Probability Theory B

From Random Experiments to Natural Language Processing – by Wilker Aziz. This book has not been published yet.

Suppose we are interested in studying our uncertainty about rolling several numbers using two 6-sided dice (see Figure B.1). The dice available to us are identical in their physical properties and their faces are perfectly symmetric such that, if we rolled them, they would each land showing one and exactly one of their faces, independently of one another, and with no specific preference for any one of the possibilities—dice like that are said to be *fair*. For this study, we design a *random experiment* in which we roll the two fair dice at the same time and do not attempt to control the result in any way, we then record the sum of the numbers they show. Figure B.2 shows all possible results of this experiment.

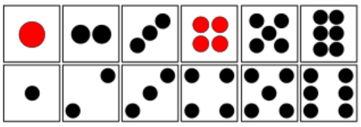

Figure B.1: A 6-sided die (plural: dice) is a cubic object with faces numbered as shown (top: Asian-style; bottom: Western-style). — By Nanami Kamimura, Derivative work: SiPlus, CC BY-SA 3.0, via Wikimedia Commons.

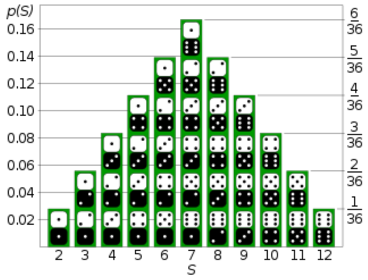

Figure B.2: Probability space associated with the sum of two fair 6-sided dice. — By Tim Stellmach, Own work, Public domain, via Wikemidia Commons.

Each pair of faces in the figure is what we call an *outcome*, together all 36 pairs form this experiment's *sample space*. In this experiment, we care about a specific property of outcomes, namely, the sum of the numbers we rolled—the possible values of this property are shown on the horizontal axis. A set that contains all outcomes in the sample space sharing a specific value of the property of interest is what we call an *event*—events are shown as stacks of outcomes over the ticks of the horizontal axis. We say we have *observed an event* if we have observed any one of its outcomes as a result of random experimentation—also called a random *draw* or *trial*. The vertical axes display the likelihood of observing each event expressed as the number of ways in which it can be obtained (*i.e.*, the number of outcomes in the event) out of the total number of outcomes possible in this experiment (*i.e.*, the size of the sample space). These quantities are what we call *probabilities*—a numerical description of our uncertainty about events. In this context, one can interpret the probability of an event as the *sample frequency* with which we would observe it should we repeat the experiment indefinitely.[1]

This chapter is about probability theory, which deals with the formal foundations of how uncertainty can be quantified. Probability theory is not concerned with giving an interpretation to probability (this is a concern of philosophy) nor does it give us mechanisms to derive probability values from experience or observations about reality (this is a concern of statistics), rather it gives this concept (*i.e.*, probability) a rigorous treatment in terms of a small but powerful set of axioms. Probability theory provides solid mathematical foundations for statistics, decision theory, statistical mechanics, and, of course, machine learning, data analysis, and applications thereof.

ILOs

After completing this chapter, you should be able to

- define a probability space
- calculate probability queries
- explain random variables and probability distributions

[1] Some people find the idea of infinite repetitions of an experiment rather unintuitive and/or hard to generalise to events that do not repeat easily or at all (*e.g.*, an observation about today, our sun exploding, *etc.*). They prefer to think of probability as one's personal quantification of belief given the information accessible to oneself (*e.g.*, the dice are fair, the experimenter does not interfere with the results, *etc.*). *Probabilities as frequencies* and *probabilities as personal beliefs* are the two most common interpretations of probability, but they are not the only two. Fortunately, probability theory does not depend on the interpretation we give to probability, so we can be pragmatic and pick the interpretation that better suits the application scenario.

B.1 Probability Spaces

Probability theory is about assigning probability to subsets of elements of a special set, which we call **sample space** of a random experiment. This section will concentrate on countable sets (often referred to as *discrete sets*), but the theory can be extended to include uncountable sets too.

We use Ω to denote a sample space, and $\omega \in \Omega$ to denote members of that space. A member is usually referred to as an **outcome** or a **sample**.

A set is an unordered collection of elements. Examples: the countably finite set of 'strictly positive odd numbers smaller than 10' $\{1, 3, 5, 7, 9\}$, the countably infinite set of 'natural numbers' $\mathbb{N}$, and the uncountable set of 'real numbers' $\mathbb{R}$.

The notation $\omega \in \Omega$ says that the element ω is a member of the set Ω. The symbol $\notin$ indicates the opposite of that.

Example B.1.1

Examples of random experiments and their sample spaces:

- in a coin toss, the coin can land showing 'heads' (H) or 'tails' (T), thus $\Omega = \{\mathrm{H}, \mathrm{T}\}$;
- a 6-sided die can roll an integer from 1 (included) to 6 (included), thus $\Omega = \{1, 2, 3, 4, 5, 6\}$;
- rolling a 10-sided die followed by rolling a 6-sided die yields pairs of numbers where the first number ranges from 1 (included) to 10 (included) and the second number ranges from 1 to 6, thus $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 10\} \times \{1, 2, 3, 4, 5, 6\}$.

In the last example, we use a shortcut, namely, rather than writing down the 60 pairs of integers in the sample space, we use the Cartesian product of the sample spaces of the individual die rolls.

The Cartesian product of sets A and B, denoted $A \times B$ and defined as $\{(a, b) : a \in A \text{ and } b \in B\}$, is the set of all pairs (a, b), where $a \in A$ and $b \in B$. The definition can be extended to more than two sets in an analogous way.

Exercise B.1.1

What's the sample space associated with drawing 1 card from a standard 52-playingcard deck?

In a sample space we enumerate individual outcomes of a random experiment (*e.g.*, the possible results of a die roll). However, rather than outcomes themselves, we often care about properties of these outcomes. For example, we may be interested in whether the outcome of a die roll is an even number. We call an **event** any set of outcomes that is a subset of the sample space.

The notation $A \subseteq \Omega$ denotes inclusion, that is, A is a set whose members are elements of Ω. We say that A is a subset of Ω or, equivalently, that Ω contains A.

If Ω is a sample space, any subset $A \subseteq \Omega$ is an event with respect to Ω.

Example B.1.2

Examples of events:

- roll a 6-sided die and get 1: $\{1\}$;
- toss a coin and get whatever outcome: $\{\mathrm{H}, \mathrm{T}\}$ — this event happens to be the entire sample space for the coin toss;
- roll a six-sided die and get an even number bigger than 2: $\{4, 6\}$.
- roll a six-sided die and get a 10: $\emptyset$ — a 10 is impossible (see Figure B.1), the empty set denotes such impossible events.

The empty set denoted $\emptyset$ or $\{\}$ is the unique set having no elements in it. For any set A, denoted $\forall A$, i) $\emptyset \subseteq A$—the empty set is a subset of A, ii) $A \cup \emptyset = A$—the union of A with the empty set is A, iii) $A \cap \emptyset = \emptyset$—the intersection of A with the empty set is empty, iv) $A \times \emptyset = \emptyset$—the Cartesian product of A with the empty set is empty.

Exercise B.1.2

Represent the following events for the sample space of Exercise B.1.1:

1. pick an 'A' of $\diamondsuit$;
2. pick an 'A';
3. pick a $\diamondsuit$.

An **event space** associated with a sample space Ω is a set $\mathscr{A}$ such that:

a. $\Omega \in \mathscr{A}$;
b. if $A \in \mathscr{A}$, then $\Omega \setminus A \in \mathscr{A}$;
c. if $A, B \in \mathscr{A}$, then $A \cup B \in \mathscr{A}$.

The relative complement of A in Ω, denoted by $\Omega \setminus A$ or $\Omega - A$, and also called the set-theoretic difference of Ω and A, is the set of all elements that are members of Ω, but not members of A.

These axioms (*i.e.*, properties that must hold) may seem a bit arbitrary at first, but they are quite reasonable, let's reword them here:

a. the sample space is an event in the event space;
b. if we can observe A, then it must be possible to observe the complement of A in Ω—we say $\mathscr{A}$ is *closed under complementation*;
c. if we can observe A and we can observe B, then it must be possible to observe their union—we say $\mathscr{A}$ is *closed under countable unions*;

The union of A and B, denoted by $A \cup B$, is the set of all things that are members of A or of B or of both.

The intersection of A and B, denoted by $A \cap B$, is the set of all things that are members of both A and B.

These properties also have two subtle consequences, namely,

d. if $A, B \in \mathscr{A}$, then $A \cap B \in \mathscr{A}$—we say $\mathscr{A}$ is *closed under countable intersections*;
e. by axiom (b) $\Omega \setminus A \in \mathscr{A}$, by axiom (a) $\Omega \in \mathscr{A}$, thus making $A = \Omega$ it follows that $\emptyset \in \mathscr{A}$.

The intersection $A \cap B$ can be *paraphrased* as $\Omega \setminus ((\Omega \setminus A) \cup (\Omega \setminus B))$. This is an instance of DeMorgan's laws from set theory.

The former can be shown by paraphrasing set intersection in terms of operations like unions and complements. Properties (b–d) state what operations take events as inputs and return events as outputs. Their significance may not be immediately obvious, but soon we will design a special function whose domain is the event space, then it will be convenient to know that complement, union and intersection can never take us out of the event space.

In this course, we will always make the implicit assumption that for a sample space of interest Ω, a valid event space $\mathscr{A}$ exists. For countable sample spaces, in particular, we will always take the event space to be the powerset of the sample space, denoted $\mathscr{P}(\Omega)$.

The powerset of a countable set Ω, denoted $\mathscr{P}(\Omega)$, is the set of all subsets of Ω, which thus includes the empty set and Ω itself. For any discrete sample space Ω, the power set $\mathscr{P}(\Omega)$ always satisfies the conditions of a valid event space.

We can now develop the so called **probability measure**. Start by associating an event space $\mathscr{A}$ with a sample space Ω, then a probability measure $\mathbb{P} : \mathscr{A} \longrightarrow \mathbb{R}$ maps each event in the event space to a real number which we call a **probability**. For a probability measure, it must hold that:

1. $\mathbb{P}(A) \geq 0$ for all $A \in \mathscr{A}$;
2. $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$ for a countable collection $\{A_1, \ldots, A_n\} \subseteq \mathscr{A}$ of pairwise disjoint events;
3. $\mathbb{P}(\Omega) = 1$.

In mathematics, a measure is a function that maps elements from a set of sets (such as events in an event space) to real numbers and for which crucial formal properties hold. The notion of measure generalises common notions such as length, area, volume, and, of course, probability. In some textbooks, the probability measure $\mathbb{P}$ is denoted Pr, which is possibly more convenient for a handwritten essay.

Let's digest those:

1. the smallest probability value attainable by any event is 0;
2. the total probability assigned to n pairwise disjoint events is the sum of the probability values assigned to each of the n events;
3. the event which is the set of all possible outcomes is assigned a total probability value of 1.

If $A_i \cap A_j = \emptyset$, then A_i and A_j have no elements in common and are said to be disjoint sets.

Suppose we are rolling a 6-sided die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. There are many events we can be interested in:

- 'rolling 1' represented by the set $\{1\}$;
- 'rolling 2' represented by the set $\{2\}$;
- 'rolling either 1 or 2' represented by the set $\{1, 2\}$;
- 'rolling an odd number' represented by the set $\{1, 3, 5\}$;
- 'rolling more than 2' represented by the set $\{3, 4, 5, 6\}$;
- 'rolling whatever' represented by the set $\{1, 2, 3, 4, 5, 6\}$, which is exactly equivalent to the complete sample space Ω;
- *etc.*

There are so many things about this sample space that we may be interested in, that we just assume we may potentially be interested in any event in $\mathscr{P}(\Omega)$. We can then talk about a probability measure $\mathbb{P}$ for the event space, whatever probability values this measure assigns to the events, this measure must be such that:

1. the smaller probability value assigned to any single event in the event space is 0;
2. if we take disjoint events such as 'rolling 1' $\{1\}$ and 'rolling 2' $\{2\}$, the total probability value assigned to the event 'rolling 1 or 2' $\{1, 2\}$ is the sum of probability values $\mathbb{P}(\{1, 2\}) = \mathbb{P}(\{1\} \cup \{2\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{2\})$;
3. the event 'rolling whatever' has probability $\mathbb{P}(\Omega) = 1$.

The probability measure is a positive measure (*i.e.*, no event can have negative probability) bounded such that the probability of the universe (*i.e.*, the event that consists of the entire sample space) is 1.

It may not look obvious at first, but properties 1–3 imply that:

4. $\mathbb{P}(\emptyset) = 0$;
5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for events $A, B \in \mathscr{A}$;
6. $\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A)$ for an event $A \in \mathscr{A}$.

Let's digest those too:

4. says that the empty event has probability 0, that is, if we *observed* a random experiment something must have happened;
5. is a generalisation of property (2) which does not require disjoint events;
6. is also called the *complement rule*.

The inclusion-exclusion principle is a technique which generalises property (2) and extends it to more than 2 events.

We now define the concept of a **probability space**. A probability space is a triple $(\Omega, \mathscr{A}, \mathbb{P})$ consisting of a sample space Ω, an event space $\mathscr{A}$, and a probability measure $\mathbb{P}$.

Example B.1.3

Here we present the probability space of a *fair coin flip*.

A coin can land heads (H) or tails (T), thus the sample space for the probability space of this random experiment is $\Omega = \{\mathrm{T}, \mathrm{H}\}$.

For the event space $\mathscr{A}$ we will use every possible subset of Ω, that is, the powerset of Ω. That is, $\mathscr{A} = \{\emptyset, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}\}$.

Probability theory does not tell us where probability values come from, it only prescribes the formal properties of probability spaces. There are infinitely many probability measures that lead to valid probability spaces for an arbitrary experiment involving coin flips (we will describe one now, and another one in the next example), all those measures have one thing in common, they comply with axioms 1–3.

Fortunately, we are concerned with a very specific coin flip, namely, the flip of a *fair coin*. Because the coin is fair, we can establish the probability of an event using the event's *cardinality*—the number of outcomes in it—relative to the cardinality of the sample space. That is, for an event $A \in \mathcal{A}$, it holds that $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$. Note that this definition cannot be used for sets of infinite size—for those we will develop other analytical tools later on.

The cardinality of a countably finite set A, denoted $|A|$, is the number of elements in it. The cardinality of the empty set is defined to be 0. Examples: $|\emptyset| = 0$, $|\{10\}| = 1$, $|\{\text{red}, \text{blue}\}| = 2$.

Now that we have a mechanism to assign probability to events, we can characterise the entire probability measure. For example, the event 'coin lands heads', denoted $\{\mathrm{H}\}$, is a set that contains a single outcome—its size or cardinality is 1—thus its probability in this probability space is $|\{\mathrm{H}\}|/|\Omega| = 1/2$. The complete probability measure for this probability space is shown in Table B.1.

Table B.1: Probability measure for a fair coin flip.

$A \in \mathcal{P}(\Omega)$	$\mathbb{P}(A)$
$\emptyset$	0
$\{\mathrm{H}\}$	$1/2$
$\{\mathrm{T}\}$	$1/2$
$\{\mathrm{H}, \mathrm{T}\}$	1

Note that, to specify Table B.1, we did not have to refer back to the axioms 1–3 (nor properties 4–6), but the resulting measure, as it turns out, complies with them. That is so because the function that assigns probability $\frac{|A|}{|\Omega|}$ to $A \in \mathcal{P}(\Omega)$ prescribes a valid probability measure, so long as the sample space is countably finite. This does not mean, however, that this measure is always appropriate to describe the random experiment of interest, as the next example shall demonstrate.

Example B.1.4

Here we present the probability space of a *crooked coin flip*, this particular coin was designed to land heads twice as often as it lands tails (*i.e.*, the odds of landing heads is 2 : 1).

Odds convey information about the likelihood of a particular outcome in gambling. Odds of an outcome ω are denoted $n : m$ (and pronounced *n to m*), where n is the number of events that produce that outcome, and m is the number of events that do not. We can convert the odds of ω to the probability of the event $\{\omega\}$ via $\frac{n}{n+m}$. Example rolling a fair 6-sided die: the odds of rolling a 6 is 1 : 5.

As before, we have a coin flip, thus $\Omega = \{\mathrm{T}, \mathrm{H}\}$. As before we use the powerset of Ω as the event space $\mathcal{A} = \{\emptyset, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}\}$.

Unlike before, this coin is crooked, so the measure that assigns probability $\frac{|A|}{|\Omega|}$ to an event $A \in \mathcal{A}$ is of no help here. We do, however, have access to a key piece of information, which, when combined with the axioms of probability 1–3 (and properties 4–6), allows us to work out the probability measure for this probability space.

The information we have access to is the odds of obtaining heads, which is 2 : 1. This means the event 'landing heads' $\{\mathrm{H}\}$ has probability $\mathbb{P}(\{\mathrm{H}\}) = 2/3$. Using axiom 2, we can write $\mathbb{P}(\{\mathrm{H}, \mathrm{T}\}) = \mathbb{P}(\{\mathrm{H}\}) + \mathbb{P}(\{\mathrm{T}\})$, we also know, from axiom 3, that $\mathbb{P}(\{\mathrm{H}, \mathrm{T}\}) = 1$ for $\{\mathrm{H}, \mathrm{T}\}$ is the entire sample space Ω. This means that $1 = 2/3 + \mathbb{P}(\{\mathrm{T}\})$, and thus $\mathbb{P}(\{\mathrm{T}\}) = 1/3$ (property 6 would have been a shortcut to derive the same result). With little work left, we can justify the probability measure in Table B.2.

Table B.2: Probability measure for a crooked coin flip with odds 2 : 1 for heads.

$A \in \mathcal{P}(\Omega)$	$\mathbb{P}(A)$
$\emptyset$	0
$\{\mathrm{H}\}$	$2/3$
$\{\mathrm{T}\}$	$1/3$
$\{\mathrm{H}, \mathrm{T}\}$	1

As the last two examples demonstrate, there are different strategies to working out the probability measure that best describes a given random experiment. And—one can never stress it enough—probability theory does not tell us which probability measure we should use, it only tells us what properties must hold for every probability measure. It is on us to find in the application domain (or in the description of a given random experiment) the information that will constrain us to an appropriate choice. Next we will develop some powerful tools to manipulate probability measures, these tools will power a framework for the compact specification of very complex probability spaces.

Exercise B.1.3

Describe the probability space of two *fair coin* flips. Do assume the experimenter has no interest in interfering with the experiments and the coins land completely independently of one another. For practice, evaluate the probability measure for every event in the powerset of the sample space.

The probability of a countable union of events $A_1, \ldots, A_n$ tells us the probability with which *any* of the events $A_1, \ldots, A_n$ should occur—recall, an event occurs if one of its outcomes occurs. Now we turn to the probability with which *all* of these events should occur. The **joint probability** of a countable set of events $\{A_1, \ldots, A_n\}$ is

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) \, . \tag{B.1}$$

All we need is to evaluate the intersection of the events and, because event spaces were defined very carefully, the intersection is guaranteed to be in the domain of the probability measure.

Example B.1.5

Consider experiments involving *two coin flips*.

Let the sample space be $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$, and the event space be $\mathcal{A} = \mathcal{P}(\Omega)$.

The event 'tossing at least one heads' is the set $A = \{\text{HH}, \text{HT}, \text{TH}\}$. The event 'tossing at least one tails' is the set $B = \{\text{TT}, \text{TH}, \text{HT}\}$. If we are interested in any of the two events, we are interested in any outcome that indicates that A or B or both occurred, that is, we are interested in outcomes in $A \cup B$—an event with probability $\mathbb{P}(A \cup B) = \mathbb{P}(\{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}) = 1$.

If we are interested in scenarios where it can be said that *both* events occurred, then we are interested in any outcome that indicates that *A as well as B* occurred, those can only be outcomes that are shared by both sets, that is outcomes in $A \cap B$—an event with probability $\mathbb{P}(A \cap B) = \mathbb{P}(\{\text{HT}, \text{TH}\})$.

Note that we do not know enough about the coins to specify the numerical values of the probability measure.

In many situations we start from a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and, upon observing an event A_j, we exclude outcomes not in A_j from further consideration. This essentially results in a closely related probability space $(A_j, \mathscr{A}, \mathbb{P}(\cdot|A_j))$ whose probability measure is the so called **conditional probability measure**, which assigns probability

$$\mathbb{P}(A_i|A_j) := \frac{\mathbb{P}(A_i \cap A_j)}{\mathbb{P}(A_j)} \tag{B.2}$$

to an event $A_i \in \mathscr{A}$ conditioned on an event $A_j \in \mathscr{A}$ with $\mathbb{P}(A_j) > 0$.

Example B.1.6

Consider the experiment involving *two coin flips*, Example B.1.5, and use the probability measure shown in Table B.3 (note that missing values can be inferred by using the axioms of probability).

Via axiom (2), the event 'tossing heads first' $H_1 = \{\mathrm{HH}, \mathrm{HT}\}$ has probability $\mathbb{P}(H_1) = \mathbb{P}(\{\mathrm{HH}\} \cup \{\mathrm{HT}\}) = {}^1\!/_3 + {}^1\!/_6 = {}^1\!/_2$. As the probability is greater than zero, we can condition on it and obtain the probability space $(H_1, \mathscr{A}, \mathbb{P}(\cdot|H_1))$ whose conditional probability measure is (partly) listed in Table B.4. The first column lists the events in the event space, the second column is obtained by intersecting an event with H_1, and the third column by application of the definition of conditional probability.

Table B.3: Example of probability measure for two coin flips.

$A \in \mathscr{A}$	$\mathbb{P}(A)$
{HH}	${}^1\!/_3$
{HT}	${}^1\!/_6$
{TT}	${}^1\!/_3$
{TH}	${}^1\!/_6$

Table B.4: Example of probability measure for two coin flips conditioned on 'tossing heads first' H_1.

$B \in \mathscr{A}$	$B \cap H_1$	$\mathbb{P}(B\|H_1)$
{HH}	{HH}	${}^1\!/_3\boldsymbol{\nabla} \cdot {}^1\!/_2 = {}^2\!/_3$
{HT}	{HT}	${}^1\!/_6\boldsymbol{\nabla} \cdot {}^1\!/_2 = {}^1\!/_3$
{TT}	$\emptyset$	0
{TH}	$\emptyset$	0
...	...	...

Exercise B.1.4

Complete the specification of the conditional probability measure in Table B.4 by evaluating it for every event in the event space $\mathscr{A}$.

Last, but not least, two events A_i, A_j are said to be **independent** if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \times \mathbb{P}(A_j)\,. \tag{B.3}$$

Independence of two events is denoted $A_i \perp\!\!\!\perp A_j$ (the notation $A_i \perp A_j$ is also common). Moreover, for $A_i \perp\!\!\!\perp A_j$ and $\mathbb{P}(A_j) > 0$, it holds:

$$\mathbb{P}(A_i|A_j) = \frac{\mathbb{P}(A_i \cap A_j)}{P(A_j)} \overset{\text{indep.}}{=} \frac{\mathbb{P}(A_i) \times \cancel{\mathbb{P}(A_j)}}{\cancel{\mathbb{P}(A_j)}} = \mathbb{P}(A_i)\,. \tag{B.4}$$

Independence will turn out a useful concept when we design probability measures over very complex event spaces.

Exercise B.1.5

Is the event 'tossing two heads' independent of the event 'tossing heads first' in the probability space of the experiment *two coin flips* with probability measure given by Table B.3?

B.2 Discrete Random Variables

You may have noticed two laborious things about probability spaces. First, we need to start from a sample space, even though sometimes we only care about specific properties of outcomes. For example, in an experiment involving drawing 5-hand cards from a standard 52-playingcard deck, all we might care about is the number of cards of club suit ♣. Second, for a probability space $(\Omega, \mathcal{A} = \mathcal{P}(\Omega), \mathbb{P})$, specifying the probability measure roughly requires listing all the events in the event space and working their probabilities out one by one. While this is doable for small sample spaces (*e.g.*, 1 or 2 coin flips), it quickly gets rather tedious (*e.g.*, in role playing games it is not unusual to roll 5 six-sided dice and add some character-specific constants to compute physical damage inflicted in a battle), or essentially impossible (*e.g.*, the length of the byte sequence that represents a file in a modern computer is a natural number, finite, but potentially unbounded). To help create efficient analytical tools that compactly characterise complex probability spaces, we will need some tools to change the interface with which we interact with the basic elements of probability theory.

There was one instance in which we characterised a probability measure with the help of an auxiliary function: when dealing with fair coin flips we told you that the function $|A|/|\Omega|$ specifies a measure that is not only appropriate to capture the properties of the random experiment (*i. e.*, that the coin is not crooked) but also, crucially, that measure properly complies with the axioms of probability theory. This may not look like a lot at first, but think about it. Is it easier to list the $2^4 = 16$ events in the event space $\mathcal{A} = \mathcal{P}(\{\mathrm{H}, \mathrm{T}\} \times \{\mathrm{H}, \mathrm{T}\})$ and their probabilities, or to compute the probability of any event $A \in \mathcal{A}$ on demand by simply storing the relation $\mathbb{P}(A) = |A|/|\Omega|$? Random variables will helps us design tools as convenient as this simple functional relation between the size of the event and the size of the sample space, but for the most diverse probability spaces.

The first thing on our way to these amazing tools is—wait for it—the sample space. In probability theory, the sample space is too fundamental a concept, so this must be bad news. Take a moment to recover from this. A sample space is a set, a fairly general object that hosts whatever we are interested in. This generality is purposeful, it is what allows us to reason about experiments whose outcomes are the most diverse things imaginable, but it stands on the way to other desiderata (*e.g.*, having a simple parametric mechanism to relate outcomes to their probability). The answer to this is to introduce a map from an arbitrary sample space (*e.g.*, 5-card hands drawn from a standard 52-playingcard deck) to a more convenient (numerical) set. A **discrete random variable** X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$

i. is a function $X : \Omega \longrightarrow \mathcal{X}$ from the sample space Ω to a *countable* subset $\mathcal{X}$ of $\mathbb{R}$,
ii. such that, for any $x \in \mathcal{X}$, the set defined as $\{\omega \in \Omega : X(\omega) = x\}$, also denoted $X^{-1}(x)$, is an event in the event space $\mathcal{A}$.

The sample space Ω is also called the domain of the random variable. The subset $\mathcal{X}$ of $\mathbb{R}$ to which the outcomes $\omega \in \Omega$ are mapped to is called the *range* (or image) of the random variable. An element $x \in \mathcal{X}$ is called an outcome of the random variable. The set $X^{-1}(x) \in \mathcal{A}$ is the event mapped to outcome x.[2] This mapping allows us to work on a numerical space ($\mathcal{X} \subset \mathbb{R}$) no matter the nature of the sample space Ω.

[2] Mathematically, for $x \in \mathcal{X}$, $X^{-1}(x)$ is known as the *fiber* over x under X.

For the range, we will generally use a calligraphic counterpart to the letter used for the random variable. Soon we will have experiments with multiple random variables, and, to distinguish sample spaces, we will normally subscript Ω with the name of the random variable associated with it. Examples: the random variable X has domain Ω_X and image $\mathcal{X} \subset \mathbb{R}$, the random variable Y has domain Ω_Y and image $\mathcal{Y} \subset \mathbb{R}$.

Example B.2.1

We have three urns, each contains balls that are blue and/or red. We draw a ball from each urn in sequence.

The sample space is the set of all sequences of size three $\Omega = \{\mathrm{b}, \mathrm{r}\}^3$ ('b' as a short for blue and 'r' as a short for red) where an element of the sequence is either blue or red.

For a sequence $\omega \in \Omega$, we use ω_i to indicate the colour of the

ith ball in the sequence, the number of red balls drawn by the person can be captured by a random variable: $R : \Omega \to \mathbb{R}$ such that $R(\omega) = \sum_{i=1}[\omega_i = \mathrm{r}]$.

The **probability distribution** of a random variable X is denoted by P_X and is defined as the function

$$P_X(X = x) \coloneqq \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}) . \tag{B.5}$$

A *lot* has just happened. First things first. The notation $X = x$ is another way of writing $X^{-1}(x)$ or, equivalently, $\{\omega \in \Omega : X(\omega) = x\}$. The probability distribution assesses the probability of the event that X maps to x. The quantity $P_X(X = x)$ is pronounced 'probability that the random variable X takes on the value x'.

Sometimes we will be interested in the set of outcomes of X to which the probability distribution P_X assigns non-zero probability. This set is called the *support* of the distribution P_X and defined as:

$$\mathrm{supp}(X) = \{x \in \mathcal{X} : P_X(X = x) > 0\} \subseteq \mathcal{X} . \tag{B.6}$$

Random variables are so convenient that we will normally never bother defining the sample space (nor the event space). Instead, we will often define the random variable in a more or less declarative way, rather than through formulae. For example, let R take on the number of red balls in a sequence of balls drawn from three adjacent urns, each urn containing a mixture of blue and red balls.

The **cumulative distribution function** (or cdf for short) of a random variable X is given by

$$F_X(a) \coloneqq P_X(X \leq a) = \sum_{x \leq a} P_X(X = x) . \tag{B.7}$$

Note that, unlike the probability distribution, the cdf takes real values (not events) as inputs. That is so because $F_X(a)$ always evaluates the probability of the event $\{\omega \in \Omega : X(\omega) \leq a\}$, also denoted $X \leq a$ for short.

Finally, we get to the concept that will power the compact laws that we will be using to describe various probability measures. A **probability mass function** (or pmf for short) can be defined for a random variable X as follows:

$$f(x) \coloneqq P_X(X = x) . \tag{B.8}$$

The pmf is defined with a specific probability distribution in mind, and it gets an arbitrary name. The pmf creates a standard real function interface to a probability distribution, allowing us to drop all the notational devices necessary for probability distributions and random variables. Its significance does not come from any perceived brevity of notation, but rather from the fact that we can now prescribe functions (in $\mathbb{R}$) that in turn characterise probability distributions, that in turn characterise probability measures.

Oftentimes, a pmf relates the probability mass of x to the value x itself and some fixed quantity θ called the pmf's **parameter**.

The Iverson bracket, denoted $[\alpha]$ is a compact way to map the result of a boolean predicate α to a real number. It evaluates to 1 when α is true, and to 0, otherwise. For example, with $\omega = \mathrm{rbr}$, $[\omega_1 = \mathrm{r}]$ evaluates to 1, but $[\omega_2 = \mathrm{r}]$ evaluates to 0.

It is also common to use P for the probability distribution of the random variable X. As we did before, for sample spaces, we subscript P with X (as in P_X) for clarity.

Notation matters. • In addition to the clear (albeit lengthy) notation $P_X(X = x)$, we find in the litearture many uses of $P(X = x)$, $P_X(x)$ and $P(x)$. The last three can be confusing without additional information. Of the three briefer options, $P(X = x)$ is the least ambiguous. The subscript X in P_X helps us remember what random variable this distribution is the probability distribution of, which becomes particularly helpful when manipulating multiple random variables. • A random variable is a function and when we write $X = x$ we are in fact instantiating an event (*i.e.*, a set of outcomes from the sample space), if we drop the '$X =$' part (as in the briefer forms $P_X(x)$ and $P(x)$) it is hard to tell that we mean to provide the event $X^{-1}(x) = \{\omega \in \Omega : X(\omega) = x\}$, as opposed to the real number x, as an argument of the probability measure (if you are used to programming languages: P_X(real number) would raise a 'type error', since probability measures take *events* built from Ω as arguments). Consider the 3-urns example, $R = 2$ in that probability space means $\{\mathrm{rrb}, \mathrm{rbr}, \mathrm{brr}\}$, whereas 2 is just a number. • A notation like $P(x)$ does not help remember the random variable name (*i.e.*, we dropped the subscript X) and violates the definition of the probability measure (since its argument should be an event), hence this is the least preferred option. In this book, we will use the notation $P_X(X = x)$ to evaluate the probability of the event $X = x$ under the probability distribution P_X of the random variable X.

The pmf can be denoted by $p(x)$, another lowercase letter (*e.g.*, $f(x)$), a Greek letter (*e.g.*, $\pi(x)$), a person's name (*e.g.*, Bernoulli(x)), amongst many other ways. We may subscript it with X for clarity (*e.g.*, $p_X(x)$, $f_X(x)$, $\pi_X(x)$). Unlike the probability distribution P_X, the pmf takes real values as arguments (*i.e.*, its domain is the random variable's image $\mathcal{X} \subset \mathbb{R}$).

The parameter of the pmf is sometimes denoted explicitly, common choices are $f(x|\theta)$, $f(x;\theta)$ and $f_\theta(x)$, but sometimes it is also simply assumed clear from context.

Example B.2.2

The uniform probability mass function $\text{Uniform}(x|N)$ assigns probability mass $\frac{[x\in\{1,\ldots,N\}]}{N}$ to $x \in \{1, \ldots, N\}$ and 0 to any other value.

Why do we need pmfs when we have distributions? The pmf is associated with a given random variable, so we can stop carrying the notation for rvs around. The pmf also helps us specify the probability distribution by a compact parametric function, and, because we can name the pmf, we can use names that help us remember that functional form. See examples of pmfs in Chapter C, they nicely illustrate this point.

Example B.2.3

The Bernoulli distribution is the distribution of a random variable X that takes on values in $\{0, 1\} \subset \mathbb{R}$ with probability mass function:

$$\text{Bernoulli}(x|\theta) = \theta^x(1-\theta)^{1-x} \,. \tag{B.9}$$

Example B.2.4

The Geometric distribution is the distribution of a random variable X that takes on values in $\mathbb{N}_0 \subset \mathbb{R}$ with probability mass function:

$$\text{Geometric}(x|\theta) = (1-\theta)^{1-x}\theta \,. \tag{B.10}$$

It can be used to model the number of failures until a success.

We can use subscripts to name pmfs in a self-evident way without having to introduce new names/letters. For example, we can use $f_X(x)$ and $f_Y(y)$ for the pmfs of X and Y, respectively. We can use $f_{Y|X=x}(y)$ to denote the pmf that prescribes the conditional distribution $P_{Y|X=x}$ of the random variable Y given $X = x$. It is also common to use $f_{Y|X}(y|x)$, though this is a slight (but intelligible) abuse of the conditioning notation.

The elementary results from probability theory all extend to random variables.

Random variables $X_1, \ldots, X_N$, which we abbreviate as X_1^N are said to be jointly distributed with probability distribution $P_{X_1^N}$ if

$$P_{X_1^N}(X_1^N = x_1^N) = \mathbb{P}(\{\omega \in \Omega : X_1(\omega) = x_1, \ldots, X_N(\omega) = x_N\}) \,. \tag{B.11}$$

Implicit in this definition is the fact that these random variables must share an underlying sample space, and that some underlying event space exists. In applications, we usually do not care about the formalities of the sample space, we only care about quantities that we can capture with random variables. Random variables allow us to take these great shortcuts without running the risk of compromising the formal validity of our probability spaces, but also without having to carefully specify their every aspect.

From the joint distribution, we can also recover the distribution of the individual variables by an operation called **marginalisation**. Suppose we have two random variables X and Y which are jointly distributed with probability distribution P_{XY}, we can obtain marginal probabilities by 'summing away' the possible values of one of the random variables:

$$P_X(X = x) = \sum_{y\in\mathcal{Y}} P_{XY}(X = x, Y = y) \,. \tag{B.12}$$

Let X and Y be random variables with joint distribution P_{XY}, the probability of $X = x$ conditioned on $Y = y$ is given by

$$P_{X|Y}(X = x|Y = y) = \frac{P_{XY}(X = x, Y = y)}{P_Y(Y = y)} \tag{B.13}$$

and the conditional distribution of X given $Y = y$ is denoted by $P_{X|Y=y}$.

Two random variables X and Y with joint distribution P_{XY} are said to be independent (denoted by $X \perp\!\!\!\perp Y$) if $\forall x \in \text{supp}(X)$ and $\forall y \in \text{supp}(Y) : P_{XY}(X = x, Y = y) = P_X(X = x)P_Y(Y = y)$, also denoted $P_{XY} = P_X P_Y$.

We close this section with two important results that will help us design probability distribution for complex random experiments and infer the result of probability queries.

The **chain rule** of probabilities:

$$\begin{aligned} P_{XY}(X = x, Y = y) &= P_X(X = x)P_{Y|X}(Y = y|X = x) &\text{(B.14a)}\\ &= P_Y(Y = y)P_{X|Y}(X = x|Y = y) &\text{(B.14b)}\end{aligned}$$

and by induction

$$P_{X_1^N}(X_1 = x_1, \dots, X_N = x_N) = \prod_{n=1}^{N} P_{X_n|X_{<n}}(X_n = x_n|X_{<n} = x_{<n}) \,, \tag{B.15}$$

where $X_{<n}$ is the (possibly empty) sequence preceding X_n. The order of enumeration is arbitrary.

The **Bayes rule** is an application of conditional probability to infer, for example, $P_{X|Y}(X = x|Y = y)$ from a joint distribution $P_{XY} = P_X \times P_{Y|X}$, where P_X and $P_{Y|X}$ are known. The result is as follows:

All results presented for probability distributions extend to probability mass functions of the corresponding random variables. That is, if we denote the pmf that prescribes the joint distribution of X and Y by $f_{XY}(x, y)$, there exist a collection of pmfs of the form $f_X(x)$, $f_Y(y)$, $f_{Y|X=x}(y)$ and $f_{X|Y=y}(x)$, such that $f_{XY}(x, y) = f_X(x)f_{Y|X=x}(y) = f_Y(y)f_{X|Y=y}(x)$.

$$P_{X|Y}(X = x|Y = y) = \frac{P_{XY}(X = x, Y = y)}{P_Y(Y = y)} \,, \tag{B.16a}$$

which follows directly from the definition of conditional probability, then we factorise the joint distribution in the numerator exploiting the fact that we know the factors P_X and $P_{Y|X}$:

$$= \frac{P_X(X = x) \times P_{Y|X}(Y = y|X = x)}{P_Y(Y = y)} \,, \tag{B.16b}$$

next, we rewrite the denominator using marginal probability

$$= \frac{P_X(X = x) \times P_{Y|X}(Y = y|X = x)}{\sum_{x' \in \mathcal{X}} P_{XY}(X = x', Y = y)} \,, \tag{B.16c}$$

and factorise the joint probability in the sum via chain rule

$$= \frac{P_X(X = x) \times P_{Y|X}(Y = y|X = x)}{\sum_{x' \in \mathcal{X}} P_X(X = x') \times P_{Y|X}(Y = y|X = x')} \,. \tag{B.16d}$$

As, by assumption, we can assess $P_X(X = x)$ and $P_{Y|X}(Y = y|X = x)$ for any x and y, the result allows us to *invert* the conditional $P_{Y|X}$.

B.3 Continuous RVs

All of probability theory extends to events defined on *uncountable* sample spaces. Good examples include continuous measurements (*e.g.*, outside temperature, distance, volume, intervals of time), proportions (*e.g.*, percentage of voters), amongst many others. As before, we have a sample space Ω, this time an *uncountable* set. We also need an event space $\mathcal{A}$ made of special subsets of Ω that can be regarded as events, the precise definition is beyond the scope of this book. For the event space, we will use the so-called Borel σ-algebra of the sample space, denoted $\mathcal{B}(\Omega)$. Think of this as a generalisation of the powerset construction for uncountable sets.[3] Finally, we have a probability measure $\mathbb{P} : \mathcal{A} \to [0, 1]$ assigning probability to events in $\mathcal{A}$. The probability measure complies with the same axioms as before.

A **continuous random variable** X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$

i. is a function $X : \Omega \to \mathcal{X}$ from the (uncountable) sample space Ω to an *uncountable* subset $\mathcal{X}$ of $\mathbb{R}$, where again $\mathcal{X} \subseteq \mathbb{R}$ is the *range* of the random variable,
ii. such that, for any set $B \in \mathcal{B}(\mathcal{X})$, the set defined as $X^{-1}[B] = \{\omega \in \Omega : X(\omega) \in B\}$ is an event in the event space $\mathcal{B}(\Omega)$.[4]

An element $x \in \mathcal{X}$ is called an outcome of the random variable. $B \in \mathcal{B}(X)$ is a subset in the range of the random variable (*e.g.*, an interval such as $(0, 1)$, or the union of disjoint intervals such as $(0, 1) \cup (2, 3)$),the notation $X \in B$, which we pronounce 'the random variable X takes on a value in B' is a shorthand for the event $\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}$.

In the continuous case, a single outcome $X = x$ always has probability 0 (we say singletons are not measurable), while sets of outcomes $X \in B$ may have non-zero probability (then we say they are measurable). An outcome $x \in \mathcal{X}$ can be assigned what we refer to as *probability density*, a non-negative quantity used in combination with Lebesgue integration to quantify the probability of events. A probability density function (pdf) $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$ is such that $\int_{\mathcal{X}} f(x)\mathrm{d}x = 1$. With it, we can characterise the probability of $X \in B$:

$$P_X(X \in B) = \int_B f(x)\mathrm{d}x \,. \tag{B.17}$$

When B is a continuous interval $(a, b) \in \mathbb{R}$, we have $P_X(X \in B) = \int_a^b f(x)\mathrm{d}x$. Instead, when B is a countable union of disjoint intervals $\bigcup_{i \in I}(a_i, b_i)$, we have $P_X(X \in B) = \sum_{i \in I} \int_{a_i}^{b_i} f(x)\mathrm{d}x$. The pdf can also be characterised in terms of a cumulative distribution function $F_X(x)$:

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) \,. \tag{B.18}$$

When B is a continuous interval $(a, b) \in \mathcal{X}$, we have $P_X(X \in B) = F_X(b) - F_X(a)$. Instead, when B is a countable union of disjoint intervals $\bigcup_{i \in I}(a_i, b_i)$, we have $P_X(X \in B) = \sum_{i \in I} F_X(b_i) - F_X(a_i)$.

As always, our goal is to prescribe a probability measure that complies with the axioms of probability theory, pdfs and cdfs are devices that help us achieve that goal. Whether we start by specifying a cdf and compute

[3] To give some intuition (though not exactly accurate), here is a hypothetical way to construct it: i) start with all open intervals $A_1, A_2, \ldots$ of Ω, ii) then gather all sets obtainable via union or intersection of any countable subset of those.

[4] Mathematically, for $B \subseteq \mathcal{X}$, $X^{-1}[B]$ is known as the *preimage* of B under X.

the pdf by differentiation, or start by specifying a pdf and compute the cdf by integration is mostly a matter of practical convenience. So long as we specify these devices coherently, we will be prescribing valid probability measures for random experiments involving continuous random variables.

Example B.3.1

The Exponential distribution is the distribution of a random variable X that takes on values in $\mathbb{R}_{>0}$ with probability density function:

$$\text{Exponential}(x|\theta) = \theta \exp(-\theta x)\ . \tag{B.19}$$

Its parameter is strictly positive (*i.e.*, $\theta > 0$) and is known as *rate*. By definition, the Exponential pdf assigns 0 density to any number $x \leq 0$. This model can be used to describe the time between events that occur continuously and independently at a constant average rate.

The Exponential cdf is known in closed-form: $F_X(x) = 1 - \exp(-\theta x)$, a result which we can verify by expressing the derivative of $F_X(x)$ with respect to x.

Example B.3.2

The Beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$ is the distribution of a random variable X that takes on values in $(0, 1)$ with probability density function:

$$\text{Beta}(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}\ , \tag{B.20}$$

where $\mathrm{B}(\alpha, \beta)$ is the Beta function is a generalisation of binomial coefficients. This model can be used to describe random proportions (or percentages).

There is no simple form for the cdf of the Beta distribution, but most computations involving it can be reliably approximated by computers.

Example B.3.3

The Normal distribution with location $\mu \in \mathbb{R}$ and scale $\sigma \in \mathbb{R}_{<0}$ is the distribution of a random variable X that takes on values in $\mathbb{R}$ with probability density function:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\ . \tag{B.21}$$

The Normal is a good model for random quantities that distributed symmetrically around a central tendency. The Normal location is also its mean, while the squared of the scale is its variance (making the scale its standard deviation). Another name for the Normal distribution is the *Gaussian* distribution. There is no simple form for the cdf of the Normal distribution, but most computations involving it can be reliably approximated by computers.

B.4 Solutions

Exercise B.1.1

$\{A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K\} \times \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$

Listing B.1: Typesetting suits in LaTeX.

- `$\clubsuit$` for $\clubsuit$;
- `$\diamondsuit$` for $\diamondsuit$;
- `$\heartsuit$` for $\heartsuit$;
- `$\spadesuit$` for $\spadesuit$.

Exercise B.1.2

1. $\{\mathrm{A}\diamondsuit\}$;
2. $\{\mathrm{A}\diamondsuit, \mathrm{A}\heartsuit, \mathrm{A}\spadesuit, \mathrm{A}\clubsuit\}$;
3. $\{2\diamondsuit, 3\diamondsuit, 4\diamondsuit, 5\diamondsuit, 6\diamondsuit, 7\diamondsuit, 8\diamondsuit, 9\diamondsuit, 10\diamondsuit, \mathrm{J}\diamondsuit, \mathrm{Q}\diamondsuit, \mathrm{K}\diamondsuit, \mathrm{A}\diamondsuit\}$

Listing B.2: Typesetting sets in LaTeX.

- `$\{\}$` for $\{\}$;
- `$\left\{\frac{1}{10}\right\}$` for $\left\{\frac{1}{10}\right\}$.

Exercise B.1.3

For a complete probability space we have to specify a sample space Ω, an event space $\mathscr{A}$, and a probability measure $\mathbb{P}$.

For a single coin flip, the sample space would be $\{\mathrm{H}, \mathrm{T}\}$, hence for two coin flips with have $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TT}, \mathrm{TH}\}$.

As the sample space is countable, we can use its powerset as the event space $\mathscr{A} = \mathscr{P}(\Omega)$. In this way, all subsets of Ω, including the empty set and Ω itself, are valid events.

For probability measure we can use any function from $\mathscr{A}$ to $\mathbb{R}$ that satisfies the basic axioms of probability. For this exercise in particular, we have enough information to give a precise characterisation of the probability measure of interest. As the coins are both fair, land independently of one another, and the experimenter does not interfere with trials, we can obtain the probability value of an event by expressing its cardinality relative to the size of the sample space, as shown in Table B.5.

Table B.5: Probability measure for two independent fair coin flips.

$A \in \mathscr{A}$	$\mathbb{P}(A)$
$\{\}$	0
$\{\mathrm{HH}\}$	1/4
$\{\mathrm{HT}\}$	1/4
$\{\mathrm{TT}\}$	1/4
$\{\mathrm{TH}\}$	1/4
$\{\mathrm{HH}, \mathrm{HT}\}$	1/2
$\{\mathrm{HH}, \mathrm{TH}\}$	1/2
$\{\mathrm{HH}, \mathrm{TT}\}$	1/2
$\{\mathrm{HT}, \mathrm{TH}\}$	1/2
$\{\mathrm{HT}, \mathrm{TT}\}$	1/2
$\{\mathrm{TH}, \mathrm{TT}\}$	1/2
$\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}\}$	3/4
$\{\mathrm{HH}, \mathrm{HT}, \mathrm{TT}\}$	3/4
$\{\mathrm{HH}, \mathrm{TH}, \mathrm{TT}\}$	3/4
$\{\mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$	3/4
$\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$	1

Exercise B.1.4

Here we list the powerset of $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TT}, \mathrm{TH}\}$, for easy of inspection events are grouped by cardinality (but note that cardinality has nothing to do with the probability measure in this exercise). I will denote by H_1 the event 'toss heads first' $\{\mathrm{HH}, \mathrm{HT}\}$, based on Table B.3 its probability is $\mathbb{P}(H_1) = 1/2$, which we compute via axiom (2).

$B \in \mathcal{A}$	$B \cap H_1$	$\mathbb{P}(B\|H_1)$
$\{\}$		0
$\{\mathrm{HH}\}$	$\{\mathrm{HH}\}$	${}^1/_3\boldsymbol{\nabla} \cdot {}^1/_2 = {}^2/_3$
$\{\mathrm{HT}\}$	$\{\mathrm{HT}\}$	${}^1/_6\boldsymbol{\nabla} \cdot {}^1/_2 = {}^1/_3$
$\{\mathrm{TT}\}$	$\emptyset$	0
$\{\mathrm{TH}\}$	$\emptyset$	0
$\{\mathrm{HH}, \mathrm{HT}\}$	$\{\mathrm{HH}, \mathrm{HT}\}$	${}^1/_2\boldsymbol{\nabla} \cdot {}^1/_2 = 1$
$\{\mathrm{HH}, \mathrm{TH}\}$	$\{\mathrm{HH}\}$	${}^1/_3\boldsymbol{\nabla} \cdot {}^1/_2 = {}^2/_3$
$\{\mathrm{HH}, \mathrm{TT}\}$	$\{\mathrm{HH}\}$	${}^1/_3\boldsymbol{\nabla} \cdot {}^1/_2 = {}^2/_3$
$\{\mathrm{HT}, \mathrm{TH}\}$	$\{\mathrm{HT}\}$	${}^1/_6\boldsymbol{\nabla} \cdot {}^1/_2 = {}^1/_3$
$\{\mathrm{HT}, \mathrm{TT}\}$	$\{\mathrm{HT}\}$	${}^1/_6\boldsymbol{\nabla} \cdot {}^1/_2 = {}^1/_3$
$\{\mathrm{TH}, \mathrm{TT}\}$	$\emptyset$	0
$\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}\}$	$\{\mathrm{HH}, \mathrm{HT}\}$	${}^1/_2\boldsymbol{\nabla} \cdot {}^1/_2 = 1$
$\{\mathrm{HH}, \mathrm{HT}, \mathrm{TT}\}$	$\{\mathrm{HH}, \mathrm{HT}\}$	${}^1/_2\boldsymbol{\nabla} \cdot {}^1/_2 = 1$
$\{\mathrm{HH}, \mathrm{TH}, \mathrm{TT}\}$	$\{\mathrm{HH}\}$	${}^1/_3\boldsymbol{\nabla} \cdot {}^1/_2 = {}^2/_3$
$\{\mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$	$\{\mathrm{HT}\}$	${}^1/_6\boldsymbol{\nabla} \cdot {}^1/_2 = {}^1/_3$
$\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$	$\{\mathrm{HH}, \mathrm{HT}\}$	${}^1/_2\boldsymbol{\nabla} \cdot {}^1/_2 = 1$

Exercise B.1.5

I will use H_2 to denote the event 'tossing two heads' $\{\mathrm{HH}\}$, and H_1 to denote the event 'tossing heads first' $\{\mathrm{HH}, \mathrm{HT}\}$. From Table B.3, the intersection of the two $H_2 \cap H_1 = \{\mathrm{HH}\}$ has probability ${}^1/_3$.

If the two events were independent, then $\mathbb{P}(H_2 \cap H_1) \overset{\text{indep.}}{=} \mathbb{P}(H_2) \times \mathbb{P}(H_1) = {}^1/_3 \times {}^1/_2 = {}^1/_6$.

As $\mathbb{P}(H_2 \cap H_1) \neq \mathbb{P}(H_2) \times \mathbb{P}(H_1)$, H_2 and H_1 are not independent events.

B.5 Additional Exercises

Exercise B.5.1

Draw a diagram such as that of Figure B.2 for the number of heads in three independent fair coin flips.

Exercise B.5.2

Describe the probability space of two coin flips, performed in order, where the first coin is crooked with odds $1 : 2$ for heads, and the second coin is fair. Evaluate the probability measure for every event in the powerset of the sample space, and indicate facts that can be used to support the probability value that you list. Valid facts include information provided in the exercise, axioms 1–3, or properties 4–6.

Exercise B.5.3

Assume $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space and $A_j \in \mathcal{A}$ is an event with

$\mathbb{P}(A_j) > 0$. Use the axioms of probability to prove that $\mathbb{P}(\cdot|A_j)$ is a probability measure.