We use the word “probability” in everyday speech in a variety of ways. For example, we might ask about the probability of rain tomorrow, or we might speculate about the probability that we will get an A on a test. However, the mathematical study of probability requires a more precise definition of the term. Agreeing on such a precise definition and its interpretation turns out to be not such an easy task. The mathematical definition of probability that we give in the next section has two main types of interpretations: frequentist and subjective.

In a frequentist interpretation of probability, it only makes sense to discuss the probability of an event resulting from a procedure that can be repeated many times (infinitely many times in theory). For example, in frequentist probability, we might ask about the probability that a particular coin will land HEADS if it is flipped. In the frequentist view, the probability of an event is interpreted as the proportion of occurrences of that event if the procedure were to be repeated many times (infinitely many times in theory).

In a subjective probability, the probability of nearly any kind of event can be discussed. From this point of view, each individual associates a probability with an event based on that individual’s belief about how
likely the event is. For example, in subjective probability, we might ask about the probability that a particular coin that has just been flipped has landed \textit{heads}. This probability may well vary depending on who is assessing it. In frequentist probability, it wouldn't make sense to ask about such a probability because while the procedure of flipping the coin can certainly be repeated, that particular flip (including its outcome) is not a repeatable procedure.

Neither of these points of view has been shown to be unequivocally superior to the other, and both are completely consistent with the standard mathematical theory. Also, each point of view has its advantages and disadvantages in different situations. Both points of view have their own complications and complexities as well. Computability of results and other factors (including many developments of Fisher\textsuperscript{1}) have led to the dominance of frequentist probability over the past 50 years or more. However, with the increase in computing power, subjective probability has been making somewhat of a comeback recently.

We will use the frequentist interpretation here almost exclusively. When we do refer to subjective probability, we will generally say so explicitly. In any case though, you should be aware that there are different interpretations of the mathematical theory of probability.

\section{1.1 Defining probability}

In order to give a mathematical definition of probability, we should first introduce a few preliminary concepts.

\textit{Random procedures}

Since we are using a frequentist interpretation of probability, we begin by defining the procedures that lead to events that we consider to have probabilities.

\textbf{Definition 1.1.1} A \textbf{random procedure} is a \textit{repeatable} procedure (whose result \textit{may not be the same on different repeats}). Each time the random

\footnote{\textsuperscript{1}Sir Ronald Aylmer Fisher (1890-1962) was a great pioneer of statistics. Among many other things, he developed much of the material covered in this text in one form or another. For more about his life and works, see his MacTutor website biography at http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Fisher.html.}
procedure occurs is called a **trial**, and a result of a trial is called an **outcome**.

The key aspect to a random procedure from the frequentist point of view is its repeatability, even if repetitions of the procedure do not always yield the same results. The classic example to keep in mind is that of flipping a coin. The procedure of flipping a coin can be repeated even though the **HEA**DS or **TAILS** result may not be the same for all the flips.

In actuality, we don’t expect anything to be *exactly* repeatable. Even for something as simple as flipping a coin, all sorts of things will be different on each flip (such as atmospheric conditions, position of the earth, the precise way that it is flipped, etc.). However, we often model situations as being repeatable. This is a simplifying assumption that turns out to be extremely useful in a wide variety of cases.

Now that we have defined the procedures of interest to us here, we introduce some vocabulary to discuss the results of such procedures.

**Definition 1.1.2** Any collection of outcomes is called an **event**. The collection of all possible outcomes of a random procedure is called the **sample space** of the random procedure. We denote events (including outcomes) by capital letters such as $A$ and $B$, and we denote the event that is the sample space of a random procedure by $S$.

Events can be described either by giving conditions that determine a collection of outcomes, or by simply stating the outcomes in the event. For example, consider a random procedure that is flipping a particular coin twice. One possible outcome of this random procedure is that the coin lands **HEA**DS on the first flip and **TAILS** on the second flip, which we will abbreviate $HT$ (using similar abbreviations for other outcomes). One possible event resulting from this random procedure is that the two flips have different outcomes. This event is the same as the collection of outcomes $\{HT, TH\}$. The sample space for this random procedure is the collection of outcomes $\{HH, HT, TH, TT\}$.

**Generating new events**

There are several ways to construct new events from given events. We now describe a few such ways.

**Definition 1.1.3** Let $A$ and $B$ be events. Then $A$ and $B$ is defined to be the event consisting of all the outcomes that are included in both $A$ and $B$. This
definition extends to more than two events as well. For example, suppose that A, B, and C are events. Then A and B and C is defined to be the event consisting of all the outcomes that are included in A, B, and C.

For example, consider the random process of flipping a particular coin three times. One event is that all three flips result in the same outcome, which is the same as \{HHH, TTT\}. Another event is that the first flip results in HEADS, which is the same as \{HHH, HHT, HTH, HTT\}. If we denote the first of these events by A and the second by B, then A and B is the event that all three flips result in the same outcome and the first flip results in HEADS. This is the same as \{HHH\}, the only outcome that the original two events have in common.

**Definition 1.1.4** Let A and B be events. Then A or B is defined to be the event consisting of all the outcomes that are included in A or B or both. This definition extends to more than two events as well. For example, suppose that A, B, and C are events. Then A or B or C is defined to be the event consisting of all the outcomes that are included in one or more of these three events.

Continuing with coin-flipping example above, A or B is the event that either all three flips result in the same outcome or the first flip results in HEADS (or both). This is the same as \{HHH, HHT, HTH, HTT, TTT\}.

We now describe one way to form a new event from a single given event.

**Definition 1.1.5** Let A be an event. Its complement \(A^c\) is the event consisting of all the outcomes not included in A.

Be sure not spell this word as “compliment”, which refers to saying something nice about someone.

As an example to illustrate the complement of an event, consider the random process of flipping a particular coin twice and the event A that both flips have the same outcome (meaning \{HH, TT\}). Then \(A^c\) is the event that both flips do not have the same outcome (meaning \{HT, TH\}).

One other concept related to the way that events combine is worth introducing here.

**Definition 1.1.6** Two events are called disjoint if they have no outcomes in common.

In symbols, two events A and B are disjoint if A and B is empty (that is, consists of no outcomes).
For example, consider again the random process of flipping a particular coin twice. Let \( A \) be the event that both flips have the same outcome (meaning \{HH, TT\}), and let \( B \) be the event that the first flip is heads and the second flip is tails (meaning \{HT\}). Then \( A \) and \( B \) are disjoint. If \( C \) is the event that at least one flip results in heads (meaning \{HH, HT, TH\}), then \( A \) and \( C \) are not disjoint, since they have the outcome \( HH \) in common.

**Discrete and continuous random procedures**

Before we begin to explore the rules of probability, we now introduce the distinction between a discrete random procedure and a continuous random procedure.

**Definition 1.1.7** If the sample space of a random procedure consists of a collection of outcomes that can be counted (at least in theory), the random procedure is called **discrete**.\(^2\) If the sample space of a random procedure consists of a continuous collection of elements, we call the random procedure **continuous**.

An example of a discrete random procedure is that of flipping some particular coin, where the outcomes in the sample space can easily be counted (since there are only two of them). A more complicated example of a discrete random procedure is that of flipping some particular coin repeatedly until heads appears. Although there are infinitely many outcomes in the sample space in this case, they can be listed out in a way that allows them to be counted (even if it would take infinitely long to do so). More specifically, we can list the outcomes as: heads first on the first flip, heads first on the second flip, heads first on the third flip, etc.

An example of a continuous random procedure is that of spinning a board game spinner that can point outward along any radius of the circle that it pivots around. Since the spinner can land pointing at any angle of the circle, this random procedure has a continuous range of outcomes.

It is an interesting mathematical theorem that continuous procedures are not discrete, and discrete procedures are not continuous. There are other kinds of procedures, such as combinations of these two, but we will not go into them here. In this chapter, all random procedures will be assumed to be discrete unless otherwise noted. We will discuss continuous

\(^2\)This isn’t quite the same as the usual more complicated mathematical definition of discrete, but it will be suitable for our purposes.
random procedures starting in Chapter 4.

For discrete random procedures, the following notation is useful.

**Definition 1.1.8** Let \( A \) be an event for a discrete random procedure. We define \(|A|\) to be the number of outcomes in \( A \).

For example, consider the random procedure that is flipping a coin twice, and let \( A \) be the event that the first flip is \text{HEADS} (meaning \{HH, HT\}). Then \(|A| = 2\), since there are two outcomes in \( A \). Also, \(|S| = 4\) in this case, since there are four outcomes in the sample space.

**Probability functions**

We can now give a precise mathematical definition of probabilities of events associated with discrete random procedures.

**Definition 1.1.9** A probability function of a random procedure is a function \( \text{Prob}(\ ) \) from the random procedure’s sample space to the real numbers satisfying:

1. \( 0 \leq \text{Prob}(A) \leq 1 \) for any event \( A \).
2. \( \text{Prob}(S) = 1 \) for the event \( S \) that is the entire sample space.
3. \( \text{Prob}(A \text{ or } B) = \text{Prob}(A) + \text{Prob}(B) \) for any disjoint events \( A \) and \( B \).
4. \( \text{Prob}(A^c) = 1 - \text{Prob}(A) \) for any event \( A \).

For any event \( A \), the real number \( \text{Prob}(A) \) is called the probability of \( A \).

Note that the third item, regarding the additivity of probabilities, applies only to disjoint events. It is emphatically not true for events that are not disjoint.

Many rules about probabilities can be derived from the properties listed in the definition. For example, probability is additive not just for any two disjoint events, but also for any finite collection of events in which any two are disjoint. This can be used to prove the following proposition, although we do not give a proof here.

**Proposition 1.1.10** Suppose that the sample space \( S \) for a random procedure consists of finitely many outcomes and that all the outcomes have the same probability. Then the probability of any individual outcome is \( 1/|S| \).
As we will soon see, this is a very useful proposition. However, we must be careful not to misinterpret it. Just because the sample space of a random phenomenon consists of only finitely many outcomes does not mean that the probability of each outcome equals 1 divided by the number of outcomes in the sample space. In order to draw this conclusion, we must also know that all the outcomes have the same probability. Without this additional piece of information, the conclusion of the proposition is not valid.

For discrete random procedures whose outcomes all have the same probability, Proposition 1.1.10 tells us how to compute the probabilities of individual outcomes, but what about probabilities of events consisting of more than one outcome? Well, because any two individual outcomes are disjoint events, we can combine Item 3 of the definition of a probability function (Definition 1.1.9 on page 6) with this proposition to obtain the following result.

**Proposition 1.1.11** Suppose that the sample space $S$ for a random procedure consists of finitely many outcomes and that all the outcomes have the same probability. Then for any event $A$,

$$\text{Prob}(A) = \frac{|A|}{|S|}.$$  

In other words, if all the outcomes have the same probability, then in order to compute the probability of an event, we need only count the number of outcomes in the event and divide by the total number of outcomes in the sample space. So for the same probability case, computing an event’s probability becomes a counting problem.

## 1.2 Conditional probability

We now introduce the concept of conditional probability, which is useful for understanding and describing relationships between events.

**Defining conditional probability**

The definition of conditional probability is as follows.
Definition 1.2.1 Let $A$ and $B$ be events. The (conditional) probability of $A$ given $B$ is defined to be the probability of $A$ if $B$ occurs. The probability of $A$ given $B$ is denoted by $\text{Prob}(A|B)$.

We illustrate this with an example.

Example 1.2.2 Consider again the random procedure that is flipping a particular coin twice, and let’s assume that we have determined that each of the four possible outcomes $HH$, $HT$, $TH$, and $TT$ occurs with probability $1/4$ for this flipping process. Also, let $A$ be the event $\{HH\}$, and let $B$ be the event that both flips give the same result (meaning $\{HH, TT\}$). Then $\text{Prob}(A) = 1/4$ but $\text{Prob}(A|B) = 1/2$ because the sample space given that both flips give the same result consists of exactly 2 outcomes with the same probability, so Proposition 1.1.10 on page 6 tells us that each outcome has probability $1/2$.

A key feature of conditional probability, which could actually be taken as the definition, is given by the following proposition, which we state without proof.

Proposition 1.2.3 For any events $A$ and $B$ associated with the same random procedure,

$$
\text{Prob}(A|B) = \frac{\text{Prob}(A \text{ and } B)}{\text{Prob}(B)}.
$$

We can use the example above to illustrate this proposition.

Example 1.2.2 (continued) Using the events $A$ and $B$ from the example above, we compute that $A$ and $B = \{HH\}$, so $\text{Prob}(A \text{ and } B) = 1/4$. Putting this together with the probability of $B$ given in the example above, the formula in Proposition 1.2.3 gives us

$$
\text{Prob}(A|B) = \frac{1/4}{1/2} = \frac{1}{2},
$$

which agrees with our previous computation of $\text{Prob}(A|B)$.

This proposition will be particularly important soon when we compute probabilities of sequences of events.

Independence of events

The notion of conditional probability allows us to define another concept in probability theory, that of independence of events.
Definition 1.2.4 Two events $A$ and $B$ are called independent if the $\Prob(A|B) = \Prob(A)$. Although it isn’t obvious, this turns out to be the same thing as $\Prob(B|A) = \Prob(B)$, so either equation can be used as the definition.

In other words, two events are independent if knowing that one occurs gives no additional information about the probability that the other occurs.

To illustrate, let’s continue the coin-flipping example above.

Example 1.2.2 (continued) We computed that $\Prob(A) = 1/4$ and that $\Prob(A|B) = 1/2$. Since these two are not equal, the events $A$ and $B$ are not independent.

On the other hand, consider the event $C$ that the first flip results in heads (that is, $\{HH, HT\}$). Since all the outcomes have the same probability, we know that $\Prob(C) = 2/4 = 1/2$, and $\Prob(C$ and $A) = \Prob(\{HH\}) = 1/4$. Also, we compute

$$\Prob(C|A) = \frac{\Prob(C$ and $A)}{\Prob(A)} = \frac{1/4}{1/2} = \frac{1}{2} = \Prob(C).$$

Since $\Prob(C|A) = \Prob(C)$, the events $A$ and $C$ are independent. ♦

How to compute the probability that both of the two independent events occur follows immediately from Proposition 1.2.3 on page 8, as we state in the following proposition.

Proposition 1.2.5 Let $A$ and $B$ be independent events associated with the same random procedure. Then

$$\Prob(A$ and $B) = \Prob(A) \Prob(B).$$

In fact, this proposition is equivalent to our definition of independence of events. Many sources actually define two events to be independent if and only if their probabilities multiply.

The independence of the events is important in this proposition. It is emphatically not true that the probabilities of any two events multiply.

1.3 Methods of computing probabilities

In this text, we use three main methods of computing probabilities of events associated with discrete random processes: counting techniques, trees, and complements. As we will see, the various methods are often most useful when used in conjunction with each other rather than individually. We begin by examining counting techniques.
Counting techniques

Counting techniques can be useful with discrete random procedures whose outcomes all have the same probability. For example, to compute the probability of any individual outcome of such a random procedure, Proposition 1.1.10 on page 6 tells us that we need only divide 1 by the number of possible outcomes in the sample space. This means that in order to compute such a probability, we need only count the number of possible outcomes in the sample space.

To illustrate, consider the case of a fair flipping of a coin, meaning one for which both outcomes have the same probability (a similar definition applying for rolling of dice, etc.). Since each outcome has the same probability, and since the total number of possible outcomes is 2, the probability of each individual outcome equals $1/2 = 0.5$. For another example, consider the case of choosing a card at random from a standard deck of playing cards (which does not contain any Jokers and so contains 52 cards). Since we choose at random, each card has equal probability. Consequently, the probability of getting any particular card equals $1/52 = 0.019$. (Note that when doing probability computations of this sort, it is usually more informative to write fractions instead of their decimal approximations.)

Proposition 1.1.11 on page 7 allows us to apply counting techniques to a somewhat wider range of problems. For example, if we choose a card at random from a standard deck of playing cards, what is the probability of obtaining an Ace? Well, there are 52 outcomes, all with the same probability, so we need only count the number of outcomes in the event of obtaining an Ace. Since there are 4 outcomes in this event (Ace of Clubs, Ace of Diamonds, Ace of Hearts, and Ace of Spades), the above proposition tells us that the probability of obtaining an Ace equals $4/52 = 0.077$.

While counting techniques are useful in their own right, they become much more so when combined with other methods of computing probabilities, such as trees.

Trees

The method of computing probabilities using trees can be applied when the random procedure in question can be broken down into a sequence of steps, each of which is itself a random procedure. To illustrate with an example, let’s consider the random procedure of flipping a fair coin three
times. A tree for this random procedure is pictured in Figure 1.3.1.

The dots that are connected by line segments in a tree are called nodes (or vertices, whose singular is vertex), and the line segments connecting them are called edges. The node on the far left, which has no edges to its left, is called the root node.

To construct a tree for a random procedure, we begin by creating a root node to represent the first step in the procedure. Then we create one edge emanating from the root node to the right (and ending in a new node) for each possible outcome of the first step in the procedure, and we label each edge with the probability of that outcome in the first step of the procedure. Each of the newly created nodes now represents a possible situation as we move to the second step in the procedure. For each newly created node, we repeat the process until we are at the end of the procedure, labeling the edges with the (unconditional) probabilities of the corresponding outcomes in each step.

Once a tree has been constructed, computing the probability of each possible outcome of the random procedure is straightforward: we simply multiply the probability of each edge in the path corresponding to that outcome. For example, consider the outcome HTT, whose corresponding path is dotted in the tree in Figure 1.3.2.

Multiplying all the probabilities on the corresponding path, we com-
compute that the probability of the outcome $HTT$ equals $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8} = 0.125$. In fact, in this particular case, is easy to see from the tree that the probability of any single outcome of this random process equals $\frac{1}{8} = 0.125$.

Why does this work? Well, the probability of the outcome $HTT$ is equal to the probability of $HT$ on the first two flips times the probability of $HTT$ given that the first two flips were $HT$ (by Proposition 1.2.3 on page 8). However, the probability of $HTT$ given that the first two flips were $HT$ is precisely the probability labeling the last edge in the path. In other words, the probability of the outcome $HTT$ is equal to the probability of $HT$ on the first two flips times the probability labeling the last edge in the path.

By the same reasoning, the probability of $HT$ on the first two flips is equal to the probability of $H$ on the first flip times the probability of $HT$ on the first two flips given that the first flip was $H$. Of course, the probability of $HT$ on the first two flips given that the first flip was $H$ is the probability labeling the second edge of the path, and the probability of $H$ on the first flip is the probability labeling the first edge of the path. Putting all this together, we find that the probability of the outcome $HTT$ equals the product of the probabilities of the edges along the corresponding path.

For a slightly more complicated example, suppose we would like to
compute the probability that two cards drawn at random from a standard deck of playing cards are both the same suit. Since drawing two cards at random can be broken into two steps, drawing one card in each step, we can use a tree to approach this problem. However, since there are 52 cards in the deck, drawing a tree with each card representing a separate outcome would quickly become unwieldy.

We can use a counting technique to “trim” the tree a bit though. Rather than thinking of each card as a separate outcome in each step of the random procedure, let’s think of each suit as a separate outcome. Since there are only four suits, this will make for only 4 edges emanating to the right from each node, which is much more manageable. The counting technique will come in when we compute the probability of each outcome of each step, as we discussed below, but first let’s draw the tree (Figure 1.3.3, with the outcomes in the event of interest circled) and then explain where the probabilities come from.

The probabilities associated with the first draw (by the way, it doesn’t matter which one we term the first in which the second draw, as long as we are consistent throughout the problem) are readily computed by counting. Since each of the 52 cards is equally likely to be drawn, the probability of obtaining a card of any given suit equals the number of cards of that suit divided by the number of cards in the deck, which equals $13/52 = 1/4 = 0.25$.

On the other hand, we need to be a little more careful with the probabilities associated with the second draw, although we can still use a counting technique. The probability of drawing a card of any given suit still equals the number of cards of that suit divided by the number of cards remaining in the deck (which is now 51), but the number of cards of the given suit depends on whether or not that suit is the same as the one for the card first drawn. If it is, and then only 12 cards of that suit remain; if not, then thirteen cards of that suit remain. Therefore the probability of drawing a card of a given suit on the second draw is either $12/51 = 0.235$ (if the suit is the same as the suit of the first card drawn) or $13/51 = 0.255$ (if not).

Since the event that we are interested in consists of 4 individual outcomes (any two of which are of course disjoint), we can compute its probability by adding the probabilities of the 4 outcomes. The probability of each individual outcome in the event equals $(1/4)(12/51) = 3/51 = 0.059$. (For those who are interested, it would be good to think of another way to understand intuitively why the probability equals $3/51$ without
Figure 1.3.3. The possible suits of two cards drawn at random.
Note, by the way, that the probability of getting two cards of the same suit is not simply \((1/4)(1/4)\), as it would be if the two draws were independent. They are not independent, since the suit of the first card drawn affects the probabilities of drawing the various suits on the second draw.

**Complements**

Sometimes both counting techniques and trees might be unwieldy for a particular problem. In this case, it is worth thinking about whether it would be easier to compute the probability of the complement of the event of interest rather than computing the probability of the event itself. Knowing the probability of the complement is enough, since the probability of the event is simply 1 minus the probability of the complement of the event.

For example, suppose we would like to know the probability that at most 7 cards will be diamonds if we draw 8 cards at random from a standard deck of playing cards. Trying to count the number of ways that this can occur, with or without drawing a tree, quickly appears to be rather unmanageable. However, the complement of the event that at most 7 cards are diamonds is the event that all 8 cards are diamonds (if that doesn’t make sense, think about it until it does).

To compute the probability that all eight cards are diamonds, imagine drawing the tree (without actually drawing it). Since this event consists of only a single outcome, we really only need to draw a single path in the tree, namely the one corresponding to this outcome. Along this path, the probability of the first edge is the probability of drawing a diamond on the first draw, which equals \(13/52 = 1/4\) (as we saw above, since each of the 52 cards is equally likely). The probability of the second edge in this path is the probability of drawing a diamond on the second draw given that we drew a diamond on the first draw, which equals \(12/51\) (since there are 12 diamonds left and each of the 51 remaining cards is equally likely to be drawn), and so on. The product of all 8 edges in the path equals

\[
\frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48} \cdot \frac{8}{47} \cdot \frac{7}{46} \cdot \frac{6}{45} = 0.0000017,
\]

which is then the probability that all 8 cards drawn were diamonds. The probability of the event that at most 7 of the cards drawn were diamonds
is therefore equal to

\[ 1 - 0.0000017 = 0.9999983. \]

In this example, using the complement made it so that instead of needing to figure out the probability of all but one of the paths in a rather large tree (and it still would have been large even if we had “trimmed” it), we only had to use a single branch of the tree, which was quite manageable. Of course, we were still using counting techniques and trees along with the complement method, but using these techniques in conjunction with each other should be becoming fairly familiar.