Probability 1 lecture notes

Clare Wallace

2025-10-02

Table of contents

Welcome to Probability 1

Welcome to Probability 1! These lecture notes contain all the mathematical content you'll need to know to succeed in Probability this year.

If you have questions about any of the content here, try one of the following:

- ask a friend!
- ask me! I like to answer emails, and I am often in my office (MCS3060): you can (and should) pop by to see if I'm around. You can do this during my official office hours (Mondays, 10-12) for a guaranteed speedy response, but you definitely shouldn't wait until then, especially if it's a short or quick question.
- Google it, or try a textbook. There are some good ones on the reading list (see below).

These notes have been developed over the years by several members of the Statistics and Probability groups, including (most recently) Debleena Thacker and Andrew Wade.

Warning

There could still be typos. If you find one, let me know about it and you can have a free bag of Skittles.

How to use these notes

The notes contain all the mathematical content for the course. In lectures, we will start at the beginning and work our way through the whole document, until we reach the end (hopefully, this will happen exactly at the end of term).

Throughout the notes, there are boxes like this one:

Try it out

You can do the "Introductory" exercises on the problem sheet already.

These contain examples you can work through to check your understanding. Wherever possible, I've also worked examples into the text, but there are some places where I want to give you an extra example. These come in purple boxes.

Content that's particularly important for the course is highlighted in red:

Key idea

Probability is cooler than statistics

while advanced material is highlighted in blue:

Advanced content

Probability is *almost surely* cooler than statistics.

You'll also find textbook recommendations, with the relevant sections:

Textbook references

If you want more help with this section, check out:

- Appendix A.1 in (Blitzstein and Hwang 2019);
- Appendix B in (Anderson, Seppäläinen, and Valkó 2018);
- or the Appendix to Chapter 0 in (Stirzaker 2003).

The library has lots of good books on introductory probability, and there are even more available online/to buy. The following four textbooks are a good starting point:

- (Blitzstein and Hwang 2019) covers the material in depth and uses simulation code to illustrate the theory.
- (Anderson, Seppäläinen, and Valkó 2018) covers just about everything in the course at about the right level of detail.
- (Stirzaker 2003) is concise and the most mathematically advanced, and will be useful for students taking 2H probability.
- (DeGroot and Schervish 2013) has a statistical perspective, covering this course as well as a lot of Statistics.

1 Axioms of probability

Goals

1. Understand elementary set theory and how to use it to formulate probabilistic scenarios and to describe the calculus of events.
2. Be familiar with the axioms of probability and their consequences, and how these properties may be deduced from the axioms.

In this chapter, we lay the foundations of probability calculus, and establish the main techniques for practical calculations with probabilities. The mathematical theory of probability is based on *axioms*, like Euclidean geometry. In classical geometry, the fundamental objects posited by the axioms are points and lines; in probability, they are *events* and their *probabilities.* The language and apparatus of set theory is used to express these concepts and to work with them.

There is a lot of ambiguity inherent in probability, because we are often using mathematical approaches to describe real-world scenarios. In some cases, there are several different ways to represent the real-world scenario as a probabilistic model, and the choices we make could affect our conclusions. In others, an unambiguous mathematical setup could have different real-world interpretations, depending on how we view it. Either way, once we have a probabilistic model, the axioms help us to ensure that the maths remains the same.

The axioms and properties of probability we develop in this chapter lay the foundations for all the rest of the theory we will build later in the course.

1.1 Sets

One of the key tools we need in this chapter is a good understanding of *set theory.* You'll see all of this much more formally in Analysis, but in this section we give a quick rundown of the essentials we need for Probability.

In essence, a set is an unordered collection of distinguishable objects; these objects can be numbers, functions, other sets, and so on—any mathematical object can belong to a set.

The formal notation for a set is an opening curly bracket, followed by a list of *elements* that belong to the set, followed by a closing curly bracket. For instance, the set containing the elements 2, 4, and 5 is denoted by

$$\{2, 4, 5\}.$$

Because the ordering of the elements is irrelevant, $\{2, 4, 5\}$ and $\{4, 5, 2\}$ denote the same set.

Definition: empty set

The set with no outcomes is called the *empty set*, and is denoted by $\emptyset$:

$$\emptyset := \{\}.$$

A set is often denoted by a capital letter such as A, B, C, and so on.

Definition: subset

For two sets A and B, we say that A is a *subset* of B, and we write $A \subseteq B$ (or $B \supseteq A$), whenever every element that belongs to A also belongs to B, that is, for all $x \in A$ we have $x \in B$.

For instance, $\{2, 4, 5\} \subseteq \{1, 2, 3, 4, 5\}$. Note that for every set A, we have $A \subseteq A$ and $\emptyset \subseteq A$. We can also use *strict* subsets, when the subset is not equal to the larger set: $\{2, 4, 5\} \subset \{1, 2, 3, 4, 5\}$.

Definition: power set

The set consisting of all subsets of a set A is called the *power set* of A, and is denoted as 2^A:

$$2^A := \{B \colon B \subseteq A\}.$$

For example, the power set of the set $A = \{1, 2, 3\}$ is

$$2^A = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

The notation 2^A alludes to the size of the power set. When A is a finite set, its power set contains $2^{|A|}$ subsets. This can be proved by constructing a bijection from 2^A to ordered $|A|$-tuples of 0s and 1s, where a 1 indicates that the corresponding element of A is in the subset.

Textbook references

If you want more help with this section, check out:

- Appendix A.1 in (Blitzstein and Hwang 2019);
- Appendix B in (Anderson, Seppäläinen, and Valkó 2018);
- or the Appendix to Chapter 0 in (Stirzaker 2003).

1.2 Sample space and events

Definition: scenarios, outcomes and sample space

Whenever we do some probability, it is based on a *scenario* in which there are various *outcomes*. We assume that we know the (set of all) possible outcomes, but we are unsure about which outcome will occur.
A *sample space* is a set of outcomes for this scenario with the property that one (and only one) of these outcomes must occur.
In this course, we will usually denote the sample space by Ω, and a generic outcome by $\omega \in \Omega$.

For instance, suppose we roll a standard six-sided die.

The most obvious sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, but if one was interested only in whether the die was odd or even, or a six or not, one could use $\Omega = \{\text{odd}, \text{even}\}$, or $\Omega = \{\text{not a } 6, 6\}$.

Often, like in the above example, we may enumerate the elements of the sample space Ω in a finite or infinite list $\Omega = \{\omega_1, \omega_2, \ldots\}$, in which case we say the set Ω is *countable* or *discrete*.

A set is said to be countable when its elements can be enumerated in a (possibly infinite) sequence. Every finite set is countable, and so is the set of natural numbers $\mathbb{N} := \{1, 2, 3, \ldots\}$. The set of integers $\mathbb{Z}$ is countable as well. The set of real numbers $\mathbb{R}$ is not countable, and neither is any interval $[a, b]$ when $a < b$.

Definition: countable

A set A is countable if either:

- A is finite, or
- there is a bijection (one-to-one and onto mapping) between A and the set of natural numbers $\mathbb{N}$.

One can prove that the set of rational numbers $\mathbb{Q}$ is countable.

When we perform an experiment we are interested in the occurence, or otherwise, of *events*. An *event* is just a collection of possible outcomes, i.e., a subset of Ω.

Key idea: Definition: events

Associated to our sample space Ω is a collection $\mathcal{F}$ of *events*:

$$A \subseteq \Omega \text{ for every } A \in \mathcal{F}.$$

We say that an event A *occurs* when the outcome that occurs at the end of the scenario is in the set A.

If Ω is discrete, we can always take $\mathcal{F} = 2^\Omega$, so that *every* subset of Ω is an event. If Ω is not discrete, we need to be a little more careful: see Section 1.4 below.

The empty set $\emptyset$ represents the *impossible event*, i.e., it will never occur. The sample space Ω represents the *certain event*, i.e., it will always occur. Most interesting events are somewhere in between.

The representation of an event as a set obviously depends on the choice of sample space Ω for the specific scenario under study, as shown by the following two examples.

Examples

1. For rolling a standard cubic die (with $\Omega = \{1, 2, 3, 4, 5, 6\}$), the event “throw an odd number” is the subset $A = \{1, 3, 5\}$ consisting of three outcomes. If we roll the die and it comes up a 3, then event A has occurred.

2. For the same scenario, but with $\Omega = \{\text{odd}, \text{even}\}$, the event ‘throw an odd number’ is the subset $A = \{\text{odd}\}$ consisting of just one outcome.

Try it out

Suppose we roll three standard six-sided dice and record the outcome of the experiment by an ordered triple (i, j, k) where $i, j, k \in \{1, 2, \dots, 6\}$. What is Ω? How big could $\mathcal{F}$ be?
Answer:
In this case $\Omega = \{1, 2, \dots, 6\}^3$ has $6^3 = 216$ individual outcomes.
The number of events (2^{216}) is enormous! One is the event that the scores on the three dice are the same: $A = \{(1, 1, 1), (2, 2, 2), \dots, (6, 6, 6)\}$.

Textbook references

If you want more help with this section, check out:

- Section 1.2 in (Blitzstein and Hwang 2019);
- Section 1.1 in (Anderson, Seppäläinen, and Valkó 2018);
- Sections 1.1 and 1.2 in (Stirzaker 2003);
- or Section 1.4 in (DeGroot and Schervish 2013).

1.3 Event calculus

Once we've defined our sample space and the set of all possible events, we need to be able to refer to combinations of events. To do so, we use standard notation from set theory.

Definition: complements

For an event $A \in \mathcal{F}$, we define its *complement*, denoted A^c (or sometimes $\bar{A}$) and read "not A", to be

$$A^c := \Omega \backslash A = \{\omega \in \Omega : \omega \notin A\}.$$

Notice that:

- the complement of A^c is A: $(A^c)^c = A$;
- there are no outcomes in *both* A and A^c: $A \cap A^c = \emptyset$;
- and every outcome is in one or the other: $A \cup A^c = \Omega$.

Key idea: event calculus

Given any two events A and B that are associated with the same sample space (i.e. $A \subseteq \Omega$ and $B \subseteq \Omega$ for the same Ω), here are some of the other events we can define, along with how we would read them out:

Notation	We say (as sets)	We say (as events)	Meaning (as events)
$A \cup B$	A union B	A or B	A occurs or B occurs or both A and B occur
$A \cap B$	A intersect B	A and B	A occurs and B occurs
$A^c := \Omega \backslash A$	A complement	not A	A does not occur
$A \backslash B$	A minus B	A but not B	A occurs but B does not
$A \subseteq B$	A is a subset of B	A implies B	if A occurs, then B must occur

(In the final row, “$A \subseteq B$” is not an event but rather a statement about how two events relate to each other. I still wanted to include it because I think it’s helpful)

Try it out

Prove that $A \backslash B = A \cap B^{\mathrm{c}}$.
Answer:
We can do this by working with the events as sets. We have

$$A \backslash B = \{\omega \in \Omega : \omega \in A,\, \omega \notin B\} = \{\omega \in \Omega : \omega \in A\} \cap \{\omega \in \Omega : \omega \notin B\} = A \cap B^{\mathrm{c}}.$$

Key idea: Definition: disjoint

We say that events A and B are *disjoint*, *mutually exclusive*, or *incompatible* if $A \cap B = \emptyset$, i.e., it is impossible for A and B both to occur.

Try it out

Consider the sample space $\Omega := \{1, 2, 3, 4, 5, 6\}$, and the events

$$\begin{aligned} A &:= \{2, 4, 6\}, \\ B &:= \{1, 3, 5\}, \\ C &:= \{1, 2, 3\}. \end{aligned}$$

In other words, A is the event “throw an even number”, B is the event “throw an odd number”, and C is the event “throw at most three”. Use some of the ideas from the table above to combine events A, B, and C in different ways. Are any of your new events disjoint?
Answer:
Some combinations:

$$\begin{aligned} A \cup B &= \Omega && \text{(even or odd)} \\ A \cap B &= \emptyset && \text{(even and odd)} \\ A^{\mathrm{c}} &= B && \text{(not even)} \\ C \backslash A &= \{1, 3\} && \text{(at most 3 but not even)} \\ A \cup C &= \{1, 2, 3, 4, 6\} && \text{(even or at most 3)} \\ A \cap C &= \{2\} && \text{(even and at most 3).} \end{aligned}$$

The events A and B are disjoint as $A \cap B = \emptyset$. We have also created two disjoint events: $C \backslash A$ and $A \cap C$. Think about why these two events would always be disjoint, however we define A and C.

Try it out

Toss a coin twice and denote the sample space by $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$. Consider the events

$$\begin{aligned}
A &:= \{\mathrm{HH}, \mathrm{HT}\} && \text{(first toss H)} \\
B &:= \{\mathrm{HT}, \mathrm{TT}\} && \text{(second toss T)} \\
C &:= \{\mathrm{HH}\} && \text{(both H).}
\end{aligned}$$

How do these events relate to each other?
Answer:
Some things you might notice:

- $C \subseteq A$, i.e., if C occurs then A must occur;
- $A \cup B = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TT}\}$ is the event that either the first toss is H, the second toss is T, or both;
- $A \cap B = \{\mathrm{HT}\}$;
- $A^{\mathrm{c}} = \{\mathrm{TH}, \mathrm{TT}\}$;
- $B \cap C = \emptyset$.

Try it out

Draw a card from a standard deck of 52 playing cards. Take Ω to consist of each of the 52 possible draws: $\Omega = \{\mathrm{A}\clubsuit, \mathrm{A}\diamondsuit, \dots, \mathrm{K}\heartsuit, \mathrm{K}\spadesuit\}$. Events in $\mathcal{F} = 2^{\Omega}$ include

$$\begin{aligned}
E &= \{\text{eight}\} = \{8\spadesuit, 8\heartsuit, 8\diamondsuit, 8\clubsuit\}, \\
S &= \{\text{spade}\} = \{A\spadesuit, 2\spadesuit, \dots, K\spadesuit\},
\end{aligned}$$

and we can combine them to form other events, such as

$$\begin{aligned}
E \cap S &= \{8\spadesuit\}, \\
E \backslash S &= \{8\heartsuit, 8\diamondsuit, 8\clubsuit\}, \\
S \backslash E &= \{A\spadesuit, 2\spadesuit, \dots, 7\spadesuit, 9\spadesuit, \dots, K\spadesuit\}.
\end{aligned}$$

As with sums ($\sum$) and products (Π) of multiple numbers, we also have shorthands for unions and intersections of multiple sets:

$$\bigcup_{i=1}^{n} A_i := A_1 \cup A_2 \cup \cdots \cup A_n$$

is the event that at least one of $A_1, A_2, \dots A_n$ occurs (or the set of all $\omega \in \Omega$ which are contained in at least one of the A_is), and

$$\bigcap_{i=1}^{n} A_i := A_1 \cap A_2 \cap \cdots \cap A_n$$

is the event that *all* of $A_1, A_2, \dots A_n$ occur (or the set of all $\omega \in \Omega$ which are in every A_i).

Occasionally, we will also need to take infinite unions and intersections over sequences of sets:

$$\begin{aligned}
\bigcup_{i=1}^{\infty} A_i &:= A_1 \cup A_2 \cup A_3 \cup \dots \\
\bigcap_{i=1}^{\infty} A_i &:= A_1 \cap A_2 \cap A_3 \cap \dots .
\end{aligned}$$

We will also sometimes need *De Morgan's Laws*: for a (possibly infinite) collection of events A_i,

a. The complement of the union is the intersection of the complements: $\left(\bigcup_i A_i\right)^{\mathrm{c}} = \bigcap_i A_i^{\mathrm{c}}$, and
b. The complement of the intersection is the union of the complements: $\left(\bigcap_i A_i\right)^{\mathrm{c}} = \bigcup_i A_i^{\mathrm{c}}$.

These could be more intuitive than they appear: the negation of "some of these things happened" is "none of these things happened", and the negation of "all of these things happened" is "some of these things did not happen".

It is often useful to visualize the sample space in a *Venn diagram*. Then events such as A are subsets of the sample space. It is a helpful analogy to imagine the probability of an event as the *area* in the Venn diagram.

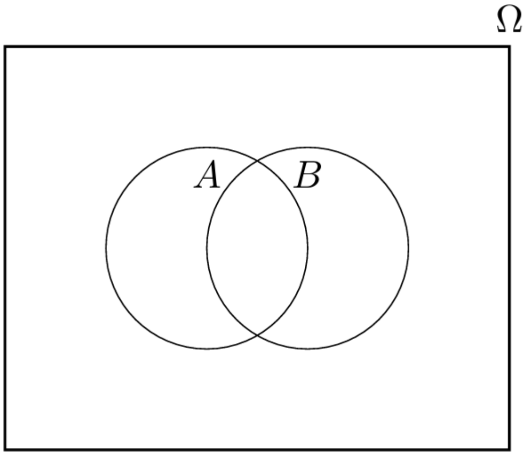


Figure 1.1: Venn diagram

Advanced content

This analogy is more apt than it first appears, since the mathematical foundations of rigorous probability theory are built on *measure theory*, which is the same theory that gives rigorous foundation to the concepts of length, area, and volume.

Textbook references

If you want more help with this section, check out:

- Section 1.2 in (Blitzstein and Hwang 2019);
- Section 1.2 in (Stirzaker 2003);
- or Section 1.4 in (DeGroot and Schervish 2013).

1.4 Sigma-algebras

In the last section we described some of the ways in which events can be combined. Now we can set out the rules for our collection of events, $\mathscr{F}$, to ensure that it's possible to *use* these different combinations.

We said that in the case where Ω is discrete, one can take $\mathscr{F} = 2^{\Omega}$.

In general, if Ω is uncountable, it is too much to demand that probabilities should be defined on *all* subsets of Ω. The reason why this is a problem goes beyond the scope of this course (see the Bibliographical notes at the end of this chapter for references), but the essence is that for uncountable sample spaces, such as $\Omega = [0, 1]$, there exist subsets of Ω that cannot be assigned a probability in a way that is consistent. The construction of such *non-measurable sets* is also the basis of the famous *Banach–Tarski paradox*.

Uncountable Ω are unavoidable: we will see an infinite coin-tossing space at the end of section Section 1.6, and other examples occur whenever we have an experiment whose outcome is modelled by a continuous distribution such as the normal distribution (more on this later).

The upshot of all this is that we can, in general, only demand that probabilities are defined for all events in some collection $\mathcal{F}$ of subsets of Ω (i.e., for some $\mathcal{F} \subseteq 2^{\Omega}$). What properties should the collection $\mathcal{F}$ of events possess? Consideration of the set operations in the previous section suggests the following definition.

Definition: -algebra

A collection $\mathcal{F}$ of subsets of Ω is called a *σ-algebra* over Ω if it satisfies the following properties.
(S1) $\Omega \in \mathcal{F}$;
(S2) $A \in \mathcal{F}$ implies that $A^{\mathrm{c}} \in \mathcal{F}$;
(S3) if $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Property **S2** says that $\mathcal{F}$ is *closed under complementation*, while **S3** says that $\mathcal{F}$ is *closed under countable unions*.

We can combine **S1** and **S2** to see that we must have $\emptyset \in \mathcal{F}$. Also note that, we can get to a finite-union version of **S3** by taking $A_{n+1} = A_{n+2} = \cdots = \emptyset$: so $\mathcal{F}$ is also closed under finite unions.

Examples

1. The power set 2^{Ω} is a σ-algebra over Ω, and in fact it is the biggest possible σ-algebra over Ω. As described above, for uncountable Ω the set 2^{Ω} may be too unwieldy, in which case we would consider a smaller σ-algebra.

2. The *trivial σ-algebra* $\{\emptyset, \Omega\}$ is the smallest possible σ-algebra over Ω. It's very nicely behaved (just two elements!) but it carries no information about the outcome of the experiment.

Try it out

Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ for the experiment of rolling a fair die. The choice of σ-algebra determines the *resolution at which we observe the experiment*, and may depend on exactly what we are interested in:

- $\mathcal{F}_0 = \{\emptyset, \Omega\}$ (carries no information);
- $\mathcal{F}_1 = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$ (if we only care whether the score is odd or even);
- $\mathcal{F}_2 = 2^{\Omega}$ (if we are interested in the exact score).

Note the inclusions $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2$.
Let us check that $\mathcal{F}_1$ is indeed a σ-algebra.
S1 is immediate: we can see it from the definition of $\mathcal{F}_1$.
For **S2**, we need that every $A \in \mathcal{F}_1$ is accompanied by its complement A^{c}; we see that this is the case.

Since $\mathcal{F}_1$ is a finite set it suffices to check **S3** for finite unions. In other words, it is enough to check that if $A, B \in \mathcal{F}_1$, then $A \cup B \in \mathcal{F}_1$ too. Since there are only two sets, this is also quick: we see that it is the case.

Textbook references

If you want more help with this section, check out:

- Section 1.2 in (Stirzaker 2003).

1.5 The axioms of probability

Key idea: Definition: probability

A *probability* $\mathbb{P}$ on a sample space Ω with collection $\mathcal{F}$ of events is a function mapping every event $A \in \mathcal{F}$ to a real number $\mathbb{P}(A)$, obeying the following axioms:
(A1) $\mathbb{P}(A) \geq 0$ for every $A \in \mathcal{F}$;
(A2) $\mathbb{P}(\Omega) = 1$; and
(A3) if A and B are *disjoint* events (i.e. if $A, B \in \mathcal{F}$ have $A \cap B = \emptyset$) then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

We call the number $\mathbb{P}(A)$ *the probability of* A.

We will see shortly that a consequence of these axioms is that the probabilities $\mathbb{P}(A)$ must lie between 0 and 1: $0 \leq \mathbb{P}(A) \leq 1$.

We can upgrade **(A3)** to a slightly more technical version:

(A4) For any infinite sequence $A_1, A_2, \ldots$ of pairwise disjoint events (so $A_i \cap A_j = \emptyset$ for all $i \neq j$),

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Key idea: A small request

If you only take one thing away from this course, please let it be this:
Probabilities are **numbers** and events are **sets**.
We can add up numbers (but not sets) and we can take unions and intersections of sets (but not numbers).

For the axioms to make sense, we can't just use any old event set $\mathcal{F}$. For one thing, we need $\Omega \in \mathcal{F}$; in fact all the events in **(A1-4)** need to be in $\mathcal{F}$. Our definition of a σ-algebra from the previous section gives us exactly the event set we need.

Key idea: Definition: probability space

If Ω is a set and $\mathcal{F}$ is a σ-algebra of subsets of Ω, and if $\mathbb{P}$ satisfies **(A1–4)** for events in $\mathcal{F}$, then the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space.*

Try it out

Consider a finite sample space $\Omega = \{\omega_1, \dots, \omega_m\}$ of size $|\Omega| = m$. Then we can define a valid probability $\mathbb{P}$ by taking any numbers $p_1, \dots, p_m$ with $p_i \geq 0$ for all i and $\sum_{i=1}^m p_i = 1$ and declaring that for any event A,

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i.$$

This satisfies the axioms **(A1–4)** (don't just believe me - check them for yourself).
By considering the event $A = \{\omega_i\}$, we see that $p_i = \mathbb{P}(\omega_i)$ is the probability of the elementary outcome ω_i.
In the simplest setting, we might assume that all the outcomes are *equally likely*, that is, $p_i = 1/m$ for all i. Note that in this case probability reduces to counting, since

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} \frac{1}{m} = \frac{|A|}{|\Omega|}.$$

As a concrete example, for tossing a *fair* die we would have $\Omega = \{1, 2, \dots, 6\}$, and $\mathbb{P}(A) = |A|/6$ so, for example,

$$\mathbb{P}(\text{score is odd}) = \mathbb{P}(\{1, 3, 5\}) = \frac{3}{6} = \frac{1}{2}.$$

We examine this setting in detail in Chapter 2.

Try it out

Consider a countably infinite sample space $\Omega = \{\omega_1, \omega_2, \dots\}$. Then we can define a valid probability $\mathbb{P}$ by taking any numbers $p_1, p_2, \dots$ with $p_i \geq 0$ for all i and $\sum_{i=1}^\infty p_i = 1$ and declaring that for any event A,

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i.$$

This definition of a probability satisfies all of the axioms **(A1-A4)**.

For this course, we will usually assume that the probability distribution is given (and satisfies the axioms), without worrying too much about how the important practical task of finding the probabilities was carried out.

Textbook references

If you want more help with this section, check out:

- Section 1.6 in (Blitzstein and Hwang 2019);
- Section 1.1 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 1.3 in (Stirzaker 2003).

1.6 Consequences of the axioms

A host of useful results can be derived from A1–4.

Key idea: Consequences of the axioms

(C1) For any two events A and B,

$$\mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

(C2) For any event A, $\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A)$.
(C3) The probability of $\emptyset$ is $\mathbb{P}(\emptyset) = 0$.
(C4) For any event A, $\mathbb{P}(A) \leq 1$.
(C5) If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$ ("monotonicity").
(C6) For any two events A and B,

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

(C7) If $A_1, A_2, \dots, A_k$ are pairwise disjoint (so $A_i \cap A_j = \emptyset$ if $i \neq j$) then

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \mathbb{P}(A_i).$$

(This property is called "finite additivity" in textbooks.)
(C8) For any events $A_1, A_2, \ldots$, (these need not be pairwise disjoint),

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

(This one is sometimes referred to as "Boole's inequality.")
(C9) If $A_1 \subseteq A_2 \subseteq \cdots$ is an *increasing sequence* of events, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

If $A_1 \supseteq A_2 \supseteq \cdots$ is a *decreasing sequence* of events, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(This property is a bit more sophisticated than the previous ones. It establishes the "continuity of probability along monotone limits:" we can take limits, as long as the events in question form a monotone sequence. It will be really important in Probability II.)

Just one more consequence to go! Before we get there, we need the following simple but extremely useful idea: partitions.

Key idea: Definition: partition

We say that the events $E_1, E_2, \dots, E_k \in \mathcal{F}$ form a (finite) *partition* of the sample space Ω if:

i. they all have positive probability, i.e., $\mathbb{P}(E_i) > 0$ for all i;

ii. they are *pairwise disjoint*, i.e., $E_i \cap E_j = \emptyset$ whenever $i \neq j$; and
iii. their union is the whole sample space: $\cup_{i=1}^{k} E_i = \Omega$.

The definition extends to countably infinite partitions. We say that $E_1, E_2, ... \in \mathcal{F}$ form an infinite partition of Ω if:

i. $\mathbb{P}(E_i) > 0$ for all i;
ii. $E_i \cap E_j = \emptyset$ whenever $i \neq j$; and
iii. $\cup_{i=1}^{\infty} E_i = \Omega$.

For example, consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. Some partitions are:

$$\begin{aligned}
\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\} \\
\{1, 2\}, \{3, 4\}, \{5, 6\} \\
\{1, 2, 3\}, \{4, 5, 6\} \\
\{1\}, \{2, 3\}, \{4, 5, 6\} \\
\{1, 2, 3, 4, 5, 6\}
\end{aligned}$$

and so on.

Key idea

(C10) If E_1, E_2, ..., E_k form a partition then

$$\sum_{i=1}^{k} \mathbb{P}(E_i) = 1.$$

These consequences have an enormous effect on the way we work with probability. In particular, it turns out that we can solve most problems without ever having to explicitly write down the outcomes in our sample space, as in the next example. In fact, some people do probability without even defining a sample space.

Try it out

Jimmy's die has the numbers 2,2,2,2,5,5. Your die has numbers 1,1,4,4,4,4. You both throw and the highest number wins. Assuming all outcomes are equally likely, what is the probability that Jimmy wins?
Answer:
The event, J, that Jimmy wins happens if either Jimmy throws a 5 (call this event F), or if you throw a 1 (call this event A). Therefore $J = A \cup F$ and by C6,

$$\mathbb{P}(J) = \mathbb{P}(A) + \mathbb{P}(F) - \mathbb{P}(A \cap F).$$

As $\mathbb{P}(F) = 1/3$, $\mathbb{P}(A) = 1/3$ and $\mathbb{P}(A \cap F) = 4/36 = 1/9$ (by counting equally likely outcomes) we have

$$\mathbb{P}(J) = 1/3 + 1/3 - 1/9 = 5/9.$$

Finite sample spaces are a great way to build up our intuition for probability calculations. However, it is surprisingly easy to end up in a situation where things start to get complicated.

Try it out

What is the probability that, in an indefinitely long sequence of tosses of a fair coin, we will eventually see heads?

Answer:

The sample space Ω is infinite and consists of all sequences $\omega = (\omega_1, \omega_2, ...)$ with $\omega_i \in \{\mathrm{H}, \mathrm{T}\}$.

What is $\mathbb{P}$? Well, it certainly would be desirable that if we restrict to just a finite sequence of n tosses, then each of the 2^n possible outcomes (sequences) should be equally likely. It is a special case of a general theorem that such a $\mathbb{P}$ exists and is unique.

Now, let $A = \{\mathrm{H}\text{ occurs}\}$. Then the only way A can *not* occur is if there are *no* heads, i.e., $A^{\mathrm{c}} = \{(\mathrm{TTT}\cdots)\}$. This is a single sequence, out of infinitely many, and it is intuitively clear that it should have probability 0. To prove this, it is enough to observe that $A^{\mathrm{c}} \subseteq \{\text{first } n \text{ tosses T}\}$, so by monotonicity **(C5)**,

$$\mathbb{P}(A^{\mathrm{c}}) \leq \mathbb{P}(\{\text{first } n \text{ tosses are T}\}) \leq 2^{-n}.$$

But this is true for any n, so we must have $\mathbb{P}(A^{\mathrm{c}}) = 0$.

Another way to see this is as follows. Consider events defined for $n = 1, 2, ...$ by

$$A_n = \{\text{first H occurs on toss } n\} = \{\omega : \omega_k = \mathrm{T}, \text{ for all } k < n, \, \omega_n = \mathrm{H}\}.$$

This means that A_1 consists of sequences H⋯, A_2 consists of sequences TH⋯, and so on.

Now the event we are interested in is $A = \cup_{n=1}^{\infty} A_n$. So, by **(A4)**,

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} 2^{-n} = 1.$$

Note that a similar argument works if the coin is biased with probability $p \in (0, 1)$ of heads.

Advanced content

In fact, the sequence space Ω in the previous example is not even countable! To see this, a sequence $(d_1, d_2, ...)$ with each $d_i \in \{0, 1\}$ is called a *dyadic expansion* of $x \in [0, 1]$ if $x = \sum_{i=1}^{\infty} 2^{-i} d_i$. For example, $(1, 0, 0, ...)$ is a dyadic expansion of $1/2$, $(1, 1, 0, 0, ...)$ is $3/4$, and so on. The map between x and $(d_1, d_2, ...)$ is almost a bijection. It is not a bijection because of possible non-uniqueness of the dyadic expansion: e.g. $(0, 1, 1, 1, ...)$ is another expansion of $1/2$. It turns out that this problem only occurs for rational x, and can be circumvented. Thus we have (essentially) a bijection between $[0, 1]$ and the space of infinite sequences of 0s and 1s, which is another name for our coin tossing space Ω. This shows that Ω is uncountable.

It is remarkable that the probability $\mathbb{P}$ on infinite sequences of coin tosses turns out to correspond (under the bijection by dyadic expansion) to nothing other than the *uniform* distribution on $[0, 1]$, that is the measure defined by lengths of intervals. This is the famous *Lebesgue* measure.

Textbook references

If you want more help with this section, check out:

- Section 1.6 in (Blitzstein and Hwang 2019);
- Section 1.4 in (Anderson, Seppäläinen, and Valkó 2018);

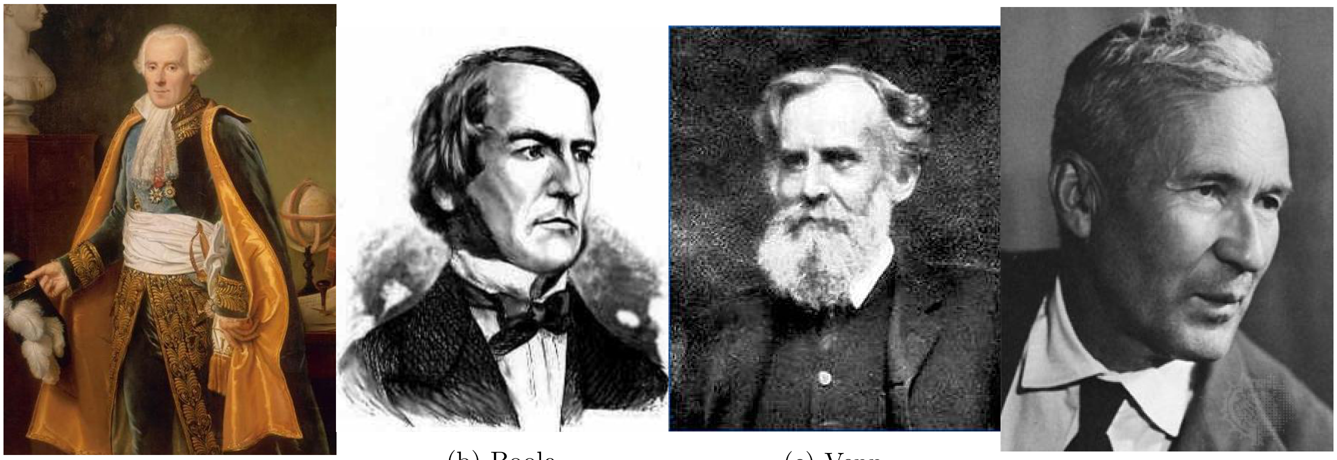
- or Sections 1.4 and 1.5 in (Stirzaker 2003).

1.7 Historical context

Sets are important not only for probability theory, but for all of mathematics. In fact, all of standard mathematics can be formulated in terms of set theory, under the assumption that sets satisfy the ZFC axioms; see for instance this Wikipedia page.

The foundations of probability have a long and interesting history (Hacking 2006; Todhunter 2014). The classical theory owes much to Pierre-Simon Laplace (1749–1827): see (Laplace 1825). However, a rigorous mathematical foundation for the theory was lacking, and was posed as part of one of David Hilbert's (1862–1943) famous list of problems in 1900 (the 5th problem). After important work by Henri Lebesgue (1875–1941) and 'Emile Borel (1871–1956), it was Andrey Kolmogorov who succeeded in 1933 in providing the axioms that we use today (see the 1950 edition of his book (Kolmogorov 1950)). This approach declares that probabilities are *measures*.

A measure μ can be defined on any set Ω with a σ-algebra of subsets $\mathscr{F}$, and the defining axioms are versions of A1 and A4. The special property of a probability measure is just that $\mu(\Omega) = 1$. Measure theory is the theory that gives mathematical foundation to the concepts of length, area, and volume. For example, on $\mathbb{R}$ the unique measure that has $\mu(a,b) = b - a$ for intervals (a,b) is the *Lebesgue measure*.



(a) Laplace

(b) Boole

(c) Venn

(d) Kolmogorov

Figure 1.2: Laplace, Boole, Venn, and Kolmogorov

George Boole (1815–1864) and John Venn (1834–1923) both wrote books concerned with probability theory (Boole 1854), (Venn 1888); both were working before the formulation of Kolmogorov's axioms.

As mentioned in Section 1.4, it is necessary in the general theory of probability to restricting events to some σ-algebra. The reason for this is that in standard ZFC set theory, when Ω is uncountable (such as $\Omega = [0,1]$ the unit interval), it follows from an argument by Vitali (1905) that many natural probability assessments, such as the continuous uniform distribution, cannot be modelled by a probability defined on *all* subsets of Ω satisfying A1–4: see for instance Chapter 1 of (Rosenthal 2007). In the case where Ω is countable, one can always define $\mathbb{P}$ on the whole of 2^{Ω}. In the case where Ω is uncountable, we usually do not explicitly mention Ω at all (when we work with continuous random variables, for example).

The formulation of the infinite coin-tossing experiment in Section 1.6 leads to the connection between coin tossing and the Lebesgue measure, as first described by Hugo Steinhaus in a 1923 paper.

An alternative approach to probability theory is to do away with axiom A4, in which case some of these technical issues can be avoided, at the expense of certain pathologies; however, in the standard approach to modern probability, based on *measure theory*, A4 is a central part of the theory.

2 Equally likely outcomes and counting principles

Goals

1. Understand the equally likely outcomes model of classical probability.
2. Know counting principles, and when and how to apply them on specific problems.

In Chapter 1 we have seen the abstract formulation of probability theory; next we turn to the question of how the probabilities themselves may be assigned.

The most basic scenario occurs when our experiment has a finite number of possible outcomes which we deem to be *equally likely*.

Such situations rarely—not to say never—occur in practice, but serve as good models in extremely controlled environments such as in gambling or games of chance. However, this situation (which will essentially come down to counting) gives us a good initial setting in which to obtain some very useful insights into the nature and calculus of probability.

2.1 Classical probability

Suppose that we have a finite sample space Ω. Since Ω is finite, we can list it as a collection of $m = |\Omega|$ possible outcomes:

$$\Omega = \{\omega_1, \dots, \omega_m\}.$$

In the equally likely outcomes model (also sometimes known as classical probability) we suppose that each outcome has the same probability:

$$\mathbb{P}(\omega) = \frac{1}{|\Omega|} \text{ for each } \omega \in \Omega,$$

or, in the notation above, $\mathbb{P}(\omega_i) = 1/m$ for each i.

The axioms of probability then allow us to determine the probability of any event $A \subseteq \Omega$: by **C7**,

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega) = \frac{|A|}{|\Omega|} \text{ for any event } A \subseteq \Omega.$$

This is a particular case of the discrete sample space discussed in Chapter 1.

Definition: Equally likely outcomes

Consider a scenario with m equally likely outcomes enumerated as $\Omega = \{\omega_1, \dots, \omega_m\}$. In the equally likely outcomes model, the probability of an event $A \subseteq \Omega$ is declared to be

$$\mathbb{P}(A) := \frac{|A|}{|\Omega|}.$$

Using this definition, we meet all of the axioms **(A1–A4)** (checking each of them comes down to what we know about counting). Remember that in the case of a finite state space, we always have the option to take $\mathcal{F} = 2^\Omega$ as our σ-algebra.

Examples

1. Draw a card at random from a well-shuffled pack, so that each of the 52 cards is equally likely to be chosen. Typical events are that the card is a spade (a set of 13 outcomes), the card is a queen (a set of 4 outcomes), the card is the queen of spades (a set of a single outcome). In the equally likely outcomes model, the probability of drawing the queen of spades (or any other specified card) is 1/52, the probability of drawing a spade is 13/52, and the probability of drawing a queen is 4/52.

2. Flip a coin and see whether it falls heads or tails, each assumed equally likely; then 'heads' or 'tails' each has probability 1/2.

3. Roll a fair cubic die to get a number from 1 to 6. Here the word 'fair' is used to mean each outcome is equally likely. Then $\Omega = \{1, \dots, 6\}$ and $\mathbb{P}(A) = |A|/6$. For example, if $A_1 = \{2\}$ (the score is 2) we get $\mathbb{P}(A_1) = 1/6$, while if $A_2 = \{1, 3, 5\}$ (the score is odd) we get $\mathbb{P}(A_2) = 3/6 = 1/2$.

4. If we roll a pair of fair dice then outcomes are pairs (i, j) so there are 36 possible outcomes. If we assume that the outcomes are equally likely, then the probability of getting a pair of 6's is 1/36, for example.

The classical interpretation of probability is the most straightforward approach we can take, just as counting can be seen as "basic" mathematics. It is a good place to start and there are many important situations where intuitively it seems reasonable to say that each outcome of a particular collection is equally likely.

To extend the theory or apply it in practice we have to address situations where there are no candidates for equally likely outcomes or where there are infinitely many possible outcomes and work out how to find probabilities to put into calculations that give useful predictions. We will come back to some of these issues later; but bear in mind that however we come up with our probability model, the same system of axioms that we saw in Chapter 1 applies.

Textbook references

If you want more help with this section, check out:

- Section 1.3 in (Blitzstein and Hwang 2019);
- or Section 1.2 in (Anderson, Seppäläinen, and Valkó 2018).

2.2 Counting principles

Given a finite sample space and assuming that outcomes are equally likely, to determine probabilities of certain events comes down to counting.

For example, in drawing a poker hand of five cards from a well-shuffled deck of 52 cards, the probability of having a 'full house' (meaning two cards of one denomination and three of another, e.g., two Kings and three 7s) is given by the number of hands that are full houses divided by the total number of hands (each hand being equally likely).

These counting problems need careful thought, and we will describe some counting principles for some of the most common situations. There is some common ground with the *Discrete Maths* course; here we have a slightly different emphasis.

2.2.1 The multiplication principle

Counting principle: Multiplication

Suppose that we must make k choices in succession where there are:

- m_1 possibilities for the first choice,
- m_2 possibilities for the second choice,
- $\vdots$
- m_k possibilities for the kth choice,

and the number of possibilities at each stage does not depend on the outcomes of any previous choices. The total number of distinct possible selections is

$$m_1 \times m_2 \times m_3 \times \cdots \times m_k = \prod_{i=1}^{k} m_i.$$

For instance, in a standard deck of playing cards, each card has a *denomination* and a *suit*. There are 13 possible denominations: A(ce), 2, 3, ..., 10, J(ack), Q(ueen), K(ing). There are 4 possible suits: ♡ (heart), ♢ (diamond), ♣ (club), ♠ (spade). Because all combinations of denomination and suit are allowed, the multiplication principle applies: there are $13 \times 4 = 52$ cards in a standard deck.

We will see many applications of counting to dealing cards from a well-shuffled deck. Counting the possibilities is affected by (i) whether the *order* of dealing is important, and (ii) how we *distinguish* the cards: e.g. we may only be interested in their colour (so all red cards are the same) or their suit or their denomination.

Examples

1. A hotel serves 3 choices of breakfast, 4 choices of lunch and 5 choices of dinner so a guest selects from $3 \times 4 \times 5$ different combinations of the three meals (assuming we opt to have all three).

2. A coffee bar has 5 different papers to choose from, 19 types of coffee and 7 different snacks. This means there are $6 \times 20 \times 8 = 960$ distinct selections of coffee, snack and paper. Of these 5 involve no coffee or snack (which the staff may object to) plus one has no coffee, snack or paper!

3. PINs are made up of 4 digits (0–9) with the exceptions that (i) they cannot be four repetitions of a single digit; (ii) they cannot form increasing or decreasing consecutive sequences, e.g. 3456 and 8765 are excluded. How many possible four-digit PINs are there?

   Ignoring restrictions there are $10^4 = 10,000$ distinct PINs. There are 10 PINs with the same digit repeated, namely 0000, 1111, ..., 9999. Increasing sequences start with $0, 1, 2, \ldots, 6$ and decreasing sequences start with $9, 8, \dots, 3$, so there are seven options for each. This leaves $10,000 - 24 = 9,976$ permitted PINs.

All of the following counting principles are effectively consequences of the multiplication principle.

2.2.2 Order matters; objects are distinct

First, we look at ordered choices of distinct objects. In this case, we distinguish between *selection with replacement*, where the same object can be selected multiple times, and *selection without replacement*, where each object can only be selected at most once.

Counting principle: Selection with replacement for ordered choices

Suppose that we have a collection of m distinct objects and we select r of them with replacement. The number of different ordered lists (ordered r-tuples) is

$$\underbrace{m \times \cdots \times m}_{r \text{ times}} = m^r.$$

Counting principle: Selection without replacement for ordered choices

Suppose that we have a collection of m distinct objects and we select $r \leq m$ of them without replacement. The number of different ordered lists (ordered r-tuples) is

$$(m)_r := \underbrace{m \times (m-1) \times (m-2) \times \cdots \times (m-r+1)}_{r \text{ terms}} = \frac{m!}{(m-r)!}.$$

The *falling factorial* notation $(m)_r$ (sometimes also denoted $m^{\underline{r}}$) is simply a convenient way to write $\frac{m!}{(m-r)!}$. In the special case where $r = m$ we set $0! = 1$ and then $(m)_m = m!$ is the number of permutations of the m objects. If m is large, and r is much smaller than m, then $(m)_r \approx m^r$.

Example

The number of ways we can deal out four cards in order from a pack of cards is $(52)_4$ and the number of ways we can arrange the four aces in order is $4!$ so the probability of finding the four aces on top of a well-shuffled deck is

$$\frac{4!}{(52)_4} = \frac{4 \times 3 \times 2 \times 1}{52 \times 51 \times 50 \times 49}.$$

This probability is approximately 3.7×10^{-6} or about 1 in $270,000$.

Try it out

There are $n < 365$ people in a room. Let B be the event that (at least) two of them have the same birthday. (We ignore leap years.)
What is $\mathbb{P}(B)$? How big must n be so that $\mathbb{P}(B) > 1/2$?
Answer:

Here the equally likely outcomes are the ordered length-n lists of possible birthdays:

$$(\text{person 1's birthday}, \text{person 2's birthday}, \dots, \text{person } n\text{'s birthday}).$$

The number of possible outcomes is

$$365 \times 365 \times \cdots \times 365 = 365^n.$$

This is the denominator in our probability.
For the numerator, we must work out how many outcomes are in B. In fact, it is easier to count outcomes in B^{c}, where everyone has a different birthday. There are

$$365 \times 364 \times \cdots \times (365 - n + 1) = (365)_n$$

of these. So

$$\mathbb{P}(B) = 1 - \mathbb{P}(B^{\mathrm{c}}) = 1 - \frac{(365)_n}{365^n}.$$

It turns out that $\mathbb{P}(B) \approx 1/2$ for $n = 23$.

Advanced content

Here's one way to get this. Note that

$$\frac{(365)_n}{365^n} = 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{n-1}{365}\right).$$

Now $1 - x \leq e^{-x}$, and in fact the inequality is very close to equality for $x = 1/365$, being close to zero. In any case,

$$\begin{aligned} \frac{(365)_n}{365^n} &\leq e^{-x}\, e^{-2x}\, e^{-3x} \cdots e^{-(n-1)x} \\ &= \exp\{-(1 + 2 + \cdots + n - 1)x\} \\ &= \exp\left\{-\frac{(n-1)n}{2 \times 365}\right\}. \end{aligned}$$

2.2.3 Order doesn't matter; objects are distinct

In this section, we move on to think about the scenario where the order in which objects are selected *doesn't* matter. This can arise in situations such as dealing a hand of cards, or separating a class into two teams.

Counting principle: Selection without replacement for unordered choices

Suppose that we have a collection of m distinct objects and we select a subset of $r \leq m$ of them without replacement. The number of distinct subsets of size r is

$$\binom{m}{r} := \frac{(m)_r}{r!} = \frac{m!}{r!\,(m-r)!}.$$

To see this, first count the number of distinct ordered lists of r objects—this is $(m)_r$. Each unordered subset has been counted $(r)_r = r!$ times as this is the number of distinct ways of arranging r different objects. Therefore the $(m)_r$ ordered selections can be grouped into collections of size $r!$, each representing a particular subset, and the result follows by dividing.

The expression $\binom{m}{r}$ is the *binomial coefficient* for choosing r objects from m and is often called 'm-choose-r'. Note that

$$\binom{m}{r} = \binom{m}{m-r}$$

as we can choose to take r objects from m in exactly the same number of ways that we can choose to leave behind r objects i.e., take $m - r$ objects.

Try it out

What is the probability of finding no aces in a four-card hand dealt from a well-shuffled deck?
Answer:
Let's answer this by treating hands as unordered selections. Then there are

$$\binom{52}{4} = \frac{52 \times 51 \times 50 \times 49}{4 \times 3 \times 2 \times 1} = 270,725$$

distinct unordered hands of four cards. The number of these with no aces is

$$\binom{48}{4} = \frac{48 \times 47 \times 46 \times 45}{4 \times 3 \times 2 \times 1} = 194,580,$$

and so the probability of finding no aces in a four card hand is

$$\binom{48}{4} \Big/ \binom{52}{4} = \frac{48 \times 47 \times 46 \times 45}{52 \times 51 \times 50 \times 49} \approx 0.7187.$$

Alternatively, we could answer this by treating the hands as *ordered* selections (the order corresponding to the order of the deal, say). Of course, this will give different numerator and denominator in our calculation, but the final answer must be the same! As ordered selections, there are

$$(52)_4 = 52 \times 51 \times 50 \times 49$$

distinct hands. The number of these with no ace is

$$(48)_4 = 48 \times 47 \times 46 \times 45.$$

Our probability is then $(48)_4/(52)_4$ which is the same as before.

In this simple example, either method is relatively straightforward, but in many examples, it is much more natural to treat the selections as ordered/undordered. For hands of cards, treating them as unordered selections usually works best. For something like rolling dice, it usually makes sense to treat them as ordered selections.

Try it out

You are dealt five cards from a well-shuffled deck. Let A be the event that exactly four cards are of the same suit. What is $\mathbb{P}(A)$?

Answer:

There are $\binom{52}{5}$ different unordered selections for the hand, and all are equally likely. How many of these unordered selections are in A? We need to describe a subset of 5 elements such that exactly 4 have the same suit. We build this up sequentially:

- We first choose the suit that we are going to use for the four cards: 4 possibilities.
- Then we choose the four denominations (unordered) for those cards: $\binom{13}{4}$ possibilities.
- All that remains is to choose the last card, which must be of a different suit than the four already chosen: $3 \times 13 = 39$ possibilities.

So the answer is

$$\mathbb{P}(A) = \frac{4 \times \binom{13}{4} \times 39}{\binom{52}{5}} \approx 0.0429.$$

Try it out

In 'Lotto Extra' you have to select 6 numbers from 1 to 49. You win the big prize if 6 randomly drawn numbers match your selection. Let W be the event that you win. Let M_4 be the event that you match exactly 4 out of 6 numbers. Find the probabilities of W and M_4.

Answer:

We model the outcomes of the Lotto draw as unordered selections, so there are $\binom{49}{6} = 13,983,816$ outcomes in total. The event W contains only one of them (your entry)! So $\mathbb{P}(W) = 1/13,983,816$. Now M_4 uses any 4 of your numbers plus any 2 of the remaining $49 - 6 = 43$ numbers. So the number of outcomes in M_4 is

$$\binom{6}{4} \times \binom{43}{2} = \frac{6 \times 5}{2 \times 1} \times \frac{43 \times 42}{2 \times 1} = 15 \times 43 \times 21 = 13,545.$$

Then $\mathbb{P}(M_4) = 13,545/13,983,816 \approx 0.001$.

Advanced content

The same counting arguments can be used when we need to divide m objects into $k > 2$ groups: arranging m distinguishable objects into k groups with sizes $r_1, \dots, r_k$ where $r_1 + \cdots + r_k = m$ can be done in

$$\binom{m}{r_1, \dots, r_k} := \frac{m!}{r_1! \cdots r_k!}$$

ways. The expression $\binom{m}{r_1, \dots, r_k}$ is called the *multinomial coefficient* (Anderson, Seppäläinen, and Valkó 2018 Example 6.7).

2.2.4 Separating objects into groups

In the final section of this chapter, we look into how we can *group* objects: either by combining different types of onect into one big group, or by separating a big group into smaller ones.

Counting principle: Two types of object

Suppose that we have m objects, r of type 1 and $m-r$ of type 2, where objects are indistinguishable from others of their type. The number of distinct, ordered choices of the m objects is

$$\binom{m}{r}.$$

For example, suppose we have four red tokens, and three black ones. Then there are $7!/(4!\,3!) = 35$ different ways to lay them out in a row. The probability that they will be alternately red and black is $1/35$ as there is only one such ordering.

To see why, note that each distinct order for laying out all of the token in a row is precisely the same as choosing 4 of the 7 positions for red ones. In other words, it is an unordered choice of 4 positions from the 7 distinct positions.

Try it out

A coin is tossed 7 times. Let E be the event that a total of 3 heads is obtained. What is $\mathbb{P}(E)$?
Answer:
Consider ordered sequences of H and T: then there are $2^7 = 128$ possible sequences, e.g. HTHTHTT. How many of them are in E? We choose the 3 places where H occurs: $\binom{7}{3} = 35$ ways to do this. The other places are taken by Ts. So the answer is

$$\mathbb{P}(E) = \frac{35}{128}.$$

Advanced content

More generally, using the positions argument again, the multinomial coefficient is the number of ordered choices of objects with k types, r_i of type i, which are indistinguishable within each type.

Counting principle: Separating into groups

The number of ways to divide m indistinguishable objects into k distinct groups is

$$\binom{m+k-1}{m} = \binom{m+k-1}{k-1}.$$

This counting principle lets us work out how many different ways there are to divide one group into smaller groups. My favourite example is a packet of Skittles: if there are 16 or 17 of them in a bag, how many different combinations of the five different flavours could we have?

We can count the number of choices with the ‘sheep-and-fences’ method. Placing all the objects in a line, separated into their groups, there are $k-1$ “fenceposts” between the k groups of sheep (or Skittles).

For example, with 6 objects in 4 groups, we could represent "three in group A, one in group B, none in group C and two in group D" with the drawing $* * * \mid * \mid\mid **$.

We draw $m + k - 1$ 'things' in total (stars and fences). This means that the number of groupings of the objects is the same as the number of choices for the locations of the $k - 1$ fences among the $m + k - 1$ 'things', or $\binom{m+k-1}{k-1} = \binom{m+k-1}{m}$.

Textbook references

If you want more help with this section, check out:

- Section 1.4 in (Blitzstein and Hwang 2019);
- Appendix C in (Anderson, Seppäläinen, and Valkó 2018);
- or Chapter 3 in (Stirzaker 2003).

2.3 Historical context

Classical probability theory originated in calculation of odds for games of chance; as well as contributions by Pierre de Fermat (1601–1665) and Blaise Pascal (1623–1662), comprehensive approaches were given by Abraham de Moivre (1667–1754) (Moivre 1756), Laplace (1749–1827) (Laplace 1825), and Sim'eon Poisson (1781–1840). A collection of these classical methods made just before the advent of the modern axiomatic theory can be found in (Whitworth 1901).

3 Conditional probability and independence

Goals

1. Know the definition of conditional probability and its properties
2. Have a solid knowledge of the partition theorem and Bayes' theorem, recognizing situations where one can apply them.
3. Understand the concept of independence.

3.1 Conditional probability

Definition: conditional probability

For events $A, B \subseteq \Omega$, the *conditional probability* of A *given* B is

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \text{whenever} \;\; \mathbb{P}(B) > 0.$$

In this course, when $\mathbb{P}(B) = 0$, $\mathbb{P}(A \mid B)$ is undefined. The usual interpretation is that $\mathbb{P}(A \mid B)$ represents our probability for A after we have observed B. Conditional probability is therefore very important for statistical reasoning, for example:

- In legal trials. How can we use DNA (or other) evidence to determine the chance that an accused person is guilty?
- Medical screening. How can we make best use of the information from large scale cancer screening programs?

Unfortunately, conditional probability is not always well understood. There are several well-known legal cases that have involved a serious error in probabilistic reasoning: see e.g. Example 2.4.5 of (Anderson, Seppäläinen, and Valkó 2018).

For example, if we roll a fair six-sided die, the conditional probability that the score is odd, given that the score is at most 3, is

$$\mathbb{P}(\text{odd} \mid \text{ at most } 3) = \frac{\mathbb{P}(\{1,3\})}{\mathbb{P}(\{1,2,3\})} = \frac{2/6}{3/6} = \frac{2}{3}.$$

Try it out

Throw three fair coins. What is the conditional probability of at least one head (event A) given at least one tail (event B)?
Answer:
Let H be the event 'all heads', T the event 'all tails'. Then $\mathbb{P}(B) = 1 - \mathbb{P}(H) = 7/8$ and $\mathbb{P}(A \cap B) =$

$1 - \mathbb{P}(H) - \mathbb{P}(T) = 6/8$ so that

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{6/8}{7/8} = \frac{6}{7}.$$

Try it out

Consider a family with two children, whose sex we do not know. The possible sexes are listed by the sample space $\Omega = \{\mathrm{BB}, \mathrm{BG}, \mathrm{GB}, \mathrm{GG}\}$, with the eldest first. Assume that all outcomes are equally likely. Consider the events

$$\begin{aligned} A_1 &= \{\mathrm{GG}\} = \{\text{both girls}\}, \\ A_2 &= \{\mathrm{GB}, \mathrm{BG}, \mathrm{GG}\} = \{\text{at least one girl}\}, \\ A_3 &= \{\mathrm{GB}, \mathrm{GG}\} = \{\text{first child is a girl}\}. \end{aligned}$$

Find $\mathbb{P}(A_1 \mid A_2)$, $\mathbb{P}(A_2 \mid A_1)$, and $\mathbb{P}(A_1 \mid A_3)$.
Answer:
We compute

$$\begin{aligned} \mathbb{P}(A_1 \mid A_2) &= \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_2)} = \frac{\mathbb{P}(\{\mathrm{GG}\})}{\mathbb{P}(\{\mathrm{GB}, \mathrm{BG}, \mathrm{GG}\})} \\ &= \frac{1/4}{3/4} = \frac{1}{3}. \end{aligned}$$

Similarly,

$$\mathbb{P}(A_2 \mid A_1) = \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} = \frac{\mathbb{P}(\{\mathrm{GG}\})}{\mathbb{P}(\{\mathrm{GG}\})} = 1,$$

and

$$\mathbb{P}(A_1 \mid A_3) = \frac{\mathbb{P}(A_1 \cap A_3)}{\mathbb{P}(A_3)} = \frac{\mathbb{P}(\{\mathrm{GG}\})}{\mathbb{P}(\{\mathrm{GB}, \mathrm{GG}\})} = \frac{1/4}{2/4} = \frac{1}{2}.$$

Try it out

Consider throwing two standard dice. Consider the events $F =$ first die shows 6, and $T =$ total is 10. Calculate $\mathbb{P}(F)$ and $\mathbb{P}(F \mid T)$. Before doing any calculation, do you expect $\mathbb{P}(F \mid T)$ to be higher or lower than $\mathbb{P}(F)$? (Hint: 10 is a high total. We'll see later that the 'average' total score on two dice is 7.)
Answer:
The possible outcomes are ordered pairs of the numbers 1 to 6, so $|\Omega| = 6^2 = 36$. In F are all outcomes of the form $(6, ?)$. There are 6 of those, so $\mathbb{P}(F) = 6/36 = 1/6$.
Now $T = \{(6,4), (5,5), (4,6)\}$ so $F \cap T = \{(6,4)\}$, and $\mathbb{P}(F \mid T) = (1/36)/(3/36) = 1/3 > 1/6$.
Similarly, if the total had been 5 we would know that F was impossible!

Textbook references

If you want more help with this section, check out:

- Section 2.2 in (Blitzstein and Hwang 2019);
- Section 2.1 in (Anderson, Seppäläinen, and Valkó 2018);

- or Section 2.1 in (Stirzaker 2003).

3.2 Properties of conditional probability

In this section, we'll meet five key properties of conditional probability.

Key idea: properties of conditional probability

(P1) For any event $B \subseteq \Omega$ for which $\mathbb{P}(B) > 0$, $\mathbb{P}(\cdot \mid B)$ satisfies axioms **A1–A4** (i.e., is a probability on Ω) and therefore also satisfies **C1–C10**.

For example, **C6** for conditional probabilities says that, if $\mathbb{P}(C) > 0$,

$$\mathbb{P}(A \cup B \mid C) = \mathbb{P}(A \mid C) + \mathbb{P}(B \mid C) - \mathbb{P}(A \cap B \mid C).$$

Key idea: properties of conditional probability: multiplication

(P2) For any events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

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

More generally, for any A, B, and C,

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(B \mid C)\,\mathbb{P}(A \mid B \cap C), \qquad \text{if } \mathbb{P}(B \cap C) > 0. \tag{3.1}$$

Some people refer to **P2** as the multiplication rule for probabilities.

Both **P1** and **P2** can be deduced from the definition of probability. For example, Equation 3.1 follows from the fact that

$$\mathbb{P}(B \mid C)\,\mathbb{P}(A \mid B \cap C) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)} \cdot \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)} = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} = \mathbb{P}(A \cap B \mid C).$$

Try it out

Derek is playing Dungarees & Dragons. He rolls an octahedral die to generate the occupant of the room he has just entered. He knows that with probability 3/8 it will be a Goblin, otherwise it will be a Hobbit. A Goblin has a 1 in 4 chance of being equipped with a spiky club. What is the chance that he encounters a Goblin with a spiky club?

Answer:

Let G be the event that the occupant is a Goblin, and let C be the event that the occupant has a spiky club. We are told that $\mathbb{P}(G) = 3/8$ and $\mathbb{P}(C \mid G) = 1/4$, so $\mathbb{P}(G \cap C) = \mathbb{P}(G)\mathbb{P}(C \mid G) = (3/8) \times (1/4) = 3/32$.

Our next property is a more general version of the multiplication rule.

Key idea: properties of conditional probability: multiplication (again)

(P3): For any events $A_0, A_1, \dots, A_k$ with $\mathbb{P}\left(\cap_{i=0}^{k-1} A_i\right) > 0$,

$$\mathbb{P}\left(\bigcap_{i=1}^{k} A_i \mid A_0\right) = \mathbb{P}(A_1 \mid A_0) \times \mathbb{P}(A_2 \mid A_1 \cap A_0) \times \cdots \times \mathbb{P}\left(A_{k-1} \mid \bigcap_{i=0}^{k-2} A_i\right) \times \mathbb{P}\left(A_k \mid \bigcap_{i=0}^{k-1} A_i\right).$$

When $k = 2$, we get **P2**; for $k = 3$, this becomes

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)\,\mathbb{P}(C \mid A \cap B).$$

We can prove this by repeatedly applying Equation 3.1 (in this case, we use it twice).

Try it out

If Derek encounters a Goblin armed with a spiky club, the Goblin will attack, causing a wound with probability 1/2. A Goblin without a spiky club will flee. If Derek encounters a Hobbit, the Hobbit will offer him a cup of tea. What is the probability that Derek is wounded by this encounter?
Answer: Let W be the event that Derek is wounded. Then

$$\mathbb{P}W = \mathbb{P}(G \cap C \cap W) = \mathbb{P}(G)\mathbb{P}(C \mid G)\mathbb{P}(W \mid C \cap G) = \frac{3}{8} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{3}{64}.$$

Key idea: properties of conditional probability: partitions

(P4) If $ E_1, E_2, …, E_k$ form a partition then, for any event A, we have

$$\mathbb{P}(A) = \sum_{i=1}^{k} P(E_i)\,P(A \mid E_i)\,. \tag{3.2}$$

More generally, if $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A \mid B) = \sum_{i=1}^{k} P(E_i \mid B)\,P(A \mid E_i \cap B)\,.$$

This result is often called the *partition theorem*, or the *law of total probability.* (If you've forgotten what a partition is, head back to Section 1.6.)

To prove **P4** is true, we first use **P2** on the right-hand side of Equation 3.2 to get

$$\sum_{i=1}^{k} P(E_i)\,P(A \mid E_i) = \sum_{i=1}^{k} P(A \cap E_i)\,.$$

But since the E_i form a partition, they are pairwise disjoint, and hence so are the $A \cap E_i$, so by **C7**

$$\sum_{i=1}^{k} P(A \cap E_i) = P(\cup_{i=1}^{k}(A \cap E_i)) = P(A \cap (\cup_{i=1}^{k} E_i))\,,$$

but since the E_i form a partition, $\cup_{i=1}^{k} E_i = \Omega$, giving the result. You should check that **P4** remains true (with $k = \infty$) for infinite partitions.

Try it out

Back in the land of Dungarees and Dragons, suppose that the Goblin player has a special token that he will play so that even unarmed Goblins will attack, rather than flee. An unarmed Goblin causes a wound with probability 1/6. Now what is the chance that Derek is wounded?

Answer:

The partition we use is G^c, $G \cap C$, and $G \cap C^c$. We know from our previous examples that $P(G^c) = 5/8$, $\mathbb{P}(G \cap C) = 3/32$, and $P(G \cap C^c) = \mathbb{P}(G)P(C^c \mid G) = 9/32$.

We also know that $P(W \mid G^c) = 0$, $P(W \mid G \cap C) = 1/2$, and, now, $P(W \mid G \cap C^c) = 1/6$. So

$$\begin{aligned}\mathbb{P}W &= P(G^c)\,P(W \mid G^c) + \mathbb{P}(G \cap C)P(W \mid G \cap C) + P(G \cap C^c)\,P(W \mid G \cap C^c)\\ &= 0 + \frac{3}{32} \cdot \frac{1}{2} + \frac{9}{32} \cdot \frac{1}{6} = \frac{3}{32}.\end{aligned}$$

Try it out

Three machines, A, B and C, produce components. 10% of components from A are faulty, 20% of components from B are faulty and 30% of components from C are faulty. Equal numbers from each machine are collected in a packet. One component is selected at random from the packet. What is the probability that it is faulty?

Answer:

Let F be the event that the component is faulty. Let M_A, M_B, M_C be the events that the component is from machines A, B, C respectively. Then M_A, M_B, M_C form a partition so

$$\begin{aligned}P(F) &= P(M_A)\,P(F \mid M_A) + P(M_B)\,P(F \mid M_B) + P(M_C)\,P(F \mid M_C)\\ &= 0.1 \times \frac{1}{3} + 0.2 \times \frac{1}{3} + 0.3 \times \frac{1}{3} = 0.2.\end{aligned}$$

The most important result in conditional probability is *Bayes' theorem*. It allows us to express the conditional probability of an event A given B in terms of the "inverse" conditional probability of B given A.

Key idea: properties of conditional probability: Bayes theorem

(P5) For any events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

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

More generally, if $\mathbb{P}(A \mid C) > 0$ and $\mathbb{P}(B \mid C) > 0$, then

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

Try it out

Suppose that in the previous example, the component was indeed faulty. What is the probability that it came from machine A?

Answer:

We have, by Bayes' theorem (**P5**),

$$P(M_A \mid F) = \frac{P(F \mid M_A)\, P(M_A)}{P(F)} = \frac{(1/10)(1/3)}{1/5} = \frac{1}{6}.$$

Try it out

Pat ends up in the pub of an evening with probability 3/10. If she goes to the pub, she will get drunk with probability 1/2. If she stays in, she will get drunk with probability 1/5. What is the probability that she gets drunk? Given that she does get drunk, what is the probability that she went to the pub?

Answer:

Let P be the event that she goes to the pub, and D be the event that she gets drunk. Then we are told that $P(P) = 3/10$, $P(D \mid P) = 1/2$ and $P(D \mid P^c) = 1/5$. Using the partition P, P^c we get

$$\mathbb{P}(D) = P(P)\, P(D \mid P) + P(P^c)\, P(D \mid P^c) = \frac{3}{10} \cdot \frac{1}{2} + \frac{7}{10} \cdot \frac{1}{5} = \frac{29}{100}.$$

Then, by Bayes' theorem (P5),

$$P(P \mid D) = \frac{P(D \mid P)\, P(P)}{\mathbb{P}(D)} = \frac{\frac{3}{10} \cdot \frac{1}{2}}{\frac{29}{100}} = \frac{15}{29}.$$

Try it out

There are three regions (A,B,C) in a country with populations in relative proportions $5:3:2$. In region A, 5% of people own a rabbit. In region B, it is 10%, and in region C, it is 15%.
i.What proportion of people nationally own rabbits? ii. What proportion of rabbit-owners come from region A?

Answer:

Let A, B, C be the events that a randomly-chosen individual comes from regions A, B, C respectively. Let R be the event that the individual is a rabbit owner. Then

$$P(R) = \mathbb{P}(A)P(R \mid A) + \mathbb{P}(B)P(R \mid B) + \mathbb{P}(C)P(R \mid C) = \frac{5}{10} \cdot \frac{1}{20} + \frac{3}{10} \cdot \frac{1}{10} + \frac{2}{10} \cdot \frac{3}{20} = \frac{17}{200}.$$

And, by Bayes' theorem,

$$P(A \mid R) = \frac{P(R \mid A)\, \mathbb{P}(A)}{P(R)} = \frac{\frac{5}{10} \cdot \frac{1}{20}}{\frac{17}{200}} = \frac{5}{17}.$$

Try it out

One of a set of n people committed a crime. A suspect has been arrested, and DNA evidence is a match. Consider the events $G =$ suspect is guilty, and $E =$ DNA evidence is a match. Suppose that we initially believe that $\mathbb{P}(G) = \alpha/n$. The probability of a 'false positive' DNA match is $P(E \mid G^c) = p$.
What is our new probability that the suspect is guilty, given the DNA evidence?

Answer:

We use the partition G, G^{c}. Then, by P4,

$$\begin{aligned} P(E) &= P(E \mid G)\,\mathbb{P}(G) + P(E \mid G^{\mathrm{c}})\,P(G^{\mathrm{c}}) \\ &= 1 \times \frac{\alpha}{n} + p \times \left(1 - \frac{\alpha}{n}\right) \\ &= \frac{\alpha + (n-\alpha)p}{n}. \end{aligned}$$

Then by Bayes' theorem (P5),

$$P(G \mid E) = \frac{P(E \mid G)\,\mathbb{P}(G)}{P(E)} = \frac{\alpha/n}{(\alpha + (n-\alpha)p)/n} = \frac{\alpha}{\alpha + (n-\alpha)p}.$$

Typically: $\alpha \approx 1$ and n is very large, and fairly easy to asses. On the other hand p is very small, and is difficult to assess as it requires a lot of information about the genetic make up of a (potentially large) group of people. When n is small it may be possible to test all of the group. A great variety of mistakes have been made in using complex evidence of this type in courts. The famous 'prosecutor's fallacy' is pretending that $P(G^{\mathrm{c}} \mid E) = P(E \mid G^{\mathrm{c}})$ (of course this is wrong).

We can also combine properties **P4** and **P5** to make a mega-property of conditional expectation: Bayes' theorem for partitions.

Definition: properties of conditional probability: Bayes theorem for partitions

For any partition A_1, ..., A_k and any B with $\mathbb{P}(B) > 0$,

$$P(A_i \mid B) = \frac{P(A_i)\,P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j)\,P(B \mid A_j)}.$$

More generally, if $\mathbb{P}(B \mid C) > 0$,

$$P(A_i \mid B \cap C) = \frac{P(A_i \mid C)\,P(B \mid A_i \cap C)}{\sum_{j=1}^{k} P(A_j \mid C)\,P(B \mid A_j \cap C)}.$$

Try it out

On any given day, it rains with probability 1/2. If it rains, Charlie the cat will go outside with probability 1/10; if it is dry, the probability is 3/5. If Charlie goes outside, what is the conditional probability that it has rained?

Answer:

Let $R =$ it rains, $C =$ Charlie goes outside. Then R, R^{c} form a partition with $P(R) = P(R^{\mathrm{c}}) = 1/2$. Also, $P(C \mid R) = 1/10$ and $P(C \mid R^{\mathrm{c}}) = 3/5$. So, by Bayes' theorem (P6),

$$\begin{aligned} P(R \mid C) &= \frac{P(C \mid R)\,P(R)}{P(C \mid R)\,P(R) + P(C \mid R^{\mathrm{c}})\,P(R^{\mathrm{c}})} \\ &= \frac{\frac{1}{10} \cdot \frac{1}{2}}{\frac{1}{10} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{1}{2}} = \frac{1}{7}. \end{aligned}$$

Textbook references

If you want more help with this section, check out:

- Sections 2.3 and 2.4 in (Blitzstein and Hwang 2019);
- Sections 2.1 and 2.2 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 2.1 in (Stirzaker 2003).

3.3 Independence of events

Tied to the idea of conditional probability is the idea of *independence*: the property that two events are unrelated, or have no bearing on each other's likelihood.

Key idea: Independence of two events

We say that two events A and B are *independent* whenever

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

We say that two events A and B are *conditionally independent* given a third event C with $\mathbb{P}(C) > 0$ whenever

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

For example, if we pick a card from a well-shuffled deck, the events "the card is red" (R) and "the card is an Ace" (A) are independent.

By counting, we have that $P(R) = \frac{26}{52} = \frac{1}{2}$ and $\mathbb{P}(A) = \frac{4}{52} = \frac{1}{13}$. Now, $A \cap R = \{A\diamondsuit, A\heartsuit\}$ so $P(A \cap R) = \frac{2}{52} = \frac{1}{26}$, We check that $\frac{1}{26} = \frac{1}{2} \cdot \frac{1}{13}$, so R and A are indeed independent.

Try it out

Roll two standard dice. Let E be the event that we have an even outcome on the first die. Let F be the event that we have a 4 or 5 on the second die. Are E and F independent?
Answer:
We will verify using a counting argument.
There are 36 equally likely outcomes, namely:

$$\Omega = \{(i, j) \colon i \in \{1, \dots, 6\} \text{ and } j \in \{1, \dots, 6\}\}$$

Of those, 3×6 are in E, and 6×2 are in F, so $P(E) = 18/36 = 1/2$ and $P(F) = 12/36 = 1/3$. Moreover, 3×2 of these outcomes belong to both E and F, so $P(E \cap F) = 6/36 = 1/6$. Indeed,

$$P(E \cap F) = 1/6 = 1/2 \times 1/3 = P(E)\,P(F)\,.$$

So E and F are independent.

Try it out

Roll a fair die. Consider the events

$$A_1 = \{2,4,6\},\ A_2 = \{3,6\},\ A_3 = \{4,5,6\},\ \text{and}\ A_4 = \{1,2\}.$$

Which pairs of events are independent?
Answer:
Note that $A_1 \cap A_2 = \{6\}$ so $P(A_1 \cap A_2) = \frac{1}{6}$, while $P(A_1)\,P(A_2) = \frac{3}{6} \cdot \frac{2}{6} = \frac{1}{6}$ too. So A_1 and A_2 are independent.
On the other hand, $A_1 \cap A_3 = \{4,6\}$ so $P(A_1 \cap A_3) = \frac{2}{6} = \frac{1}{3}$, but $P(A_1)\,P(A_3) = \frac{3}{6} \cdot \frac{3}{6} = \frac{1}{4}$. So A_1 and A_3 are *not* independent.

Never confuse disjoint events with independent events! For independent events, we have that $P(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, but for disjoint events, $P(A \cap B) = 0$ because $A \cap B = \emptyset$.

Disjointness is a property of the *sets* only (it can be seen from the Venn diagram). Independence is a property of *probabilities* (it cannot be seen from the Venn diagram).

Try it out

In the context of the previous example, $A_3 \cap A_4 = \emptyset$, so A_3 and A_4 are disjoint. They are certainly not independent, since $P(A_3 \cap A_4) = 0$ but $P(A_3)\,P(A_4) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \neq 0$.

The next theorem explains why independence is called independence:

Consider any two events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. The following statements are equivalent.

(i) $P(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

(ii) $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

(iii) $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

In other words, learning about B will not tell us anything new about A, and similarly, learning about A will not tell us anything new about B.

For conditional independence, we have a similar result.

Consider any three events A, B, and C, with $P(A \cap B \cap C) > 0$. The following statements are equivalent.

(i) $P(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C)$.

(ii) $P(A \mid B \cap C) = \mathbb{P}(A \mid C)$.

(iii) $P(B \mid A \cap C) = \mathbb{P}(B \mid C)$.

In other words, if we know C then learning about B will not tell us anything new about A, and similarly, if we know C then learning about A will not tell us anything new about B.

Consider the card-shuffling example again. The probability that our card is an Ace is $\mathbb{P}(A) = 1/13$ and the probabilitiy that it is an Ace, given it is red, is

$$P(A \mid R) = \frac{P(A \cap R)}{P(R)} = \mathbb{P}(A),$$

by independence. The 'reason' for the independence is that the proportion of aces in the deck (4/52) is the same as that of aces among the red cards (2/26).

Key idea

It is possible for two events to be conditionally independent on particular events, but not to be (unconditionally) independent. We will see an example of this when we discuss genetics, in Section 5.2.

It can be extremely useful to recognize situations where (conditional) independence can be applied. Of course, it is equally important not to assume (conditional) independence where there really are dependencies.

Definition: independence for multiple events

A (possibly infinite) collection of events $\mathcal{A} \subseteq \mathcal{F}$ are *mutually independent* if for every *finite* non-empty $\mathcal{C} \subseteq \mathcal{A}$ (that is, $\mathcal{B}$ is a finite subcollection of the events in question),

$$P\left(\bigcap_{A \in \mathcal{C}} A\right) = \prod_{A \in \mathcal{C}} \mathbb{P}(A).$$

A collection of events $\mathcal{A} \subseteq \mathcal{F}$ are *mutually conditionally independent*]{.alert}* given another event B if for every finite non-empty subcollection $\mathcal{C} \subseteq \mathcal{A}$,

$$P\left(\bigcap_{A \in \mathcal{C}} A \mid B\right) = \prod_{A \in \mathcal{C}} \mathbb{P}(A \mid B).$$

The smallest case here is to consider three events. We say that the events A, B, and C are mutually independent if all of the following equalities are satisfied:

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

Suppose we roll 4 dice and their values are independent.

To find the probability that we throw no sixes let A_i be the event 'the ith throw is not a 6'. By assumption A_1, ..., A_4 are independent so

$$P(\text{no sixes on 4 dice}) = P\left(\bigcap_{i=1}^{4} A_i\right) = \prod_{i=1}^{4} P(A_i) = \left(\frac{5}{6}\right)^4.$$

The same result is obtained from the classical model, by selection with replacement.

It is possible for events to be *pairwise* independent without being *mutually* independent, as the next example demonstrates.

Examples: Example

Toss two fair coins. The sample space is $\Omega = \{HH, HT, TH, TT\}$ and each outcome has probability $1/4$.

Let $A = \{HH, HT\}$ be the event that the first coin comes up 'heads', $B = \{HH, TH\}$ the event that the second coin comes up 'heads', and $C = \{HH, TT\}$ the event that the coins come up the same. Then since $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$ and each pairwise intersection has probability $1/4$, it is easy to see that the events are pairwise independent. However, $P(A \cap B \cap C) = P(HH) = 1/4$ which is not the same as $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = 1/8$, so the three events are *not* mutually independent.
To interpret this in words, if we consider any two of the events, the occurrence of one tells us nothing about the occurrence of the other. As soon as we consider statements involving all three events, however, we see the dependence. For example,

$$P(C \mid A \cap B) = 1,$$

since $A \cap B = \{HH\}$ and $\{HH\} \subseteq C$, compared to the unconditional probability $\mathbb{P}(C) = 1/2$.

Textbook references

If you want more help with this section, check out:

- Section 2.5 in (Blitzstein and Hwang 2019);
- Section 2.3 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 2.2 in (Stirzaker 2003).

3.4 Historical context

Bayes' theorem is named after the Reverend Thomas Bayes (1701–1761); it was published after his death, in 1763. In our modern approach to probability, the theorem is a very simple consequence of our definitions; however, the result may be interpreted more widely, and is one of the most important results regarding statistical reasoning.

Figure 3.1: Thomas Bayes

4 Interpretations of probability

Goals

1. Understand that there are different ways to interpret probability values.

This chapter covers some different ways in which we can interpret probabilities in the real world. The axioms in Chapter 1 are helpful to determine a framework for mathematical probability, but they leave us lots of room to choose a model within that framework.

We have already discussed one approach in Chapter 2: the "classical" approach, in which each outcome in the sample space is assigned the same probability. This interpretation has some obvious limitations in practice. Often we cannot find a set of outcomes that it is reasonable to think of as a priori equally likely. Therefore, it is essential to have more widely applicable models to deal with uncertainty.

In this chapter, we discuss two more approaches to determining probabilities in real-world applications: the *relative frequencies* approach and the *subjective probability* (or betting) approach. Either can be used, depending on the context, to help us to assign probabilities to events.

The goal of this chapter is to help you to develop more intuition and probabilistic thinking. For the rest of the course, we will assume that we "know" the probability of each event, without worrying too much about how it was determined.

4.1 Relative frequency interpretation

This interpretation applies to **trials** giving chance outcomes of an experiment that can be repeated indefinitely under essentially unchanged conditions and which exhibits **long term regularity**.

Suppose that we run n trials of an experiment with a known list of possible outcomes and the number of trials on which event A occurs is n_A (A is again a set of possible outcomes). The **relative frequency** of occurrence of A is n_A/n.

For example, if we toss a coin 1000 times and observe 490 heads, then the relative frequency of heads is 490/1000.

For some experiments, it may be reasonable to suppose that relative frequencies are stable for very large n.

If we toss a fair coin one billion times, we might expect that the relative frequency of heads after the first few thousand throws would remain very close to 1/2.

As a mathematical idealization, we suppose that there is a unique, empirical limiting value for n_A/n, as n tends to infinity, which we call the *relative frequency probability* of A.

For our coin, the statement $P(\text{heads}) = 1/2$ means 'if we tossed the coin an extremely large number of times, then the proportion of heads would be arbitrarily close to 1/2'.

This interpretation is widely used, especially in physics, where experiments are designed for repeatability and we can expect future trials to behave like those in the past. In this view, probability is a property of the experimental setup and may be “objectively’ ’ discovered by sufficient repetitions of the experiment.

Amongst the problems with this interpretation are:

- it is often impossible to decide what “essentially unchanged conditions’ ’ are;
- we often have no way of knowing when limiting frequencies become stable (how many trials should we do to test this?);
- we can only use it in situations that are repeatable.

4.2 Betting interpretation

A very different way of interpreting probability goes by considering probability as a quantification of someone’s (yours, mine, your neighbour, …) belief that an event will occur. There are various different ways in which we can measure this belief numerically. Here is one of the simplest.

Your **subjective probability** that A will occur is measured by the amount $£p_A$ that you would consider to be a fair price for the following gamble:

- if A occurs, you receive £1;
- if A does not occur, you receive nothing.

In this interpretation, there are no “true’ ’ probabilities. Different individuals will have different information relevant to a problem and so may validly make different probability assessments.

For instance, if you say your probability that ‘Your Team’ wins its next match is 1/2 this means that you view £1/2 as a fair price for the gamble winning you £1 if Your Team wins but otherwise nothing. Others may disagree with you.

Subjective probability ideas are often used by decision makers who have to consider problems concerning unique, non-repeatable events, based on their informed but subjective judgements. The advantages of this interpretation are that probability measures the belief of a subject, and is no longer seen as a property of the experimental setup. Potential issues are that the highest ‘buying price’ may differ from lowest ‘selling price’; a subject may have reason to misrepresent their fair price; placing the bet itself might affect the experiment.

4.3 Interpretation and the axioms

We claimed that the axioms of probability are the same regardless of the interpretation of the probabilities that we are using. **A1** and **A2** are clearly very sensible in any interpretation. The justification of **A3** (and, by extension, **A4**) needs some more thought.

A3 feels intuitive for the classical model of probability by its relation to counting: in the classical model if A contains m_A outcomes and B contains m_B outcomes, with none in common with A, then $m_{A\cup B} = m_A + m_B$. The argument is very similar for the relative frequency model and only slightly more subtle for the betting model.

Textbook references

For more information on these ideas, check out:

- Section 1.2 in (DeGroot and Schervish 2013);
- Chapter 0 in (Stirzaker 2003);
- or (Hájek 2012).

5 Some applications of probability

Goals

1. Understand the meaning of a reliability network.
2. Know how to evaluate the reliability of networks.
3. Understand the probabilistic nature of genetics.
4. Know how to derive the probability of genotypes of children given the genotypes of parents.
5. Know how and when to exploit the general rules of (conditional) probability to solve complex reliability networks and problems in genetics.

5.1 Reliability of networks

Reliability theory concerns mathematical models of systems that are made up of individual components which may be faulty.

If components fail randomly, a key objective of the theory is to determine the probability that the system as a whole works. This will depend on the structure of the system (how the components are organized). This is an important problem in industrial (or other) applications, such as electronic systems, mechanical systems, or networks of roads, railways, telephone lines, and so on.

Once we know how to work out (or estimate) failure probabilities of these systems, we can start to ask more sophisticated questions, such as: How should the system be designed to minimize the failure probability, given certain practical constraints? What is a good inspection, servicing and maintenance policy to maximize the life of the system for a minimal cost?

In this course, to demonstrate an application of the probabilistic ideas we have covered so far, we address the basic question: Given a system made up of finitely many components, what is the probability that the system works? Whether the system functions depends on whether the components function, and on the configuration of those components.

Example

The figure below shows (a) two components in series, (b) three in parallel, (c) a four component system. In each case assume that the system works if it is possible to get from the left end to the right through functioning components.

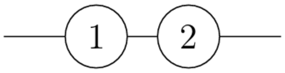


Figure 5.1: a) Two components in series

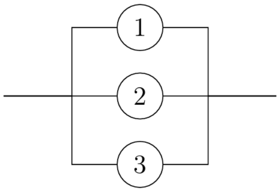


Figure 5.2: b) Three components in parallel

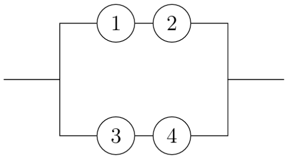


Figure 5.3: c) A system with four components

The system in (a) works if and only if both components 1 and 2 work.
The system in (b) works if any of 1, 2, 3 work.
The system in (c) works if either both 1 and 2 work, or both 3 and 4 work (or all work).

Definition: reliability network

A *reliability network* is a diagram of nodes and arcs. The nodes represent components of a multi-component system, where each node is either working or is broken, and where the entire system works if it is possible to get from the left end to the right of the diagram through working components only. Suppose the ith component functions with probability p_i, $i \in \{1, 2, \dots, k\}$, and different components are independent. The probability that the system works is then a function of the probabilities p_1, ..., p_k. We denote this function by $r(p_1, p_2, \dots, p_k)$, and call it the *reliability function*. It is determined by the layout of the reliability network.

Looking at reliability networks, and determining their reliability functions, is a source of lots of good examples to practice working with the axioms of probability (in particular **A3**, **C1**, and **C6**), as well as building up some more intuition about independence. Let's have one more example (and there are a couple on the problem sheet, too).

Try it out

Consider the three networks in the previous example. We consider the events

$$W_i = \text{component } i \text{ works}, \quad S = \text{system works}.$$

Calculate $P(S)$ for each of the networks.
Suppose that the system in (c) works. What is the conditional probability that component 1 works?
Answer:
The first step is to represent S in terms of the W_i and the operations of set theory. Then we can compute $P(S)$ using our rules for probabilities.
In (a),

$$S = W_1 \cap W_2.$$

By independence, $P(S) = P(W_1)\,P(W_2) = p_1 p_2$.
For (b), we have

$$S = W_1 \cup W_2 \cup W_3.$$

It is easiest to compute

$$P(S^c) = P(W_1^c \cap W_2^c \cap W_3^c) = P(W_1^c)\,P(W_2^c)\,P(W_3^c) = (1-p_1)(1-p_2)(1-p_3),$$

by independence. So

$$P(S) = 1 - (1-p_1)(1-p_2)(1-p_3).$$

For (c), we have

$$S = (W_1 \cap W_2) \cup (W_3 \cap W_4).$$

Then, by C6,

$$\begin{aligned} P(S) &= P(W_1 \cap W_2) + P(W_3 \cap W_4) - P(W_1 \cap W_2 \cap W_3 \cap W_4) \\ &= p_1 p_2 + p_3 p_4 - p_1 p_2 p_3 p_4. \end{aligned}$$

To find the conditional probability that component 1 works, given that system (c) works, we go back to the definition of conditional probability:

$$\begin{aligned} P(W_1 \mid S) &= \frac{P(W_1 \cap S)}{P(S)} \\ &= \frac{P((W_1 \cap W_2) \cup (W_1 \cap W_3 \cap W_4))}{P(S)} \\ &= \frac{p_1 p_2 + p_1 p_3 p_4 - p_1 p_2 p_3 p_4}{p_1 p_2 + p_3 p_4 - p_1 p_2 p_3 p_4}. \end{aligned}$$

Textbook references

If you want more help with this section, check out:

- Sections 4.1–4.4 in (Billinton and Allan 1996);
- or Chapter 9 in (Ross 2010).

5.2 Genetics

Inherited characteristics are determined by *genes*. The mechanism governing inheritance is random and so the laws of probability are crucial to understanding genetics.

Your cells contain 23 pairs of *chromosomes*, each containing many genes (while 23 pairs is specific to humans the idea is similar for all animals and plants). The genes take different forms called *alleles* and this is one reason why people differ (there are also environmental factors). Of the 23 pairs of chromosomes, 22 pairs are *homologous* (each of the pair has an allele for any gene located on this pair). People with different alleles are grouped by visible characteristics into *phenotypes*; often one allele, A say, is *dominant* and another, a, is *recessive* in which case AA and Aa are of the same phenotype while aa is distinct. Sometimes, the recessive gene is rare and the corresponding phenotype is harmful, for example haemophilia or sickle-cell anaemia.

For instance, in certain types of mice, the gene for coat colour (a phenotype) has alleles B (black) or b (brown). B is dominant, so BB or Bb mice are black, while bb mice are brown with no difference between Bb and bB.

With sickle-cell anaemia, allele A produces normal red blood cells but a produces deformed cells. Genotype aa is fatal but Aa provides protection against malarial infection (which is often fatal) and so allele a is common in some areas of high malaria risk.

To apply probability theory to the study of genetics, we use the **basic principle of genetics:** For each gene on a homologous chromosome, a child receives one allele from each parent, where each allele received is chosen independently and at random from each parent's two alleles for that gene.

Examples: Example

For example, in a certain type of flowering pea, flower colour is determined by a gene with alleles R and W, with phenotypes RR (red), RW (pink) and WW (white). The table of offspring genotype probabilities given parental genotypes is

Parental genotype		RR RR	RR RW	RR WW	RW RW	RW WW	WW WW
offspring	RR	1	1/2	0	1/4	0	0
genotype	RW	0	1/2	1	1/2	1/2	0
	WW	0	0	0	1/4	1/2	1

How to read the table: for example, parents RR and RW produce RR offspring with chance 1/2 (the RR parent must supply an R but the RW parent supplies either R or W, each with probability 1/2). When we cross red and white peas, all the offspring will be pink but when we cross red and pink peas, about half of the peas will be red, half pink. Mendel carried out experiments like these to establish the genetic basis of inheritance.

Advanced content

Similar but larger tables are relevant when there are more than two alleles.

It is extremely important to note that **genotypes of siblings are dependent unless we condition on parental genotypes**. For example, if two black mice (which may each be BB or Bb) have 100 black offspring, you may conclude that the next offspring is overwhelmingly likely to also be black, because it is very likely that at least one parent is BB.

Genes can also affect reproductive fitness, as we see in the next example.

Try it out

A gene has alleles A and a but a is recessive and harmful, so genotype aa does not reproduce while AA, Aa are indistinguishable. With proportions $1 - \lambda$, λ of AA, Aa in the healthy population, show that the probability of an aa offspring is $\lambda^2/4$.
Answer: To show this we can use the partition F_{AA}, F_{Aa} (father AA, Aa respectively) to calculate

$$P(Fa) = P(F_{AA})\,P(Fa \mid F_{AA}) + P(F_{Aa})\,P(Fa \mid F_{Aa}) = 0 + \lambda \times (1/2) = \lambda/2$$

for the event Fa that the father provides allele a.
By symmetry, the mother also supplies allele a with probability $\lambda/2$ and by independence (random mating) the probability that they both supply allele a is $\lambda^2/4$ e.g. when $\lambda \approx 1/2$, about 6% of

offspring will be aa. Over time the proportion of allele a will decrease unless Aa has a reproductive advantage over AA.

Things are slightly different for genes on the X or Y chromosomes (sex-linked genes).

These are the final chromosome pair, known as the sex chromosomes. Each may be X, a long chromosome, or Y, a short chromosome. Most of the genes on X do not occur on Y. Most people have sex determined as XX (female) or XY (male); YY is not possible.[1]

Try it out

A gene carried on the X chromosome has alleles A and a (so men have only one allele, while women have two).

- aa women are unhealthy;
- a men are unhealthy;
- otherwise the person is healthy.

A male child inherits his gene on the X chromosome from his mother (as he must get his Y from his father) with equal chance of the two alleles that the mother carries. A female child inherits her father's single allele as well as one of her mother's two alleles.
Jane is healthy. Her maternal aunt has an unhealthy son (Jane's cousin). Jane's maternal grandparents and her father are all healthy.

i. What is the probability that Jane is genotype Aa?

Now suppose that Jane has two healthy brothers.

ii. What now is the probability that Jane is genotype Aa?

Answer: We start with part i. From the information given, we can add some information to the genetic tree. The healthy men are A. A male child receives his single (X-carried) allele as a random selection from his mother's two alleles (the genotype of his father has no bearing). Thus Jane's Aunt must carry an a. She cannot have inherited this from the healthy grandfather, so the grandmother must also carry an a. This gives us the picture below.

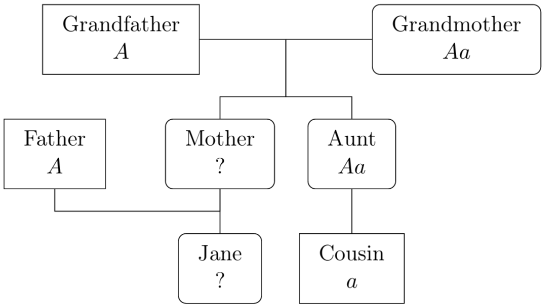


Figure 5.4: Jane's family tree, pt1

[1] X, XXX, XXY, and XYY can occur, but are very rare.

Consider the events $J = \{\text{Jane is } Aa\}$, $M_1 = \{\text{mother is } AA\}$, and $M_2 = \{\text{mother is } Aa\}$. From the tree above, we have M_1 occurs if and only if Jane's mother inherited an A from her mother, i.e., $P(M_1) = 1/2$ and $P(M_2) = 1/2$ too. Given the mother's genotype, we can work out the probabilities for Jane's inheritance. Thus, by the partition theorem,

$$\begin{aligned}P(J) &= P(M_1)\, P(J \mid M_1) + P(M_2)\, P(J \mid M_2) \\ &= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.\end{aligned}$$

Now we move on to part ii. The tree is now augmented by the additional information about Jane's siblings:

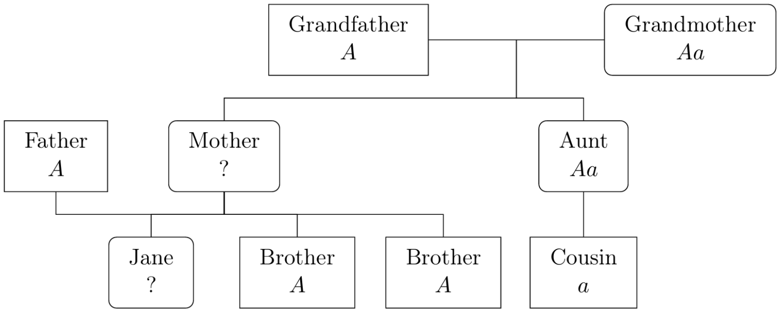


Figure 5.5: Jane's family tree, pt2

Let $B = \{\text{Two brothers are } A\}$. We want $P(J \mid B)$. Note that our knowledge of B changes our beliefs about the genotype of Jane's mother. To see this more clearly, imagine that Jane had 100 brothers, all of whom were of type A. Then we would be very nearly sure that Jane's mother was of type AA, and so Jane would be almost certainly of type AA too.
For the calculation, we use the partition theorem for conditional probabilities:

$$P(J \mid B) = P(M_1 \mid B)\, P(J \mid M_1 \cap B) + P(M_2 \mid B)\, P(J \mid M_2 \cap B)\,.$$

But given M_i, J is independent of B so

$$P(J \mid B) = P(M_1 \mid B)\, P(J \mid M_1) + P(M_2 \mid B)\, P(J \mid M_2)\,.$$

As above, we have $P(J \mid M_1) = 0$ and $P(J \mid M_2) = 1/2$. By Bayes's theorem,

$$\begin{aligned}P(M_2 \mid B) &= \frac{P(B \mid M_2)\, P(M_2)}{P(B \mid M_1)\, P(M_1) + P(B \mid M_2)\, P(M_2)} \\ &= \frac{(1/2)^2 \cdot (1/2)}{1^2(1/2) + (1/2)^2 \cdot (1/2)} = \frac{1}{5}.\end{aligned}$$

So

$$P(J \mid B) = 0 + \frac{1}{2} \cdot \frac{1}{5} = \frac{1}{10}.$$

So seeing that Jane has two healthy brothers significantly reduces the chance that Jane is carrying an a.

Textbook references

If you want more help with this section, check out

- Section V.5 in (Feller 1968);
- or Section 5.6 in (Chung and AitSahlia 2003).

5.3 Hardy-Weinberg equilibrium

Consider a population of a large number of individuals evolving over successive generations. Consider a gene (on a homologous chromosome) with two alleles A and a and genotypes $\{AA, Aa, aa\}$. Suppose the genotype proportions in the population (uniformly for males and females) at generation $n = 0, 1, 2, ...$ are

AA	Aa	aa
u_n	$2v_n$	w_n

where we have $u_n + 2v_n + w_n = 1$. Suppose also that the proportions of the alleles in the population are

A	a
p	q

where $p_n + q_n = 1$. We see that

$$p_n = \frac{2u_n + 2v_n}{2u_n + 4v_n + 2w_n} = u_n + v_n$$

and, similarly, $q_n = v_n + w_n$.

Suppose that

- the gene is *neutral*, meaning that different genotypes have equal reproductive success;
- there is *random mating* with respect to this gene, meaning that each individual in generation $n+1$ draws randomly two parents whose genotypes are independently in the proportions u_n, $2v_n$, w_n.

How do the genotype proportions evolve over successive generations?

Consider the offspring of generation 0. Let FA = event that child gets allele A from father, MA = event that child gets allele A from mother, F_{AA} = event that father is AA, F_{Aa} = event that father is Aa, F_{aa} = event that father is aa. Then

$$\begin{aligned} P(FA) &= P(F_{AA})\, P(FA \mid F_{AA}) + P(F_{Aa})\, P(FA \mid F_{Aa}) + P(F_{aa})\, P(FA \mid F_{aa}) \\ &= 1 \cdot u_0 + \frac{1}{2} \cdot 2v_0 + 0 \cdot w_0 = u_0 + v_0 = p_0. \end{aligned}$$

Similarly, $P(MA) = p_0$. In particular, since parents contribute alleles independently, the probability distribution of the genotype of an individual in generation 1 is

AA	Aa	aa
p_0^2	$2p_0(1-p_0)$	$(1-p_0)^2$

Provided that the population is large enough (see the *law of large numbers* in **?@sec-limits**) these will also be the generation 1 proportions of AA, Aa, aa, i.e.,

$$u_1 = p_0^2, \qquad v_1 = p_0(1-p_0), \qquad w_1 = (1-p_0)^2.$$

Now let $p_1 = u_1 + v_1$ be the proportion of A in the gene pool at generation 1. Substituting the values of u_1, v_1 we find that

$$p_1 = u_1 + v_1 = p_0^2 + p_0(1-p_0) = p_0,$$

i.e. the proportions of A and a in the gene pool are constant.

The same argument applies for later generations, so that $p_n = p_0$ for all n, i.e., the proportions of the two alleles in the gene pool remain constant. This means that, for $n \geq 1$,

$$u_n = p_0^2, \qquad v_n = p_0(1-p_0), \qquad w_n = (1-p_0)^2,$$

so that the proportions of the three genotypes in the population remain constant in every generation after the first. This is called the *Hardy–Weinberg equilibrium.*

5.4 Historical context

Reliability for systems of infinitely many components is related to *percolation.*

On the infinite square lattice $\mathbb{Z}^2$, declare each vertex to be *open*, independently, with probability $p \in [0,1]$, else it is *closed.* Consider the *open cluster* containing the origin, that is, the set of vertices that can be reached by nearest-neighbour steps from the origin using only open vertices. Percolation asks the question: for which values of p is the open cluster containing the origin *infinite* with positive probability? It turns out that for this model, the answer is: for all $p > p_{\mathrm{c}}$ where $p_{\mathrm{c}} \approx 0.593$.

The picture shows part of a percolation configuration, with open sites indicated by black dots and edges between open sites indicated by unbroken lines.

Percolation is an important example of a probability model that displays a *phase transition.* You may see more about it if you do later probability courses.

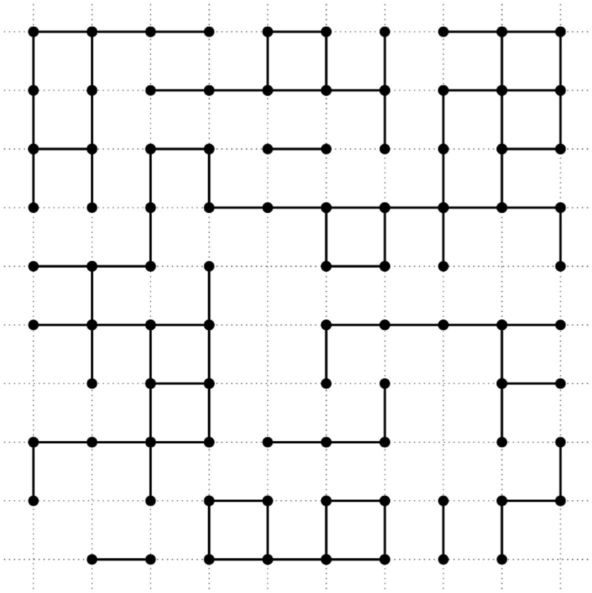


Figure 5.6: A lattice, with some edges missing

The laws governing the statistical nature of inheritance were first observed and formulated by monk Gregor Johann Mendel (1822–1884).

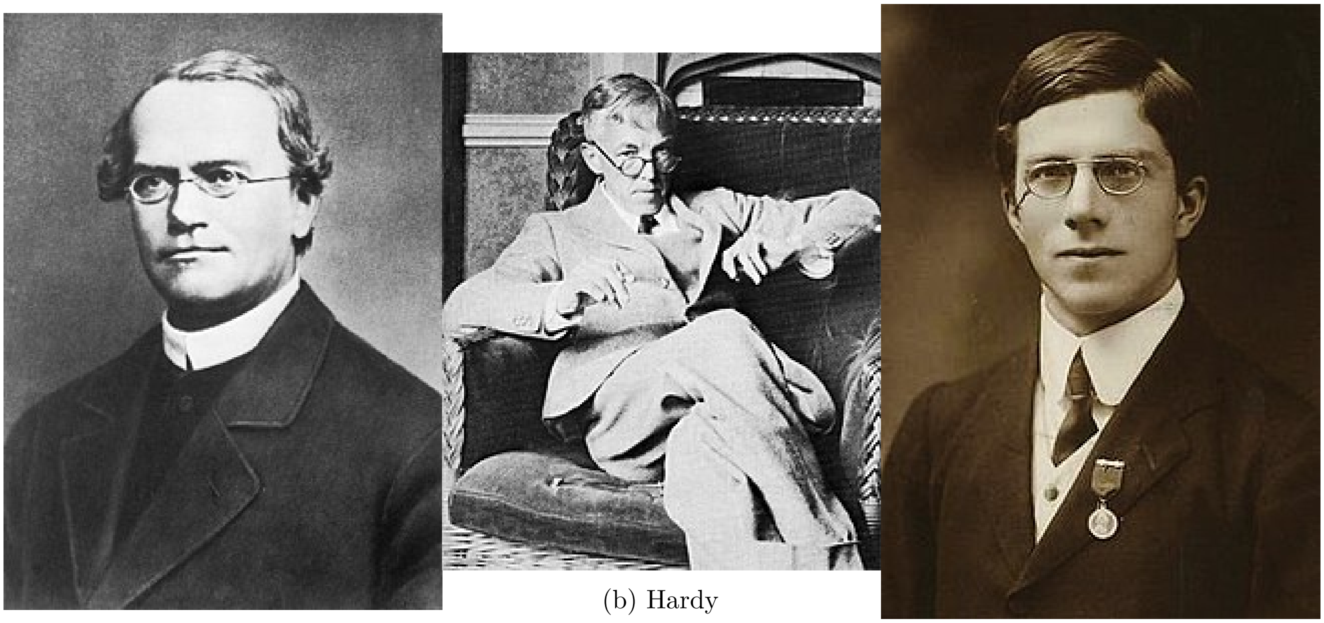
Biologist William Bateson (1861–1926)) coined the terms “genetics” and “allele”.

The Hardy of the Hardy–Weinberg law is G.H. Hardy (1877–1947), the famous mathematical analyst, who published it in 1908.

The statistician R.A. Fisher (1890–1962) made significant contributions to genetics, and much early work in statistics was concerned with genetical problems. A lot of this work contributed to a legacy of eugenics, which was used as a justification for racial discrimination.

The Wright–Fisher model formulates a random model for the evolution of genes in a population with mutation as an *urn model* (Mahmoud 2009, chap. 9).

The deep influence of probability theory on genetics has continued in recent times, with significant developments including the *coalescent* of J.F.C. Kingman.



(a) Mendel

(b) Hardy

(c) Fisher

Figure 5.7: Mendel, Hardy, and Fisher.

References

Anderson, D. F., T. Seppäläinen, and B. Valkó. 2018. *Introduction to Probability.* Cambridge University Press.

Billinton, R., and R. N. Allan. 1996. *Reliability Evaluation of Power Systems.* 2nd ed. Plenum Press.

Blitzstein, J. K., and J. Hwang. 2019. *Introduction to Probability.* Texts in Statistical Science Series. CRC Press.

Boole, G. 1854. *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities.* London: Walton; Maberly.

Chung, K. L., and F. AitSahlia. 2003. *Elementary Probability Theory.* 4th ed. Undergraduate Texts in Mathematics. Springer-Verlag, New York.

DeGroot, M. H., and M. J. Schervish. 2013. *Probability and Statistics.* 4th ed. Harlow, England: Pearson.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications, Vol. 1.* 3rd ed. Wiley, New York.

Hacking, I. 2006. *The Emergence of Probability.* Cambridge University Press.

Hájek, A. 2012. "Interpretations of Probability." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Kolmogorov, A. N. 1950. *Foundations of the Theory of Probability.* New York: Chelsea Publishing Company.

Laplace, P. S. 1825. *Essai Philosophique Sur Les Probabilitiés.* Paris: Bachelier.

Mahmoud, H. M. 2009. *Pólya Urn Models.* CRC Press, Boca Raton, FL.

Moivre, A. de. 1756. *The Doctrine of Chances: Or, a Method for Calculating the Probabilities of Events in Play.* Third. London: A. Millar.

Rosenthal, J. 2007. *A First Look at Rigorous Probability Theory.* Second. New York: World Scientific.

Ross, S. M. 2010. *Introduction to Probability Models.* 10th ed. Academic Press, Amsterdam.

Stirzaker, D. 2003. *Elementary Probability.* Second. Cambridge University Press.

Todhunter, I. 2014. *A History of the Mathematical Theory of Probability.* Cambridge University Press.

Venn, J. 1888. *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability, with Especial Reference to Its Application to Moral and Social Science.* Third. London: Macmillan.

Whitworth, W. A. 1901. *Choice and Chance.* Third. Cambridge: Deighton Bell.