).$$

**PROOF**   See Section 1.7 for the proof of this theorem. ∎

### EXAMPLE 1.6.1

Suppose $S$ is the set of all positive integers, with $P(s) = 2^{-s}$ for all $s \in S$. Then what is $P(\{5, 6, 7, 8, \ldots\})$?

We begin by noting that the events $A_n = \{5, 6, 7, 8, \ldots, n\}$ increase to $A = \{5, 6, 7, 8, \ldots\}$, i.e., $\{A_n\} \nearrow A$. Hence, using continuity of probabilities, we must have

$$
\begin{aligned}
P(\{5, 6, 7, 8, \ldots\}) &= \lim_{n \to \infty} P(\{5, 6, 7, 8, \ldots, n\}) \\
&= \lim_{n \to \infty} (P(5) + P(6) + \cdots + P(n)) \\
&= \lim_{n \to \infty} \left(2^{-5} + 2^{-6} + \cdots + 2^{-n}\right) = \lim_{n \to \infty} \left(\frac{2^{-5} - 2^{-n-1}}{1 - 2^{-1}}\right) \\
&= \lim_{n \to \infty} \left(2^{-4} - 2^{-n}\right) = 2^{-4} = 1/16.
\end{aligned}
$$

Alternatively, we could use countable additivity directly, to conclude that

$$
P(\{5, 6, 7, 8, \ldots\}) = P(5) + P(6) + P(7) + \cdots ,
$$

which amounts to the same thing. ∎

### EXAMPLE 1.6.2

Let $P$ be *some* probability measure on the space $S = R^1$. Suppose

$$
P\left((3, 5 + 1/n)\right) \geq \delta
$$

for all $n$, where $\delta > 0$. Let $A_n = (3, 5 + 1/n)$. Then $\{A_n\} \searrow A$ where $A = (3, 5]$. Hence, we must have $P(A) = P((3, 5]) \geq \delta$ as well.

Note, however, that we could still have $P((3, 5)) = 0$. For example, perhaps $P(\{5\}) = \delta$, but $P((3, 5)) = 0$. ∎

## Summary of Section 1.6

- If $\{A_n\} \nearrow A$ or $\{A_n\} \searrow A$, then $\lim_{n \to \infty} P(A_n) = P(A)$.
- This allows us to compute or bound various probabilities that otherwise could not be understood.

## EXERCISES

**1.6.1** Suppose that $S = \{1, 2, 3, \ldots\}$ is the set of all positive integers and that $P(\{s\}) = 2^{-s}$ for all $s \in S$. Compute $P(A)$ where $A = \{2, 4, 6, \ldots\}$ is the set of all *even* positive integers. Do this in two ways — by using continuity of $P$ (together with finite additivity) and by using countable additivity.

**1.6.2** Consider the uniform distribution on [0, 1]. Compute (with proof)

$$
\lim_{n \to \infty} P([1/4, \ 1 - e^{-n}]).
$$

**1.6.3** Suppose that $S = \{1, 2, 3, \ldots\}$ is the set of all positive integers and that $P$ is *some* probability measure on $S$. Prove that we must have

$$\lim_{n \to \infty} P(\{1, 2, \ldots, n\}) = 1.$$

**1.6.4** Suppose $P([0, \frac{8}{4+n}]) = \frac{2+e^{-n}}{6}$ for all $n = 1, 2, 3, \ldots$ . What must $P(\{0\})$ be?

**1.6.5** Suppose $P([0, 1]) = 1$, but $P([1/n, 1]) = 0$ for all $n = 1, 2, 3, \ldots$ . What must $P(\{0\})$ be?

**1.6.6** Suppose $P([1/n, 1/2]) \le 1/3$ for all $n = 1, 2, 3, \ldots$ .
(a) Must we have $P((0, 1/2]) \le 1/3$?
(b) Must we have $P([0, 1/2]) \le 1/3$?

**1.6.7** Suppose $P([0, \infty)) = 1$. Prove that there is some $n$ such that $P([0, n]) > 0.9$.

**1.6.8** Suppose $P((0, 1/2]) = 1/3$. Prove that there is some $n$ such that $P([1/n, 1/2]) > 1/4$.

**1.6.9** Suppose $P([0, 1/2]) = 1/3$. Must there be some $n$ such that $P([1/n, 1/2]) > 1/4$?

## PROBLEMS

**1.6.10** Let $P$ be *some* probability measure on sample space $S = [0, 1]$.
(a) Prove that we must have $\lim_{n \to \infty} P((0, 1/n)) = 0$.
(b) Show by example that we might have $\lim_{n \to \infty} P([0, 1/n)) > 0$.

## CHALLENGES

**1.6.11** Suppose we know that $P$ is *finitely* additive, but we do *not* know that it is *countably* additive. In other words, we know that $P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n)$ for any finite collection of disjoint events $\{A_1, \ldots, A_n\}$, but we do not know about $P(A_1 \cup A_2 \cup \cdots)$ for infinite collections of disjoint events. Suppose further that we know that $P$ is continuous in the sense of Theorem 1.6.1. Using this, give a *proof* that $P$ must be countably additive. (In effect, you are proving that continuity of $P$ is *equivalent* to countable additivity of $P$, at least once we know that $P$ is finitely additive.)

# 1.7 | Further Proofs (Advanced)

## Proof of Theorem 1.3.4

*We want to prove that whenever $A_1, A_2, \ldots$ is a finite or countably infinite sequence of events, not necessarily disjoint, then $P(A_1 \cup A_2 \cup \cdots) \le P(A_1) + P(A_2) + \cdots$.*

Let $B_1 = A_1$, and for $n \ge 2$, let $B_n = A_n \cap (A_1 \cup \cdots \cup A_{n-1})^c$. Then $B_1, B_2, \ldots$ are disjoint, $B_1 \cup B_2 \cup \cdots = A_1 \cup A_2 \cup \cdots$ and, by additivity,

$$P(A_1 \cup A_2 \cup \cdots) = P(B_1 \cup B_2 \cup \cdots) = P(B_1) + P(B_2) + \cdots. \qquad (1.7.1)$$

Furthermore, $A_n \supseteq B_n$, so by monotonicity, we have $P(A_n) \geq P(B_n)$. It follows from (1.7.1) that

$$P(A_1 \cup A_2 \cup \cdots) = P(B_1) + P(B_2) + \cdots \leq P(A_1) + P(A_2) + \cdots$$

as claimed. ∎

## Proof of Theorem 1.6.1

*We want to prove that when $A, A_1, A_2, \ldots$ are events, and either $\{A_n\} \nearrow A$ or $\{A_n\} \searrow A$, then $\lim_{n \to \infty} P(A_n) = P(A)$.*

Suppose first that $\{A_n\} \nearrow A$. Then we can write

$$A = A_1 \cup (A_2 \cap A_1^c) \cup (A_3 \cap A_2^c) \cup \cdots$$

where the union is disjoint. Hence, by additivity,

$$P(A) = P(A_1) + P(A_2 \cap A_1^c) + P(A_3 \cap A_2^c) + \cdots .$$

Now, by definition, writing this infinite sum is the same thing as writing

$$P(A) = \lim_{n \to \infty} \left( P(A_1) + P(A_2 \cap A_1^c) + \cdots + P(A_n \cap A_{n-1}^c) \right). \qquad (1.7.2)$$

However, again by additivity, we see that

$$P(A_1) + P(A_2 \cap A_1^c) + P(A_3 \cap A_2^c) + \cdots + P(A_n \cap A_{n-1}^c) = P(A_n).$$

Substituting this information into (1.7.2), we obtain $P(A) = \lim_{n \to \infty} P(A_n)$, which was to be proved.

Suppose now that $\{A_n\} \searrow A$. Let $B_n = A_n^c$, and let $B = A^c$. Then we see that $\{B_n\} \nearrow B$ (why?). Hence, by what we just proved, we must have $P(B) = \lim_{n \to \infty} P(B_n)$. But then, using (1.3.1), we have

$$1 - P(A) = \lim_{n \to \infty} \{1 - P(A_n)\},$$

from which it follows that $P(A) = \lim_{n \to \infty} P(A_n)$. This completes the proof. ∎