Queen's University
School of Computing

CISC 203: Discrete Mathematics for Computing II
Lecture 3: Discrete Probability
Winter 2019

1 Probability Theory Fundamentals

Fundamentally, a computer is a deterministic machine. By "deterministic", we mean that if we perform the same actions multiple times, we will produce the same result each time. Each time we run a (simple, sequential) program, the CPU executes the same instructions and the same values are stored in memory. Indeed, if this were not the case, then computers would not be a very useful tool at all. Why, then, do computer scientists care about probability theory; a field where performing the same action multiple times can result in multiple, distinct outcomes occurring with varying odds?

Well, for starters, not all programs are sequential, as we assumed previously. Some programs are parallelized, with different parts of the program running simultaneously and possibly finishing their computations in a different order each time the program is run. Other programs are randomized, so that each run of the program results in a potentially different output. In fact, some programs rely on parallelization or randomization to produce output in a reasonable amount of time (or at all). Thus, it is important for computer scientists to have a working knowledge of probability theory in order to understand more advanced topics like parallel computing and randomized algorithms.

There are two branches of probability theory: discrete probability and continuous probability. Discrete probability is probability theory in the context of finite or countable probability spaces, where we can compute probabilities of multiple events via summation. Continuous probability, on the other hand, is probability theory in the context of uncountable probability spaces, where we can compute probabilities of multiple events via integration. In keeping with the theme of this course, we will focus here only on discrete probability; continuous probability is a whole other beast, and discussion on that topic will be left for another course.

1.1 Definitions

To discuss discrete probability, we must first define the basic elements of probability theory. When we perform an experiment, like flipping a coin, rolling a die, or drawing a playing card from a deck, there exists a certain set of outcomes; for instance, flipping a heads, rolling a 6, or drawing an ace of spades. A collection of these outcomes is called an **event**. The set of all outcomes is called the **sample space**, and it is usually denoted by the symbol Ω.

Example 1. The sample space for flipping a single coin consists of the two outcomes

$$\Omega_{\mathrm{C}} = \{\text{Ⓗ}, \text{Ⓣ}\}.$$

The sample space for rolling a single die consists of the six outcomes

$$\Omega_{\mathrm{D}} = \{\text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅}\}.$$

The sample space for drawing a single playing card from a deck consists of the 52 outcomes

$$\Omega_{\text{PC}} = \left\{ \begin{array}{lllllllllllll} \text{A}\clubsuit, & \text{K}\clubsuit, & \text{Q}\clubsuit, & \text{J}\clubsuit, & 10\clubsuit, & 9\clubsuit, & 8\clubsuit, & 7\clubsuit, & 6\clubsuit, & 5\clubsuit, & 4\clubsuit, & 3\clubsuit, & 2\clubsuit \\ \text{A}\diamondsuit, & \text{K}\diamondsuit, & \text{Q}\diamondsuit, & \text{J}\diamondsuit, & 10\diamondsuit, & 9\diamondsuit, & 8\diamondsuit, & 7\diamondsuit, & 6\diamondsuit, & 5\diamondsuit, & 4\diamondsuit, & 3\diamondsuit, & 2\diamondsuit \\ \text{A}\heartsuit, & \text{K}\heartsuit, & \text{Q}\heartsuit, & \text{J}\heartsuit, & 10\heartsuit, & 9\heartsuit, & 8\heartsuit, & 7\heartsuit, & 6\heartsuit, & 5\heartsuit, & 4\heartsuit, & 3\heartsuit, & 2\heartsuit \\ \text{A}\spadesuit, & \text{K}\spadesuit, & \text{Q}\spadesuit, & \text{J}\spadesuit, & 10\spadesuit, & 9\spadesuit, & 8\spadesuit, & 7\spadesuit, & 6\spadesuit, & 5\spadesuit, & 4\spadesuit, & 3\spadesuit, & 2\spadesuit \end{array} \right\}.$$

When we perform an experiment and observe an outcome, we say that the event containing that outcome "occurs". For instance, if we roll a die and get a two, the event {⚁} occurred. At the same time, the event that we rolled an even number—that is, the event {⚁, ⚃, ⚅}—also occurred.

With our notions of events and sample spaces, we can now define the concept of **probability** itself.

Definition 2 (Probability—equal likelihood)**.** The probability that an event E occurs, where E is taken as a subset of a sample space Ω consisting of a finite number of equally-likely outcomes, is

$$\mathbb{P}[E] = \frac{|E|}{|\Omega|}.$$

In our definition of probability, we used the phrase "equally-likely outcomes". This means that all outcomes have an equal probability of being observed. For instance, if we flip a fair coin, we get that the probability of flipping a heads or a tails is the same; that is, $\mathbb{P}[\{\text{Ⓗ}\}] = \mathbb{P}[\{\text{Ⓣ}\}] = 1/2$. Likewise, if we roll a fair die, we get that

$$\mathbb{P}[\{\text{⚀}\}] = \mathbb{P}[\{\text{⚁}\}] = \mathbb{P}[\{\text{⚂}\}] = \mathbb{P}[\{\text{⚃}\}] = \mathbb{P}[\{\text{⚄}\}] = \mathbb{P}[\{\text{⚅}\}] = 1/6.$$

This definition is sometimes called the "classical definition" of probability, and it dates back to an 1812 work by the French scientist Pierre-Simon Laplace titled *Théorie analytique des probabilités*.

Naturally, though, we cannot always assume that all outcomes are equally likely in every experiment we perform. The world just isn't that nice. Therefore, we must generalize our definition of probability to account for outcomes that may be observed either more or less frequently than other outcomes. Fortunately, it is easy for us to generalize our definition: we just take the probability of an event to be the sum of the probabilities of each outcome in the event.

Definition 3 (Probability—general)**.** The probability that an event E occurs, where E is taken as a subset of a sample space Ω consisting of a countable number of outcomes, is

$$\mathbb{P}[E] = \sum_{\omega \in E} \mathbb{P}[\omega].$$

At this point, you may be wondering what the symbol "$\mathbb{P}$" is meant to denote. This symbol represents a **probability measure**; given a sample space Ω, a probability measure $\mathbb{P} : \Omega \to \mathbb{R}$ maps elements of the sample space (that is, outcomes) to real numbers (that is, probabilities). In other words, a probability measure calculates the probability of some outcome being observed. Note that, even though a probability measure maps to the set $\mathbb{R}$, we denote the probability of an outcome using just the set $[0, 1]$.

Probability measures abide by a couple of important properties, sometimes referred to as "Kolmogorov's probability axioms" after the Soviet mathematician Andrey Kolmogorov. These properties are as follows.

Proposition 4. *Let Ω be a sample space consisting of a countable number of outcomes, and let ω be an outcome. Then, for all probability measures $\mathbb{P}$ on Ω, the following are true:*

1. $0 \leq \mathbb{P}[\omega] \leq 1$ *for all* $\omega \in \Omega$, *and*
2. $\sum_{\omega \in \Omega} \mathbb{P}[\omega] = 1$.

As a consequence of these properties, we get that $\mathbb{P}[\emptyset] = 0$ (the probability of nothing from the sample space happening is 0) and $\mathbb{P}[\Omega] = 1$ (the probability of something from the sample space happening is 1). We say that an event with probability 0 occurs "almost never" and an event with probability 1 occurs "almost surely".

Remark. Using probability measures, we can define **probability distributions**. In fact, we've already seen one probability distribution: the **uniform distribution** assigns the probability $\mathbb{P}[\omega] = 1/n$ to each of n outcomes $\omega \in \Omega$. We will see many more probability distributions later in these notes.

Combining sample spaces and probability measures, we get the last fundamental definition we require: the notion of a **probability space**, which is simply the mathematical formalization of an experiment.

Definition 5 (Probability space)**.** A probability space $(\Omega, \mathbb{P})$ is a tuple containing a sample space Ω consisting of a countable number of outcomes and a probability measure $\mathbb{P} : \Omega \to \mathbb{R}$.

If you're tired of seeing definitions by now, that's understandable. Let's shift gears by looking at a few basic examples of probability.

Example 6. Consider the "flipping coins" sample space, Ω_C = {Ⓗ, Ⓣ}. If we flip two fair coins in sequence, the probability space that models this experiment consists of the sample space $\Omega = \Omega_\mathrm{C} \times \Omega_\mathrm{C}$ = {ⒽⒽ, ⒽⓉ, ⓉⒽ, ⓉⓉ} together with the probability measure $\mathbb{P}[\omega] = 1/4$ for all $\omega \in \Omega$.

What is the probability of flipping two fair coins and having the first coin show tails? This is given by $\mathbb{P}$[{ⓉⒽ, ⓉⓉ}] $= 1/4 + 1/4 = 1/2$.

Example 7. Consider the "rolling dice" sample space, Ω_D = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}. If we roll two fair dice in sequence, the probability space that models this experiment consists of the sample space $\Omega = \Omega_\mathrm{D} \times \Omega_\mathrm{D}$ together with the probability measure $\mathbb{P}[\omega] = 1/36$ for all $\omega \in \Omega$.

What is the probability of rolling two fair dice and obtaining a sum of 7? This is given by

$\mathbb{P}$[{⚀⚅, ⚁⚄, ⚂⚃, ⚃⚂, ⚄⚁, ⚅⚀}] $= (1/36 \times 6) = 6/36 = 1/6.$

Example 8. Consider the "drawing cards" sample space, Ω_PC. For each card c in the deck, we have a probability of $\mathbb{P}[c] = 1/52$ that we will draw c.

A "full house" is a poker hand that consists of five cards: three cards are of one rank (or value), and the other two cards are of another rank. For example, Q♦ Q♥ Q♠ 7♣ 7♥ is a full house. For each hand h, we have a probability of $\mathbb{P}[h] = 1/C(52, 5)$ that we will draw h. What is the probability that h is a full house?

To determine this, we must calculate the number of ways we can draw the required cards. First, we choose the rank of the first set of cards: we can do this in $C(13, 1)$ ways. Then, we choose three out of the four suits of that rank in $C(4, 3)$ ways. Next, we choose the rank of the second set in $C(12, 1)$ ways, and the suits of that rank in $C(4, 2)$ ways. Considering the entire deck, we have $C(52, 5)$ ways to draw five cards overall. Therefore, the probability of drawing a full house is

$$\frac{C(13,1)C(4,3)C(12,1)C(4,2)}{C(52,5)} = \frac{(13)(4)(12)(6)}{2\,598\,960} = 0.00144.$$

1.2 Properties

Now that we are familiar with basic elements like sample spaces, events, and probabilities, we can investigate a few properties that are important in probability theory. However, most of these properties will not be particularly novel; in fact, you've likely seen these properties before, just in a different form.

Recall that a sample space contains outcomes, and an event is a collection of outcomes from some sample space. This sounds remarkably familiar to the relationship between sets, elements, and subsets. Indeed, probability theory can be defined formally using set theory: for instance, if we consider a sample space Ω to be a set, we can define an event as $E \subseteq 2^\Omega$ (a subset of the power set of Ω). Generally, we can treat outcomes, events, and sample spaces in the same way we treat elements and sets; we can apply operations to events and outcomes, and we can translate many set-theoretic properties into the language of probability theory. Let's try translating some of these properties now.

Remark. Going forward, unless otherwise stated, assume that we have a generic sample space denoted by Ω.

1.2.1 Combinations

Although considering single events is straightforward, we can't do very much with such basic knowledge. If we establish a method of considering multiple events together, then we can really leverage the power of probability theory to obtain interesting results. In fact, we've already seen examples where we consider multiple events together, so let's formalize some techniques for doing so.

We will begin by considering combinations of events. Here, by "combinations", we don't mean the counting technique; rather, we mean a set of multiple events, or a union of events.

Let's start off easy by again considering a single event. Most times, we only care about the probability of an event E occurring. But, given this information, what is the probability that E does not occur? In other words, what is the probability that $\overline{E}$ occurs, where $\overline{E}$ is the **complementary event** of E?

Theorem 9. *Let $E \in \Omega$ be an event occurring with probability $\mathbb{P}[E]$. The probability that the event $\overline{E}$ occurs is $\mathbb{P}[\overline{E}] = 1 - \mathbb{P}[E]$.*

Proof. By Proposition 4, the sum of the probabilities of all outcomes in Ω must be equal to 1. Since each outcome is either in E or $\overline{E}$, but not in both, we have $\sum_{\omega \in \Omega} \mathbb{P}[\omega] = 1 = \mathbb{P}[E] + \mathbb{P}[\overline{E}]$. Therefore, $\mathbb{P}[\overline{E}] = 1 - \mathbb{P}[E]$. □

From this result, we see that combining the probabilities of the events E and $\overline{E}$ results in a probability of 1; we are certain that either E will occur or E will not occur.

Let's now consider two events E and F, occurring with probability $\mathbb{P}[E]$ and $\mathbb{P}[F]$, respectively. What is the probability that either of these events will occur? By Definition 3, the probability that either event will occur should be equal to the sums of the probabilities of all outcomes in both E and F—and indeed, it is, provided that E and F do not share outcomes. But, just like our combinatorial sum rule did not work for sets that shared elements, so too does this rule not work for events that share outcomes. Again, we run the risk of double counting!

To fix the way we sum probabilities of outcomes, recall our fix for the combinatorial sum rule: the inclusion-exclusion principle counted the elements contained in two sets using the formula $|A \cup B| = |A| + |B| - |A \cap B|$. If we swap out the set-theoretic notation for probability-theoretic notation, we can use the exact same formula to determine the probability that either of two events will occur: namely,

$$\mathbb{P}[E \cup F] = \mathbb{P}[E] + \mathbb{P}[F] - \mathbb{P}[E \cap F].$$

After seeing the above formula, you might rightfully ask how we determine $\mathbb{P}[E \cap F]$. Keep this question in mind; we will see three methods of doing so in the following sections. For now, marvel in the fact that the same idea behind the inclusion-exclusion principle also applies to the probabilities of events.

Example 10. Let b be a randomly-generated bit string of length four, and define events E = "b contains all `1`s" and F = "b starts with two `0`s". Assuming bits are assigned randomly, we have $\mathbb{P}[E] = 1/16$ and $\mathbb{P}[F] = 1/4$.

What is $\overline{E}$? This is the event where b does not contain all `1`s; that is, where b contains at least one `0`. By Theorem 9, we have that $\mathbb{P}[\overline{E}] = 1 - \mathbb{P}[E] = 15/16$. This makes sense, since there are 15 bit strings of length four containing at least one `0`.

What is $\mathbb{P}[E \cup F]$? This is the event where b either contains all `1`s or starts with two `0`s. Note that both of these events cannot occur at once, so we do not need to worry about overcounting. We can simply calculate $\mathbb{P}[E \cup F] = \mathbb{P}[E] + \mathbb{P}[F] = 1/16 + 1/4 = 5/16$.

Just like we did in our lecture on combinatorics, we can generalize our "probabilistic inclusion-exclusion principle" from two events to m events. In doing so, we obtain two general results, depending on whether our events $E_1, \ldots, E_m$ share outcomes or not. We will begin with the generalization where none of our m events share outcomes. (As we will see in the next section, we call such events "disjoint", and so this result is known as the "disjoint union" result.)

Theorem 11 (Disjoint union)**.** *Let $E_1, \ldots, E_m \in \Omega$ be events occurring with probability $\mathbb{P}[E_1], \ldots, \mathbb{P}[E_m]$, respectively. Then*

$$\mathbb{P}\left[\bigcup_{i=1}^{m} E_i\right] = \sum_{i=1}^{m} \mathbb{P}[E_i].$$

Proof. Omitted. □

In this easier generalization, since none of our events $E_1, E_2, \ldots, E_m$ share outcomes, the probability that any of these events will occur is simply the sum of each individual event's probability. Moreover, we know that this sum cannot exceed 1 by the properties given in Proposition 4; since events consist of outcomes, and since each of the m events come from the same sample space Ω, the probabilities of all outcomes (and, therefore, all events) must sum to at most 1.

What if our m events share outcomes? In this case, we can generalize our "probabilistic inclusion-exclusion principle" in exactly the same way we generalized the combinatorial inclusion-exclusion principle. (We won't repeat ourselves here.)

However, if we apply the same idea behind the formula in Theorem 11 to this new scenario where events can share outcomes, then we get a handy result that allows us to obtain a quick-and-dirty upper bound on the probability of some union of events occurring. This result is referred to in some texts as the "union bound", but here we refer to it as **Boole's inequality**—named for the English mathematician George Boole, the very same person for whom Boolean algebra is named.

Theorem 12 (Boole's inequality)**.** *Let $E_1, \ldots, E_m \in \Omega$ be events occurring with probability $\mathbb{P}[E_1], \ldots, \mathbb{P}[E_m]$, respectively. Then*

$$\mathbb{P}\left[\bigcup_{i=1}^{m} E_i\right] \leq \sum_{i=1}^{m} \mathbb{P}[E_i].$$

Proof. Omitted. □

Essentially, Boole's inequality bounds the probability of some union of events occurring by taking the "probabilistic inclusion-exclusion principle" and dropping the terms that account for shared outcomes. Since we are now potentially overcounting, we get an upper bound instead of an exact value.

Remark. Interestingly, there is a middle ground between Boole's inequality and the "probabilistic inclusion-exclusion principle". If, instead of dropping all terms that account for shared outcomes, we only drop some of the terms, then we get a family of inequalities that provide both upper and lower bounds on the probability of some union of events occurring. These are called **Bonferroni inequalities**, named for the Italian mathematician Carlo Emilio Bonferroni.

1.2.2 Disjointness and Independence

If combining events led to "unions of events", then it surely makes sense also to talk about "intersections of events". In this sense, just like the probability of a "union of events" gave the chance of one of the events occurring, the probability of an "intersection of events" gives the chance of all events occurring.

We'll begin with the case where two events cannot occur together. In our discussion on complementary events, we stated that E and $\overline{E}$ cannot share the same outcome simultaneously. Likewise, we got two different generalizations of the "probabilistic inclusion-exclusion principle" depending on whether our set of events shared outcomes.

If two events E and F do not share outcomes (that is, if $E \cap F = \emptyset$), we say that E and F are **disjoint events**. Moreover, since E and F do not share outcomes and, therefore, cannot occur together, we have $\mathbb{P}[E \cap F] = 0$.

Example 13. Let b be a randomly-generated bit string of length four, and define events $E =$ "b begins with a `1`" and $F =$ "b begins with a `0`". Assuming bits are assigned randomly, we have $\mathbb{P}[E] = 1/2$ and $\mathbb{P}[F] = 1/2$, since the outcome of the first bit being a `1` is equally as likely as the outcome of the first bit being a `0`.

What is $\mathbb{P}[E \cap F]$? This is the event where the first bit is simultaneously `1` and `0`; obviously impossible, since the bit string b cannot begin with both a `1` and a `0`. Therefore, E and F are disjoint, so $\mathbb{P}[E \cap F] = 0$.

Remark. Be careful: knowing that $\mathbb{P}[E \cap F] = 0$ does not necessarily imply E and F are disjoint. Consider, as a fun example, the events $E =$ "CISC 203 students do not learn about discrete probability" and $F =$ "CISC 203 students learn about basket-weaving" from the sample space $\Omega =$ "things students learn".

What if, instead of disjoint events, we have two events E and F that can occur together but don't affect one another? In this case, we say that E and F are **independent events**. Independence in this sense does not mean that the two events are mutually exclusive; it just means the probability of one event occurring does not alter the probability of the other event occurring. In other words, the probability of two independent events occurring is just the probabilities of each event occurring taken together, or $\mathbb{P}[E \cap F] = \mathbb{P}[E] \times \mathbb{P}[F]$.

In a set of independent events $\{E_1, \ldots, E_m\}$, we say that each event is **mutually independent**. If we have a weaker relationship where every k-element subset of events is independent, then we say that each event is ***k*-wise independent**.

The classical illustration of independence is given by the "gambler's fallacy". Imagine a gambler walks into a casino and starts playing the tables, only to find themselves on a winning streak. This luck might trick the gambler into thinking that either (i) their luck will continue and they will keep winning, or (ii) their luck will run out and they will lose on the next round. In reality, the gambler has no way of knowing which of the two events will occur, since each round is independent of the others; the gambler's past success does not influence the next round.

Example 14. Let b be a randomly-generated bit string of length four, and define events $E =$ "b begins with a `1`" and $F =$ "b contains an even number of `1`s". Assuming bits are assigned randomly, we have $\mathbb{P}[E] = 1/2$ and $\mathbb{P}[F] = 1/2$, since of the 16 bit strings of length four, eight of them contain an even number of `1`s.

The events E and F are not disjoint, but they are independent. This is because the property of b beginning with a `1` does not exclude b from also containing an even number of `1`s, and vice versa. To see why E and F are independent, consider the set of all bit strings of length four that both begin with `1` and contain an even number of `1`s: $\{\texttt{1001}, \texttt{1010}, \texttt{1100}, \texttt{1111}\}$. The probability of randomly generating one of these bit strings is $4/16 = 1/4$, which is the same as $\mathbb{P}[E] \times \mathbb{P}[F] = 1/2 \times 1/2 = 1/4$.

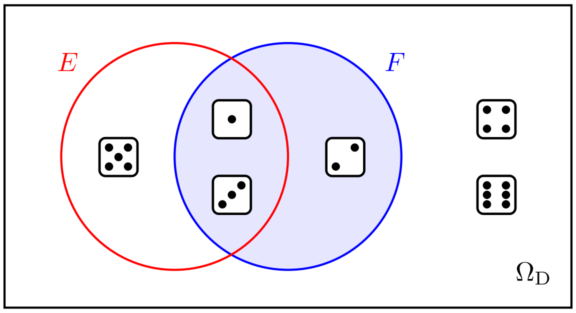
1.3 Conditional Probability

Now that we've discussed events that don't affect one another—independent events—let's move on to discussing events that potentially affect one another. We begin with a couple of simple motivating examples.

The Turing Award is the highest honour in computer science, given to people who have made outstanding research contributions to the field. The odds are very low that, in any given year, an average computer scientist will receive this award. However, suppose that an average computer scientist wrote a paper proving, once and for all, that $\mathsf{P} \neq \mathsf{NP}$. All of a sudden, that particular computer scientist's chance of receiving the Turing Award would skyrocket. The fact that the computer scientist wrote the famous paper influences their chance of receiving the award.

As a more concrete example, consider rolling a single die. As we saw earlier, the sample space for this experiment is $\Omega_{\mathrm{D}} = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$. Let's define two events: E denotes the event "rolled an odd number", and F denotes the event "rolled a number less than 4". A friend rolls the die and tells you that they saw a number less than 4. What is the probability that they also rolled an odd number?

Let's look at this scenario diagrammatically, where E and F are represented by circles and each outcome is placed in the appropriate location.



Immediately, we see that rolling a 4 or a 6 means neither of the events E nor F occur, so both of those outcomes lie outside of each circle. We also see that both circles contain three outcomes, so $\mathbb{P}[E] = 3/6 = 1/2$ and $\mathbb{P}[F] = 3/6 = 1/2$. However, we don't care about all of these outcomes; we only care about the outcomes where our friend rolled an odd number. If we already know that F occurred—that is, we already know our friend rolled a number less than 4—then that narrows our sample space down to {⚀, ⚁, ⚂}. Of these outcomes, the only rolls that would make E occur are ⚀ and ⚂. Therefore, we conclude that the probability of E occurring, given that F occurred, is now $2/3$; a $1/6$ increase from the original probability!

If we have a scenario where the probability of some event E occurring is affected by another event F occurring, then we say that E is **conditional** on F. We also have special terminology and notation for discussing the changed probability of E after F occurs: instead of saying "the probability of E", we say "the probability of E given F", and we write this as $\mathbb{P}[E \mid F]$.

How do we calculate $\mathbb{P}[E \mid F]$? Following our die-roll example, we consider only the outcomes in the event known to have occurred ($\mathbb{P}[F]$, or "rolling a number less than 4") and, from there, we determine the outcomes shared by both events ($\mathbb{P}[E \cap F]$, or "rolling an odd number less than 4"). After this, it's a simple matter of taking a ratio of the number of outcomes shared by both E and F versus the total number of outcomes in F.

Definition 15 (Conditional probability)**.** Let $E, F \in \Omega$ be events occurring with probability $\mathbb{P}[E]$ and $\mathbb{P}[F]$, respectively, with the added condition that $\mathbb{P}[F] > 0$. The conditional probability that E occurs given F is

$$\mathbb{P}[E \mid F] = \frac{\mathbb{P}[E \cap F]}{\mathbb{P}[F]}.$$

Interestingly, we present the formula for $\mathbb{P}[E \mid F]$ as a definition and not as a theorem to be proved. This is because, in Kolmogorov's interpretation of probability—and yes, that is the same Kolmogorov who defined the axioms we saw in Proposition 4—the formula denoting the conditional probability of E given F is intuitively true. Thus, it doesn't make sense to prove the result. (If this leaves you dissatisfied, don't worry; there are other interpretations of probability that handle this result differently.)

Observe that our definition of conditional probability generalizes our method of calculating $\mathbb{P}[E \cap F]$ from independent events to any pair of events. By rearranging the formula given in Definition 15, we obtain a complete set of intersection results:

- $\mathbb{P}[E \cap F] = 0$, if E and F are disjoint;
- $\mathbb{P}[E \cap F] = \mathbb{P}[E] \times \mathbb{P}[F]$, if E and F are independent; and
- $\mathbb{P}[E \cap F] = \mathbb{P}[E \mid F] \times \mathbb{P}[F]$, if E and F are not independent (namely, if E is conditional on F).

Conditional probability brings about an alternate definition of independence: two events E and F are independent if $\mathbb{P}[E] = \mathbb{P}[E \mid F]$ (or, equivalently, if $\mathbb{P}[F] = \mathbb{P}[F \mid E]$). From this, we can rederive our formula for the probability of two independent events occurring using our formula for conditional probability, which reinforces the fact that the independent-events formulation is a special case of conditional probability.

Example 16. Let b be a randomly-generated bit string of length four, and define events E = "b contains two consecutive `1`s" and F = "b both starts and ends with a `1`". Assuming bits are assigned randomly, we have

$\mathbb{P}[E] = 1/2$, since the bit strings containing consecutive `1`s are $\{\texttt{0011}, \texttt{0110}, \texttt{1100}, \texttt{0111}, \texttt{1110}, \texttt{1011}, \texttt{1101}, \texttt{1111}\}$. We also have $\mathbb{P}[F] = 1/4$, since both the first and last bits of b have a $1/2$ probability of being assigned a `1`.

Assume that F occurred. What is $\mathbb{P}[E \mid F]$? Since we know that b both starts and ends with a `1`, we can narrow our sample space to all bit strings of length four that satisfy such a property: $\{\texttt{1001}, \texttt{1011}, \texttt{1101}, \texttt{1111}\}$. We also know that $\mathbb{P}[E \cap F] = 3/16$, since the set of bit strings satisfying both properties is $\{\texttt{1011}, \texttt{1101}, \texttt{1111}\}$. Therefore, by our formula for conditional probability, we get $\mathbb{P}[E \mid F] = (3/16)/(1/4) = 3/4$.

1.4 Bayes' Theorem

When students deal with conditional probabilities for the first time, they often make one crucial mistake: they assume that $\mathbb{P}[E \mid F] = \mathbb{P}[F \mid E]$ for any pair of events E and F. Unfortunately, this symmetry does not always hold, and making such a mistake could lead to confusing calculations or incorrect answers.

Fortunately, in the 18th century, an English minister and statistician named Rev. Thomas Bayes developed a formula that allows us to convert between the values $\mathbb{P}[E \mid F]$ and $\mathbb{P}[F \mid E]$. Bayes never got a chance to share this formula with the world himself, as he died two years before his paper was read at the Royal Society. However, his results were corroborated by Laplace (our old friend from earlier in these notes), who rediscovered the formula on his own, and Bayes' important work is still used to this day.

Theorem 17 (Bayes' theorem)**.** *Let $E, F \in \Omega$ be events occurring with probability $\mathbb{P}[E]$ and $\mathbb{P}[F]$, respectively, with the added condition that $\mathbb{P}[F] > 0$. The conditional probability that E occurs given F is*

$$\mathbb{P}[E \mid F] = \frac{\mathbb{P}[F \mid E]\ \mathbb{P}[E]}{\mathbb{P}[F]}.$$

Proof. By the definition of conditional probability, we know that $\mathbb{P}[E \mid F] = \mathbb{P}[E \cap F] \ / \ \mathbb{P}[F]$. We also know that $\mathbb{P}[F \mid E] = \mathbb{P}[F \cap E] \ / \ \mathbb{P}[E]$. Clearly, $\mathbb{P}[E \cap F] = \mathbb{P}[F \cap E]$.

Rearranging, we get that $\mathbb{P}[E \cap F] = \mathbb{P}[E \mid F] \times \mathbb{P}[F]$ and $\mathbb{P}[E \cap F] = \mathbb{P}[F \mid E] \times \mathbb{P}[E]$, so

$$\mathbb{P}[E \mid F] \times \mathbb{P}[F] = \mathbb{P}[F \mid E] \times \mathbb{P}[E].$$

Divide both sides of this expression by $\mathbb{P}[F]$ to get the desired formula. □

Bayes' theorem comes in particularly handy in the field of computing, especially when dealing with applications of computers that require probabilistic decisions to be made. By this, we mean if the computer is given a set of "known" data, then we want the computer to draw conclusions on new data based (in whole or in part) on what it learned from the given data. These data sets have particular names: training data and testing data, respectively. Historical examples of training computers to make probabilistic decisions usually involved spam filtering or image recognition. Nowadays, the hot topics are artificial intelligence and machine learning, and Bayes' theorem appears frequently in those domains.

Example 18. In an email system, the spam filter determines which emails should be delivered to the recipient and which emails should be delivered to the junk bin. It is important for the spam filter to work correctly; nobody is happy if an email from the boss gets blocked, just like nobody is happy to hear that a Nigerian prince wants to wire-transfer \$500 000 to their bank account.

Suppose we have a spam filter that has been trained on certain "spammy" words, like PASSWORD. Such words appear more frequently in spam emails than in regular emails, so the spam filter marks them as suspicious. Assume that the word PASSWORD appears in 5/6 of all emails and 1/2 of all spam emails used by the filter for training, and that the filter's training email set consisted of $\mathbb{P}[\text{spam}] = 2/3$ and $\mathbb{P}[\text{real}] = 1/3$. Then, the probability that an email containing the word PASSWORD is truly a spam email is

$$\mathbb{P}[\text{spam} \mid \textsc{password}] = \frac{\mathbb{P}[\textsc{password} \mid \text{spam}]\ \mathbb{P}[\text{spam}]}{\mathbb{P}[\textsc{password}]} = \frac{(1/2) \times (2/3)}{(5/6)} = \frac{2}{5}.$$

2 Random Variables

Thus far, when we have been discussing events, we have had to define our events explicitly; for example, when rolling a die, we may want to define the event $E =$ "rolled a 2". This can get tedious if we have many events to define. Imagine having to define separate events for drawing each card from a standard deck of playing cards or, worse, picking a number between 1 and 1 000 000.

Instead of defining our events explicitly, we could define some kind of mathematical object that takes an outcome as an argument and gives us the corresponding probability as its unique result. This sounds familiar... almost like a function. Indeed, what we consider in this section is exactly the same as a function, just labelled with a different name. We call this special function a **random variable**.

Definition 19 (Random variable)**.** A random variable $X : \Omega \to A$ is a function mapping outcomes from a sample space Ω to a set A.

Before we go any further, let's address the question that is sure to appear at some point in this discussion: who decided to call this thing a "random variable"? It's *not* a variable (it's a function) and it's *not* random (the function deterministically maps elements from one set to another set)! The person responsible for this name was the Italian mathematician Francesco Cantelli, who first used the term "random variable"—or "*variabile casuale*" in Italian—in a 1916 paper; the term has stuck around ever since.

Okay, now let's get the discussion back on track. In our definition of a random variable, we see that X maps outcomes in Ω to a set A. This set A can be anything we like, though we often simply take A to be the set of real numbers, $\mathbb{R}$; this allows us to map outcomes to probabilities in the range $[0, 1]$ as we have been doing up to this point. However, we can just as easily take A to be any other set. Consider:

- If A is the set of natural numbers, $\mathbb{N}$, then we can count outcomes.
  - The number of Ⓗ s seen in a sequence of f coin flips.
  - The number of ⚁ s rolled in a sequence of r dice rolls.
- If A is the set of binary digits, $\mathbb{B} = \{0, 1\}$, then we can indicate outcomes.
  - The appearance of a ⚄ in a sequence of r dice rolls.
  - The appearance of a A♠ in a hand of c playing cards.
- If A is the set of Cartesian coordinates, $\mathbb{R}^2$, then we can generate points on a plane.

As you can see, random variables give us a lot more flexibility when dealing with probability applications.

Since X is a function, we typically write $X[\omega]$ when we want to determine the value corresponding to some outcome $\omega \in \Omega$.

Example 20. Consider $X : (\Omega_{\mathrm{D}} \times \Omega_{\mathrm{D}}) \to \mathbb{N}$, where we map pairs of dice rolls to the number of pips common to both dice. Then $X[\{\text{⚃⚁}\}] = 2$, $X[\{\text{⚀⚀}\}] = 1$, $X[\{\text{⚂⚄}\}] = 3$, and so on. (You might notice that X behaves similarly to the minimum function.)

At times, we may wish to determine the probability of X taking a certain value. This is equivalent to us determining the probability of some event occurring. For instance, taking X to be defined as in Example 20, suppose we roll a die and we want to determine the probability that the number of common pips is 4. We would determine the probability of the event $E = \{\omega \in (\Omega_{\mathrm{D}} \times \Omega_{\mathrm{D}}) \mid X[\omega] = 4\}$.

To save space, instead of writing out all the details of the event using set notation, we can overload the probability measure symbol $\mathbb{P}$ in the following way. Suppose again that we roll a die and we want to determine the probability that the number of common pips is 4. We can write this simply as $\mathbb{P}[X = 4]$, where X is again defined as in Example 20.

Putting things all together in our example, we have

$$\mathbb{P}[X = 4] = \mathbb{P}[\{\omega \in (\Omega_{\mathrm{D}} \times \Omega_{\mathrm{D}}) \mid X[\omega] = 4\}] = \mathbb{P}[\{\text{⚃⚄}, \text{⚃⚅}, \text{⚄⚃}, \text{⚅⚃}\}] = \frac{4}{36} = \frac{1}{9}.$$